Review of Statistical Learning

Mathematical Tools for ITS (11MAI)

Mathematical tools, 2020

Jan Přikryl

11MAI, lecture 5

Monday, November 2, 2020

version: 2020-11-02 00:41

Department of Applied Mathematics, CTU FTS

Lectue Contents

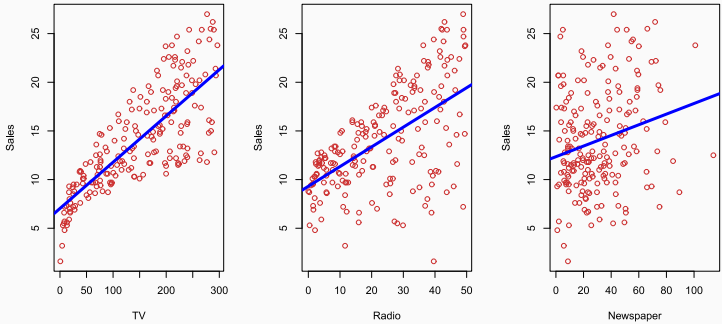
What you should alredy know from 11MAMY

A part of 11MAMY (4th semester) lectures concentrates on methods of "statistical learning", based on the text by James et al., Introduction to Statistical Learning.

In the next lectures we will discuss some of the topics in more detail. This lecture presents a very concentrated (but still quite long) summary of prerequisities, i.e. knowledge that you – at least in theory – possess. It is based on slides from 11MAMY.

We will speak mostly about *regression and classification*, *model selection*, *cross validation*, and *supervised and unsupervised learning*.

Co je statistické učení?



Zde jsou ukázány `Prodeje (Sales)` versus `TV`, `Rozhlas (Radio)` a `Noviny (Newspaper)` s modrou přímkou lineární regrese pro každý případ jednotlivě. Můžeme předpovídat `Prodeje` pomocí těchto tří diagramů? Možná se nám to povede lépe, použijeme-li nějaký model:

$$\texttt{Prodeje} \approx f(\texttt{TV},\texttt{Rozhlas},\texttt{Noviny})$$

Označení

Zde jsou `Prodeje` *odpověď* nebo *cílová hodnota*. Odpověď obvykle značíme y.
`TV` je *charakteristika*, *vlastnost*, *vstup* nebo *prediktor*, označíme ji x_1.
Podobně označíme `Rozhlas` jako x_2 a tak dále.

Na *vstupní vektor* můžeme pak souhrnně odkazovat jako na

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Náš model nyní zapíšeme jako

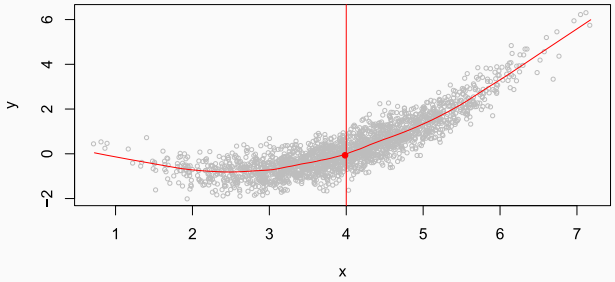
$$\mathbf{y} = f(\mathbf{x}) + \epsilon,$$

kde ϵ zachycuje chyby měření a jiné nepřesnosti. Funkci $f(X)$ nazýváme regresní funkce.

K čemu je $f(X)$ dobré?

Máme p různých prediktorů a každou sadu prediktorů změříme N-krát.

- S dobrým f můžeme dělat předpovědi Y v nových bodech $X = x$.
- Můžeme přijít na to, které složky $X = (X_1, X_2, \ldots, X_p)$ jsou pro pochopení Y důležité a které jsou irelevantní. Tak např. `Stáří` a `Roky vzdělávání` mají velký vliv na `Příjem`, ale `Rodinný stav` typicky ne.
- V závislosti na složitosti funkce f můžeme být schopni pochopit, jak každá složka X_j vektoru X ovlivňuje Y.



Existuje zde ideální $f(X)$? Konkrétně, co je dobrou hodnotou $f(X)$ pro libovolně zvolenou hodnotu X, řekněme $X = 4$? V bodě $X = 4$ může být mnoho hodnot Y. Dobrá hodnota funkce f je

$$f(4) = E(Y|X = 4).$$

$E(Y|X = 4)$ znamená *očekávanou hodnotu* (průměr) hodnot Y pro dané $X = 4$.

Tato ideální funkce $f(x) = E(Y|X = x)$ se nazývá *regresní funkce*.

Regresní funkce $f(x)$

- Je také definována pro vektor X; např.
  $f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$
- Je to *ideální* nebo *optimální* prediktor Y vzhledem ke střední kvadratické chybě:
  $f(x) = E(Y|X = x)$ je funkce, která minimalizuje $E[(Y - g(x))^2|X = x]$ přes všechny funkce g ve všech bodech $X = x$.
- $\epsilon = Y - f(x)$ je *neredukovatelná* (neodstranitelná) chyba — tj. i kdybychom znali $f(x)$, stejně bychom dělali chyby v předpovídání, neboť v každém bodě $X = x$ typicky existuje rozložení možných hodnot Y.
- Pro každý odhad $\hat{f}(x)$ funkce $f(x)$ máme

$$E[(Y - \hat{f}(x))^2|X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{redukovatelné}} + \underbrace{\text{var}(\epsilon)}_{\textit{neredukovatelné}}$$

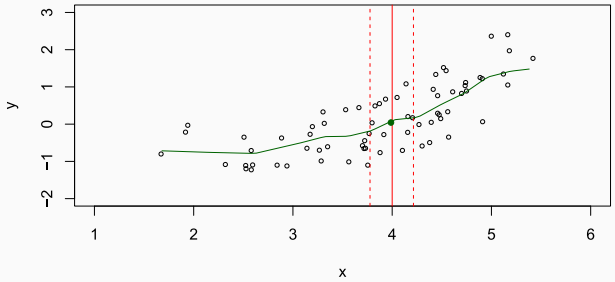
Jak odhadnout f – Metoda k nejbližších sousedů (k-NN)

Typicky máme málo bodů pro $X = x$ přesně (pokud vůbec nějaké). Nemůžeme tedy spočítat $\mathrm{E}[Y|X = x]$!

Zmírněme definici a položme

$$\hat{f}(x) = \text{average}(Y|X \in \mathcal{N}(x))$$

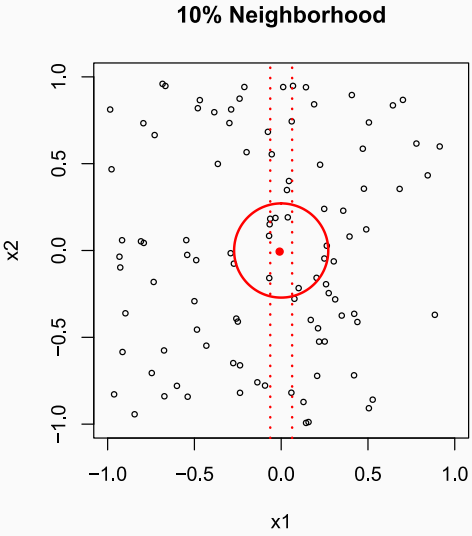
kde $\mathcal{N}(x)$ je nějaké vhodné *okolí* bodu x.

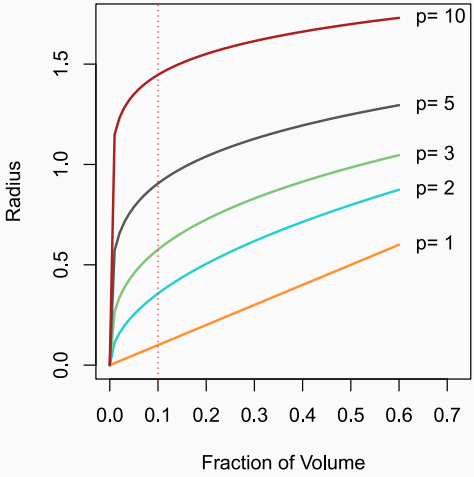


Jak odhadnout f – pokračování

- Metoda nejbližších sousedů může být *docela dobrá* pro malá p a spíše velká N.
- Metoda nejbližších sousedů může být *velmi špatná*, je-li p velké. Důvod: *prokletí dimensionality*. Ve více dimenzích mají nejbližší sousedé tendenci být hodně daleko.
  - Abychom snížili rozptyl, potřebujeme ke zprůměrování získat rozumný podíl N hodnot y_i, např. 10 %.
  - 10 % okolí ve vysokých dimenzích už nemusí být lokální, takže ztrácíme ducha odhadu $\mathrm{E}[Y|X = x]$ lokálním průměrováním.

Prokletí dimensionality





Parametrické a strukturované modely

Lineární model je důležitý příklad parametrického modelu:

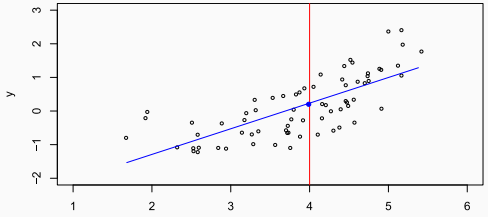
$$f_{\mathrm{L}}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

- Lineární model se specifikuje prostřednictvím $p + 1$ parametrů $\beta_0, \beta_1, \ldots, \beta_p$.
- Parametry odhadneme prokládáním modelu trénovacími daty.
- Ačkoli lineární model téměř *nikdy není správný*, často slouží jako dobrá a interpretovatelná aproximace neznámé skutečné funkce $f(X)$.

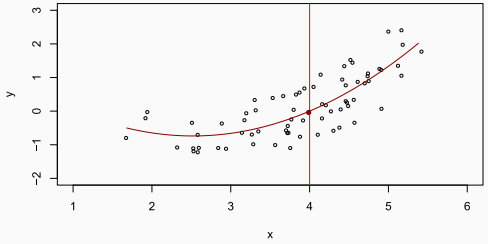
Nelineární model (zde kvadratický) může vystihnout data přesněji:

$$f_{\mathrm{N}}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \cdots + \beta_{q-1} X_p + \beta_q X_p^2.$$

Lineární model $\hat{f}_{\mathrm{L}}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ zde dává rozumnou aproximaci:



Kvadratický model $\hat{f}_{\mathrm{Q}}(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ prochází daty o něco lépe:



Posouzení přesnosti modelu

Předpokládejme, že prokládáme nějakými trénovacími daty $\texttt{Tr} = \{x_i, y_i\}_1^N$ model $\hat{f}(x)$ a chceme vědět, jak dobře si vede.

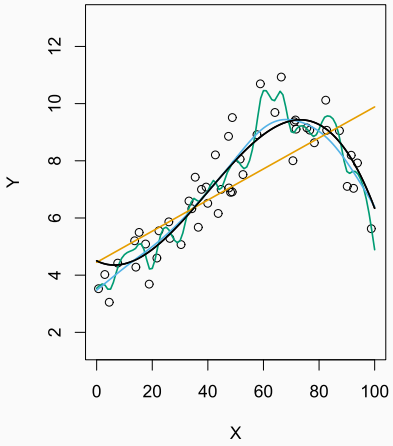
- Mohli bychom vypočítat střední kvadratickou chybu předpovědi přes $\texttt{Tr}$:

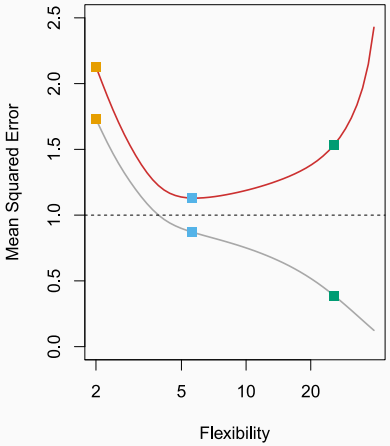
$$\text{MSE}_{\texttt{Tr}} = \text{Ave}_{i \in \texttt{Tr}}[y_i - \hat{f}(x_i)]^2$$

To může stranit více přeurčeným modelům.

- Místo toho bychom měli, pokud je to možné, vypočítat tu chybu pomocí nových *testovacích dat* $\texttt{Te} = \{x_i, y_i\}_1^M$:

$$\text{MSE}_{\texttt{Te}} = \text{Ave}_{i \in \texttt{Te}}[y_i - \hat{f}(x_i)]^2$$





Černá křivka je skutečnost. Červená křivka vpravo je $\mathrm{MSE}_{\mathtt{Te}}$, šedá křivka je $\mathrm{MSE}_{\mathtt{Tr}}$. Oranžová, modrá a zelená křivka (a čtverečky těchto barev) odpovídají aproximacím různé flexibility.

Kompromis mezi zkreslením a rozptylem

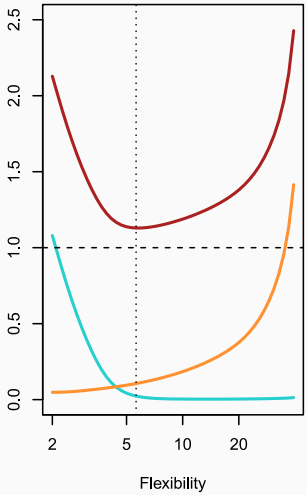
Předpokládejme, že jsme nějakými trénovacím daty `Tr` proložili model $\hat{f}(x)$ a nechť (x_0, y_0) je testovací pozorování vyvozené z této populace. Jestliže skutečný model je $Y = f(X) + \epsilon$ (kde $f(X) = \mathrm{E}[Y|X = x]$), pak

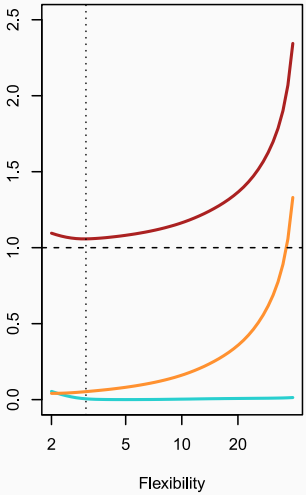
$$\mathrm{E}\left[y_0 - \hat{f}(x_0)\right]^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

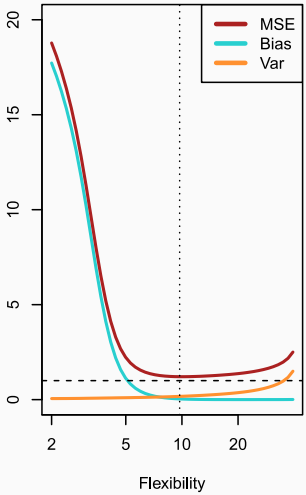
Očekávání počítá průměr přes variabilitu y_0 a rovněž variabilitu v `Tr`. Poznamenáváme, že $\mathrm{Bias}(\hat{f}(x_0)) = \mathrm{E}[\hat{f}(x_0)] - f(x_0)$.

Typické je, že jak roste *flexibilita* $\hat{f}$, roste její rozptyl a zkreslení (bias) se snižuje. Takže volba flexibility založená na střední testovací chybě odpovídá *kompromisu mezi zkreslením a rozptylem*.

Kompromis mezi zkreslením a rozptylem na našich třech příkladech







Klasifikační úlohy

Proměnná odpovědi Y je zde *kvalitativní* — např. email je jeden z prvků $\mathcal{C} = (\texttt{spam}, \texttt{ham})$ ($\texttt{ham}$ = dobrý email), třída číslic je jedna z $\mathcal{C} = \{0, 1, \ldots, 9\}$.

Naše cíle jsou:

- Vytvořit klasifikátor $C(X)$, který přiřadí značku třídy z $\mathcal{C}$ budoucímu neoznačenému pozorování X.
- Ohodnotit nejistotu v každé klasifikaci.
- Porozumět roli různých prediktorů mezi složkami $X = (X_1, X_2, \ldots, X_p)$.

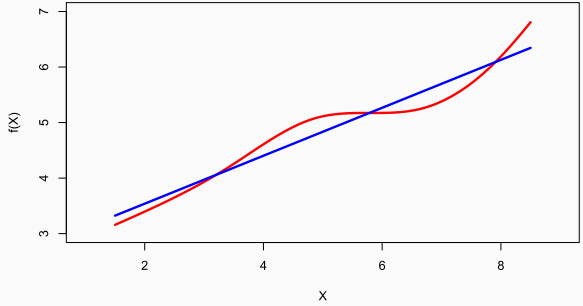
Stejně jako u regrese lze použít průměrování přes nejbližší sousedy.
A pro rostoucí dimenze se to stejně tak hroutí. Avšak dopad na $\hat{C}(x)$ je menší, než na $\hat{p}_k(x), k = 1, 2, \ldots, K$.

Lectue Contents

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Lineární regrese

- Lineární regrese je jednoduchý přístup k učení s učitelem (supervizovanému učení). Předpokládá, že závislost Y na $X_1, X_2, \ldots, X_p$ je lineární.
- Skutečné regresní funkce nejsou nikdy lineární!



- Ačkoli se může zdát přehnaně zjednodušená, je lineární regrese extrémně užitečná jak svou koncepcí, tak prakticky.

Jednoduchá lineární regrese s jediným prediktorem X

- Budeme uvažovat model
$$Y = \beta_0 + \beta_1 X + \epsilon,$$
kde β_0 a β_1 jsou dvě neznámé konstanty, které představují **regresní konstantu** (absolutní člen) a **sklon** (směrnici), říká se jim také **regresní koeficienty** nebo **parametry**,
- ϵ je chybový člen, často $\epsilon \approx \mathcal{N}(0, \sigma^2)$ (šum modelu).
- Jsou-li dány nějaké odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ koeficientů modelu, předpovídáme budoucí prodeje pomocí vzorce
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$
kde $\hat{y}$ označuje předpověď Y na základě $X = x$. Symbol **stříška** označuje odhadnutou hodnotu.

Odhad parametrů metodou nejmenších čtverců

- Nechť $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ je předpověď Y založená na i-té hodnotě X. Pak $e_i = y_i - \hat{y}_i$ představuje i-té **reziduum**.
- Definujeme **reziduální součet čtverců** (RSS) jako

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

nebo ekvivalentně jako

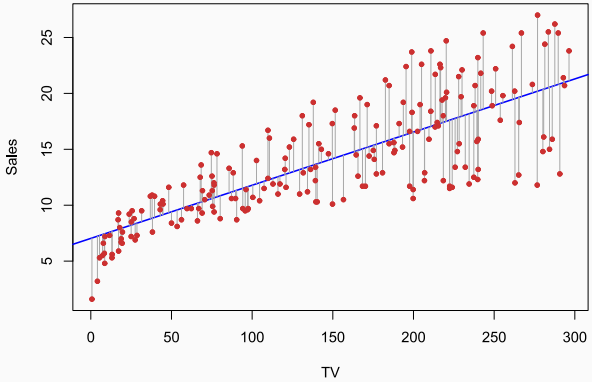
$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- Metoda nejmenších čtverců volí $\hat{\beta}_0$ a $\hat{\beta}_1$ tak, aby hodnota RSS byla minimální. Dá se ukázat, že tyto minimalizující hodnoty jsou

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

kde $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ jsou výběrové průměry.

Příklad: reklamní údaje



Výsledek metody nejmenších čtverců pro regresi položky `sales` vůči `TV`. V tomto případě lineární aproximace zachycuje podstatu vzájemného vztahu, i když na levém konci grafu je poněkud závadná.

Posouzení přesnosti odhadů koeficientů

- Směrodatná chyba odhadu odráží to, jak se odhad mění při opakovaném vzorkování. Máme

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$

  kde $\sigma^2 = \mathrm{var}(\epsilon)$.
- Tyto směrodatné chyby se mohou použít k výpočtu **intervalů spolehlivosti**. Interval spolehlivosti 95 % se definuje jako takový rozsah hodnot, že s pravděpodobností 95 % bude tento obor obsahovat skutečnou neznámou hodnotu daného parametru. Pro β_1 má tvar

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Znamená to, že je přibližně 95 % možnost, že interval

$$\left\langle \hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\, \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1) \right\rangle$$

bude obsahovat skutečnou hodnotu β_1 (ve scénáři, kdy jsme dostali opakované vzorky jako je současný vzorek).

Pro naše reklamní data je 95 % interval spolehlivosti $\langle 0{,}042, 0{,}053 \rangle$.

Testování hypotéz

Směrodatné chyby mohou být také použity k *testování hypotéz* o koeficientech.

- Nejběžnější test hypotézy je testování **nulové hypotézy** tvaru

  H_0: Mezi X a Y není žádný vzájemný vztah

  vůči **alternativní hypotéze**

  H_A: Mezi X a Y existuje nějaký vztah.

- Matematicky to odpovídá testování

$$H_0 : \beta_1 = 0$$

  vůči

$$H_A : \beta_1 \neq 0,$$

  neboť pokud $\beta_1 = 0$, model se redukuje na $Y = \beta_0 + \epsilon$ a X s Y není propojeno.

- K testu nulové hypotézy vypočítáme t-**statistiku** danou vztahem

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

- Ta bude mít t-rozdělení s $n - 2$ stupni volnosti, za předpokladu, že $\beta_1 = 0$.
- Pomocí statistického softwaru se snadno vypočítá pravděpodobnost, že budeme pozorovat jakoukoli hodnotu rovnou $|t|$ nebo větší. Tato pravděpodobnost se nazývá p-**hodnota**.

Výsledky pro reklamní údaje

	Koeficient	Směr. chyba	t-statistika	p-hodnota
Regresní konstanta	7,0325	0,4578	15,36	$< 0{,}0001$
TV	0,0475	0,0027	17,67	$< 0{,}0001$

Posouzení celkové přesnosti modelu

- Vypočítáme **reziduální směrodatnou chybu**

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\,\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  kde **reziduální součet čtverců** je $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- **R kvadrát** neboli koeficient determinace je

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}},$$

  kde $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ je **celkový součet čtverců**.
- Dá se ukázat, že v této jednoduché lineární regresní situaci je $R^2 = r^2$, kde r je korelace mezi X a Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2)}}.$$

Výsledky pro reklamní údaje

Veličina	Hodnota
Reziduální směrodatná chyba	3,26
R^2	0,612
F-statistika	312,1

Vícenásobná lineární regrese

Vícenásobná lineární regrese zahrnuje dva či více regreosrů.

- Náš model je v tomto případě

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- Koeficient β_j interpretujeme jako *průměrný* vliv jednoho jednotkového růstu X_j na Y, za předpokladu, že **všechny ostatní prediktory se nemění**. V příkladu s reklamou nabývá model tvaru

$$\texttt{prodeje} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{rozhlas} + \beta_3 \times \texttt{noviny} + \epsilon.$$

Interpretace regresních koeficientů

- Ideální scénář je tehdy, když prediktory nejsou korelovány – **vyvážený plán**:
  - Každý koeficient může být odhadnut a testován odděleně.
  - Jsou možné interpretace typu **"jednotková změna v X_j je spojena se změnou Y o β_j, přičemž všechny ostatní proměnné zůstávají beze změny"**.
- Korelace mezi prediktory působí problémy:
  - Rozptyl všech koeficientů má tendenci růst, někdy dramaticky.
  - Interpretace se stávají hazardními – když se změní X_j, změní se všechno ostatní.
- Měli bychom se vyhnout **kauzalitě** v pozorovaných datech.

Rozhodování o důležitých proměnných

- Nejpřímější přístup se nazývá regrese se **všemi podmnožinami** nebo s **nejlepší podmnožinou**: počítáme aproximace metodou nejlepších čtverců pro všechny možné podmnožiny a pak z nich vybereme pomocí nějakého kritéria, které vyvažuje trénovací chybu s velikostí modelu.
- Často však nemůžeme vyšetřit všechny možné modely, protože je jich 2^p; například pro $p = 40$ existuje přes miliardu modelů!
  Místo toho potřebujeme automatizovaný přístup, který prohledává nějakou jejich podmnožinu.

Dopředná selekce

- Začni s **nulovým modelem** — modelem, který obsahuje regresní konstantu, ale žádné prediktory.
- Prolož p jednoduchých lineárních regresí a přidej k nulovému modelu tu proměnnou, která vede k nejnižšímu RSS.
- Přidej k tomu modelu proměnnou, která vede k nejnižšímu RSS mezi všemi modely s dvěma proměnnými.
- Pokračuj, dokud není splněno nějaké zastavovací kritérium, například že všechny zbývající proměnné mají p-hodnotu nad nějakou hranicí.

Zpětná eliminace

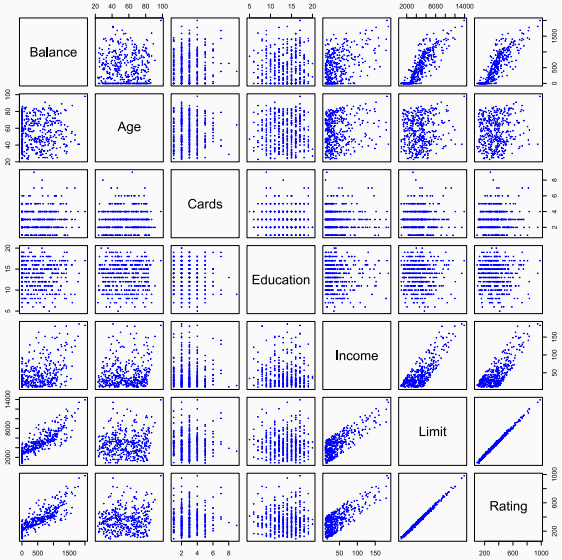
- Začni se všemi proměnnými v modelu.
- Odeber proměnnou s největší p-hodnotou – to jest proměnnou, která je nejméně statisticky významná.
- Prolož nový model s $p-1$ proměnnými a odeber proměnnou s největší p-hodnotou.
- Pokračuj do splnění nějakého zastavovacího kritéria. Můžeme například zastavit, když všechny zbývající proměnné mají významnou p-hodnotu definovanou jako nějaká hranice významnosti.

Kvalitativní prediktory

- Některé prediktory nejsou **kvantitativní**, ale jsou **kvalitativní**, nabývají hodnot v diskrétní množině.
- Nazývají se také **kategoriální** (nikoliv kategorické) prediktory nebo **proměnné faktory**.
- Viz například matici bodových grafů s údaji o kreditních kartách na následujícím slajdu.

  Kromě sedmi kvantitativních proměnných, jež jsou v matici uvedeny, jsou v datech čtyři kvalitativní proměnné: `gender` (pohlaví), `student` (studentský status), `status` (rodinný stav) a `ethnicity` (původ – kavkazský, afroamerický (AA) nebo asijský).

Údaje o kreditních kartách



Example

Vyšetřete rozdíl v zůstatku na kreditní kartě mezi muži a ženami, přičemž budete ignorovat ostatní proměnné.

Utvoříme novou proměnnou

$$x_i = \begin{cases} 1 & \text{je-li } i\text{-tá osoba žena,} \\ 0 & \text{je-li } i\text{-tá osoba muž.} \end{cases}$$

Výsledný model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{je-li } i\text{-tá osoba žena,} \\ \beta_0 + \epsilon_i & \text{je-li } i\text{-tá osoba muž.} \end{cases}$$

Údaje o kreditních kartách — pokračování

Interpretace?

Výsledky pro model podle pohlaví:

	Koef.	SE	t-statistika	p-hodnota
Regresní konst.	509,80	33,13	15,389	$< 0{,}0001$
`gender` [žena]	19,73	46,05	0,429	0,6690

- Při více než dvou úrovních utvoříme dodatečné fiktivní proměnné. Tak například pro proměnnou `ethnicity` utvoříme dvě fiktivní proměnné. První by mohla být

$$x_{i1} = \begin{cases} 1 & \text{je-li } i\text{-tá osoba asijského původu,} \\ 0 & \text{není-li } i\text{-tá osoba asijského původu,} \end{cases}$$

a druhá by mohla být

$$x_{i2} = \begin{cases} 1 & \text{je-li } i\text{-tá osoba kavkazského původu,} \\ 0 & \text{není-li } i\text{-tá osoba kavkazského původu.} \end{cases}$$

- Pak mohou být v regresní rovnici použity obě tyto proměnné, takže dostaneme model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{je-li } i\text{-tá osoba Asiat} \\ \beta_0 + \beta_2 + \epsilon_i & \text{je-li } i\text{-tá osoba běloch} \\ \beta_0 + \epsilon_i & \text{je-li } i\text{-tá osoba Afroameričan.} \end{cases}$$

- Fiktivních proměnných bude vždy o jednu méně než je počet úrovní. Úroveň bez fiktivní proměnné — v tomto příkladu Afroameričani — je známa jako **výchozí úroveň**.

Výsledky pro etnickou příslušnost

	Koef.	SE	t-statistika	p-hodnota
Regresní konst.	531,00	46,32	11,464	$< 0{,}0001$
`ethnicity` [asijská]	-18,69	65,02	-0,287	0,7740
`ethnicity` [kavkazská]	-12,50	56,68	-0,221	0,8260

Odstraníme předpoklad aditivity: **interakce** a **nelinearita**

Interakce:

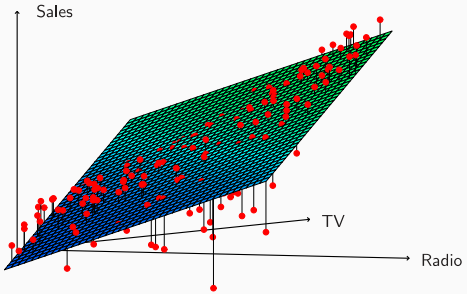
- V naší předchozí analýze reklamních dat jsme předpokládali, že vliv zvýšení prostředků jednoho reklamního média na `sales` (prodeje) nezávisí na objemu prostředků vynaložených na zbylá média.
- Tak například, lineární model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

říká, že průměrný efekt jednotkového vzrůstu reklamních nákladů v `TV` na `sales` je vždy β_1, nezávisle na množství prostředků vynaložených na `radio`.

- Ale předpokládejme, že peníze vynaložené na rozhlasovou reklamu ve skutečnosti zvyšují efektivitu TV reklamy, takže koeficient sklonu pro `TV` by měl s růstem hodnoty `radio` růst.
- V této situaci, máme-li dán pevný rozpočet \$100 000, investice poloviny do `radio` a poloviny do `TV` může zvýšit `sales` více, než použití celé částky na `TV` nebo na `radio`.
- V marketingu se tomuhle říká efekt **synergie**, ve statistice se o tom mluví jako o efektu **interakce**.

Interakce v reklamních datech?



Když je úroveň `TV` nebo `radio` nízká, pak jsou skutečné hodnoty `sales` nižší, než jak předpovídá lineární model.

Ale když se reklama rozdělí mezi tato dvě média, pak má model tendenci `sales` podhodnocovat.

Model nabývá tvaru

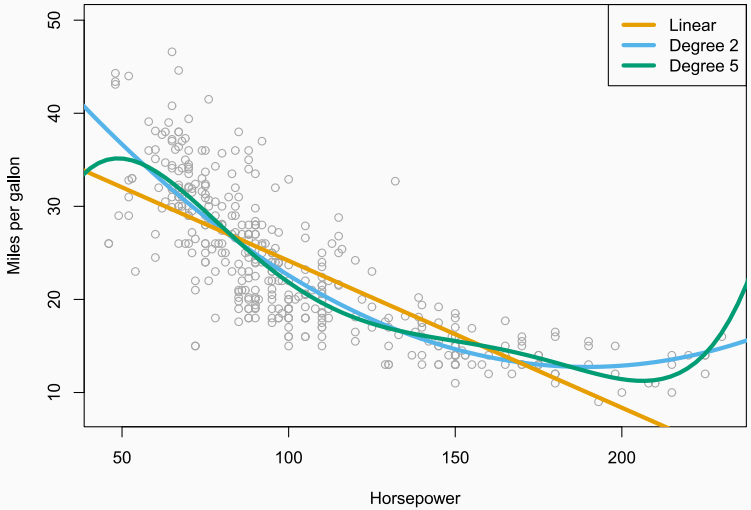
$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

Výsledky:

	Koef.	SE	t-statistika	p-hodnota
Regres. konst.	6,7502	0,248	27,23	$< 0{,}0001$
TV	0,0191	0,002	12,70	$< 0{,}0001$
radio	0,0289	0,009	3,24	0,0014
TV × radio	0,0011	0,000	20,73	$< 0{,}0001$

- Někdy se stane, že interakční člen má velmi malou p-hodnotu, ale přidružené hlavní efekty (v tomto případě `TV` a `radio`) nikoliv.
- **Princip hierarchie:** *Pokud zahrneme do modelu nějakou interakci, měli bychom také zahrnout hlavní efekty, a to i tehdy, nejsou-li p-hodnoty spojené s jejich koeficienty významné.*
- Odůvodněním tohoto principu je skutečnost, že interakce se v modelu bez hlavních efektů obtížně interpretují — jejich smysl se změní.
- Speciálně, interakční členy také obsahují hlavní efekty i tehdy, když model nemá členy s hlavními efekty.

polynomiální regrese údajů o automobilech



Obrázek naznačuje, že

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

může dávat lepší aproximaci.

	Koeficient	SE	t-statistika	p-hodnota
Regres. konst.	56,9001	1,8004	31,6	$< 0,0001$
horsepower	-0,4662	0,0311	-15,0	$< 0,0001$
horsepower^2	0,0012	0,0001	10,1	$< 0,0001$

Lectue Contents

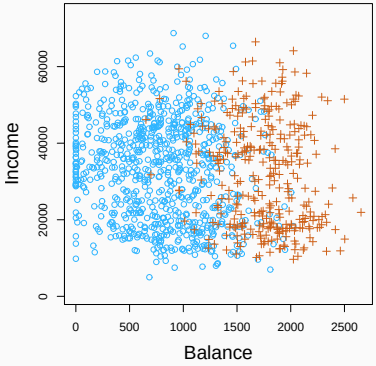
- Kvalitativní proměnné nabývají hodnot z neuspořádané množiny $\mathcal{C}$, například

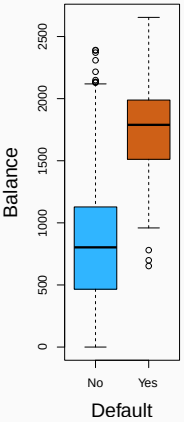
$$\texttt{eye color} \in \{\texttt{brown}, \texttt{blue}, \texttt{green}\}$$
$$\texttt{email} \in \{\texttt{spam}, \texttt{ham}\}.$$

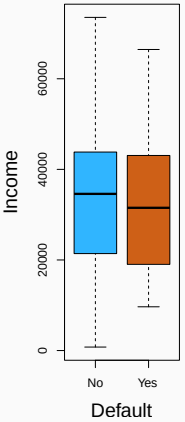
- Pro daný vektor charakteristik X a kvalitativní odpověď Y nabývající hodnot z množiny $\mathcal{C}$ spočívá klasifikační úloha ve vytváření funkce $C(X)$, která jako vstup bere vektor charakteristik X a předpovídá hodnotu Y, tj. $C(X) \in \mathcal{C}$.
- Často nás spíše zajímají odhady **pravděpodobností**, že X patří do té které kategorie v $\mathcal{C}$.

  Je například hodnotnější mít odhad pravděpodobnosti, že pojistný nárok je podvodný, než klasifikaci podvodný nebo ne.

Příklad: Neplatiči kreditních karet







`Default` = neplatič

Můžeme použít lineární regresi?

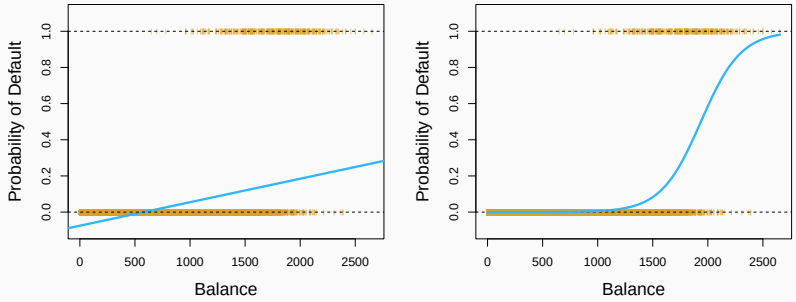
Předpokládejme, že pro klasifikační úlohu `Default` kódujeme

$$Y = \begin{cases} 0 & \text{jestliže } \texttt{No} \\ 1 & \text{jestliže } \texttt{Yes}. \end{cases}$$

Můžeme prostě provést lineární regresi Y vzhledem k X a klasifikovat jako `Yes`, jestliže $\hat{Y} > 0{,}5$?

- V tomto případě binárního výstupu odvádí lineární regrese jako klasifikátor dobrou práci a je ekvivalentní **lineární diskriminační analýze**, kterou budeme probírat později.
- Protože v dané populaci $\mathrm{E}[Y|X = x] = \mathrm{P}(Y = 1|X = x)$, mohli bychom si myslet, že regrese je pro tuto úlohu perfektní.
- Lineární regrese však může produkovat hodnoty pravděpodobnosti menší než nula nebo větší než jedna. Vhodnější je zde **logistická regrese**.

Lineární versus logistická regrese



Svisle: Pravděpodobnost nesplácení

Oranžové značky označují odpověď Y (0 nebo 1). Lineární regrese neodhaduje $P(Y = 1|X)$ dobře. Logistická regrese se zdá být pro tuto úlohu zcela vhodná.

Nyní předpokládejme, že odpověď Y může nabývat tří hodnot. Do pohotovostní místnosti se dostaví pacient a my jej musíme klasifikovat podle jeho symptomů:

$$Y = \begin{cases} 1 & \text{pokud } \texttt{mrtvice} \\ 2 & \text{pokud } \texttt{předávkování léky} \\ 3 & \text{pokud } \texttt{epileptický záchvat} \end{cases}$$

Toto kódování naznačuje, že je zde uspořádání, což ve skutečnosti implikuje, že rozdíl mezi `mrtvice` a `předávkování léky` je stejný jako mezi `předávkování léky` a `epileptický záchvat`.

Lineární regrese zde není vhodná.

Vhodnější jsou **vícetřídní logistická regrese** nebo **diskriminační analýza**.

Logistická regrese

Budeme pro zkrácení psát $p(X) = \mathrm{P}(Y = 1|X)$ a budeme uvažovat použití proměnné `balance` k předpovídání `default`.

Logistická regrese používá výraz

$$p(X) = \frac{\mathrm{e}^{\beta_0+\beta_1 X}}{1 + \mathrm{e}^{\beta_0+\beta_1 X}}$$

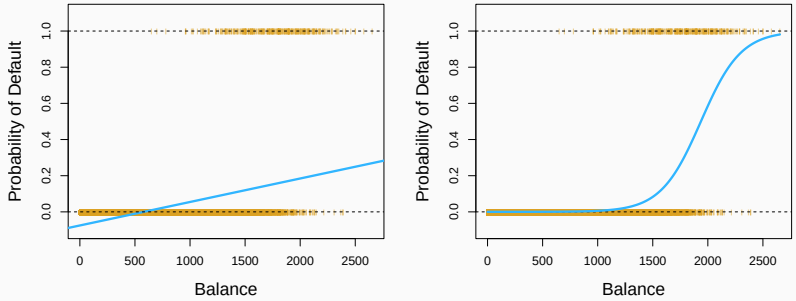
($\mathrm{e} \approx 2{,}71828$ je matematická konstanta — Eulerovo číslo). Je snadné vidět, že bez ohledu na hodnoty β_0, β_1 nebo X bude $p(X)$ nabývat hodnot mezi 0 a 1.

Jednoduchá úprava dává

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

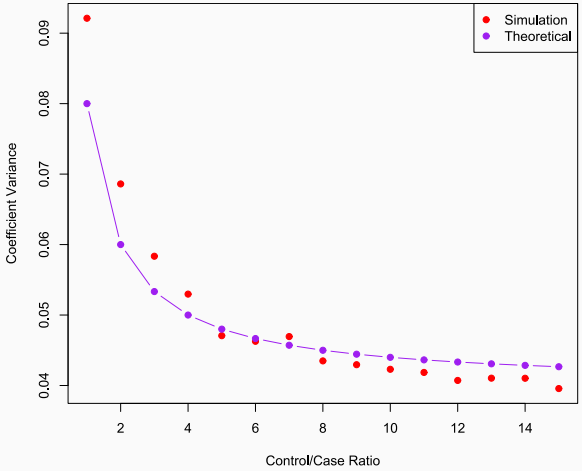
Tato monotónní transformace se nazývá **log odds** (logaritmus rizika) nebo **logitová transformace** (logit) $p(X)$.

Lineární versus logistická regrese



Logistická regrese zaručuje, že náš odhad $p(X)$ bude ležet mezi 0 a 1.

Snižování rozptylu v nevyvážených binárních datech



Vzorkování více kontrol než případů snižuje rozptyl odhadů parametrů. Ale po dosažení poměru zhruba 5 k 1 se snižování rozptylu zastavuje.

Diskriminační analýza

Při *diskriminační analýze* modelujeme pravděpodobnostní rozdělení X v každé třídě odděleně a pak použijeme **Bayesovu větu** k obrácenému pohledu na věc a k získání $P(Y|X)$.

Použijeme-li v každé třídě normální (Gaussovo) rozdělení, vede to k lineární nebo kvadratické diskriminační analýze.

Nicméně je tento přístup zcela obecný a mohou být použita rovněž jiná rozdělení. My se soustředíme na normální rozdělení.

Bayesova věta pro klasifikaci

Thomas Bayes byl známý matematik, jehož jméno je spojeno s velkou podoblastí statistického a pravděpodobnostního modelování. Zde se soustředíme na jeden jednoduchý výsledek známý jako **Bayesova věta**:

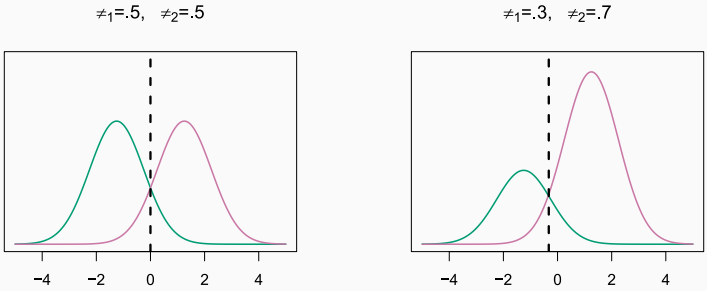
$$\mathrm{P}(Y = k|X = x) = \frac{\mathrm{P}(X = x|Y = k) \cdot \mathrm{P}(Y = k)}{\mathrm{P}(X = x)}$$

Pro diskriminační analýzu se to zapisuje trochu odlišně:

$$\mathrm{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}, \quad \text{kde}$$

- $f_k(x) = \mathrm{P}(X = x|Y = k)$ je **hustota** X ve třídě k. Budeme zde používat normální hustoty, odděleně v každé třídě.
- $\pi_k = \mathrm{P}(Y = k)$ je marginální nebo **apriorní** pravděpodobnost pro třídu k.

Klasifikace na nejvyšší hustotu



Nový bod klasifikujeme podle toho, která hustota je nejvyšší.

Jsou-li apriorní pravděpodobnosti odlišné, bereme to rovněž v úvahu, a porovnáváme $\pi_k f_k(x)$. V obrázku napravo upřednostňujeme purpurovou třídu — rozhodovací hranice se posunula doleva.

Proč diskriminační analýza?

- Jsou-li třídy dobře oddělené, jsou odhady parametrů u logistického regresního modelu překvapivě nestabilní. Lineární diskriminační analýza tímto problémem netrpí.
- Pokud n je malé a rozdělení prediktorů X je v každé ze tříd přibližně normální, je lineární diskriminační model opět stabilnější než logistický regresní model.
- Lineární diskriminační analýza je oblíbená v situacích, kdy máme více než dvě třídy odpovědí, protože rovněž poskytuje zobrazení dat v méně dimenzích.

Lineární diskriminační analýza při $p = 1$

Gaussova hustota má tvar

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}.$$

Zde je μ_k střední hodnota a σ_k^2 je rozptyl (ve třídě k). Budeme předpokládat, že všechny hodnoty $\sigma_k = \sigma$ jsou stejné.

Dosadíme-li toto do Bayesova vzorce, dostaneme poměrně komplikovaný výraz pro $p_k(x) = \mathrm{P}(Y = k|X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu_\ell}{\sigma}\right)^2}}$$

Naštěstí je zde možnost zjednodušení a krácení.

Diskriminační funkce

Abychom klasifikovali v hodnotě $X = x$, potřebujeme zjistit, která z hodnot $p_k(x)$ je největší. Zlogaritmujeme a odstraníme členy, které nezávisejí na k, a zjistíme tak, že toto je ekvivalentní přiřazení x do třídy s největším **diskriminačním skóre**:

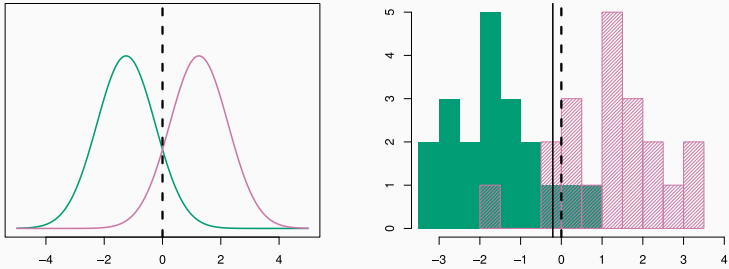
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Všimněte si, že $\delta_k(x)$ je **lineární** funkce x.

Jestliže máme $K = 2$ třídy a $\pi_1 = \pi_2 = 0{,}5$, pak se dá ukázat, že **rozhodovací hranice** je v bodě

$$x = \frac{\mu_1 + \mu_2}{2}.$$

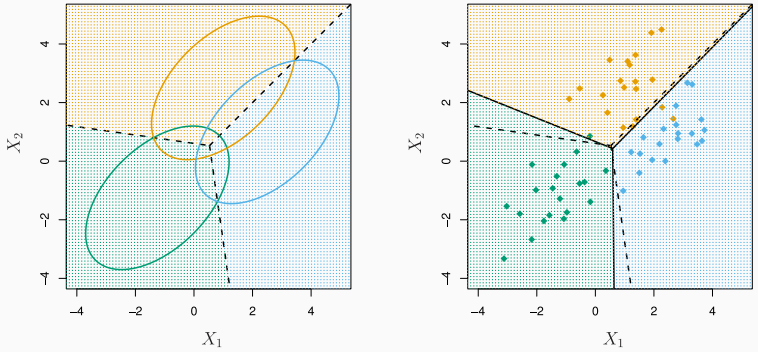
(Zkuste to ukázat sami.)



Příklad s $\mu_1 = -1{,}5$, $\mu_2 = 1{,}5$, $\pi_1 = \pi_2 = 0{,}5$ a $\sigma^2 = 1$.

V typické situaci tyto parametry neznáme; máme jen trénovací data. V takovém případě prostě parametry odhadneme a dosadíme je do příslušného vzorce.

Ilustrace: $p = 2$ a $K = 3$ třídy



Je zde $\pi_1 = \pi_2 = \pi_3 = 1/3$.

Čárkované úsečky jsou známy jako **Bayesovy rozhodovací hranice**. Pokud by byly známy, poskytovaly by nejméně chyb se špatnou klasifikací mezi všemi možnými klasifikátory.

Od $\delta_k(x)$ k pravděpodobnostem

Jakmile máme odhady $\hat{\delta}_k(x)$, můžeme je převést na odhady pravděpodobností tříd:

$$\hat{\mathrm{P}}(Y = k|X = x) = \frac{\mathrm{e}^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} \mathrm{e}^{\hat{\delta}_l(x)}}.$$

Klasifikace na největší $\hat{\delta}_k(x)$ tak znamená klasifikaci do třídy, pro niž je $\hat{\mathrm{P}}(Y = k|X = x)$ největší.

Jestliže $K = 2$, klasifikujeme do třídy 2, pokud $\hat{\mathrm{P}}(Y = k|X = x) \geq 0{,}5$, jinak do třídy 1.

LDA na kreditních datech

		Skutečný stav nesplácení		
		Ne	Ano	Celkem
Předpovězený	Ne	9644	252	9896
stav nesplácení	Ano	23	81	104
	Celkem	9667	333	10000

$(23 + 252)/10000$ chyb — míra chybné klasifikace je 2,75%!

Některá upozornění:

- Toto je **trénovací chyba**, model je možná přeurčený. Tady nás to příliš neznepokojuje, protože zde $n = 10000$ a $p = 4$.
- Pokud bychom klasifikovali podle apriorní pravděpodobnosti – vždy do třídy `Ne` v tomto případě — udělali bychom $333/10000$ chyb, neboli pouze 3,33%.
- Na skutečných `Ne` děláme $23/9667 = 0{,}2\%$ chyb; na skutečných `Ano` děláme $252/333 = 75{,}7\%$ chyb!

Typy chyb

Míra falešných pozitiv: Podíl negativních příkladů, které jsou klasifikovány jako pozitivní — 0,2% v našem příkladu.

Míra falešných negativ: Podíl pozitivních příkladů, které jsou klasifikovány jako negativní — 75,7% v našem příkladu.

Tuto tabulku jsme vytvořili tak, že jsme klasifikovali do třídy `Ano`, pokud

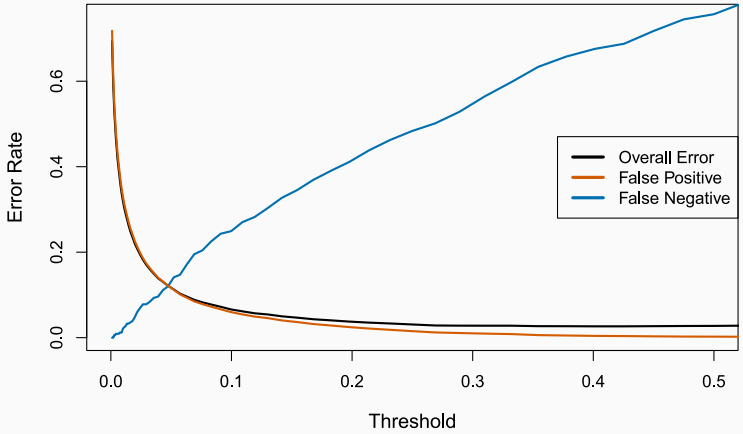
$$\hat{\mathrm{P}}(\texttt{Neplatič} = \texttt{Ano} | \texttt{Zůstatek}, \texttt{Student}) \geq 0{,}5.$$

Ty dvě míry chyb můžeme pozměnit tak, že změníme prahovou hodnotu z 0,5 na nějakou jinou hodnotu v intervalu $[0, 1]$:

$$\hat{\mathrm{P}}(\texttt{Neplatič} = \texttt{Ano} | \texttt{Zůstatek}, \texttt{Student}) \geq \textit{práh}$$

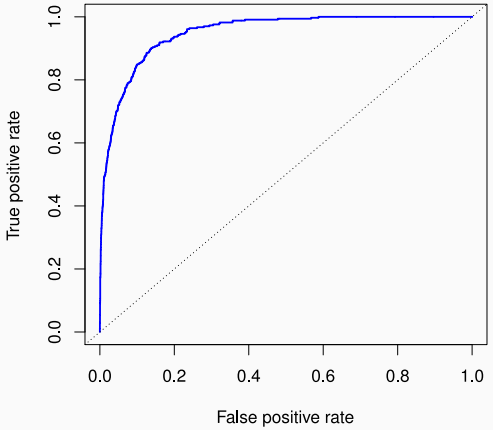
a měníme *práh*.

Měníme *prahovou hodnotu*



Abychom snížili míru falešných negativ, můžeme chtít snížit prahovou hodnotu na 0,1 nebo méně.

Křivka ROC



Graf ROC zobrazuje obě míry současně. Někdy se používá **AUC** neboli **area under the curve** (oblast pod křivkou) k vyhodnocení celkové účinnosti. Větší **AUC** je dobré.

Logistická regrese versus LDA

Pro úlohu s dvěma třídami se dá ukázat, že pro LDA platí

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \cdots + c_p x_p.$$

Má tedy stejný tvar jako logistická regrese.

Rozdíl je v tom, jak jsou odhadovány parametry.

- Logistická regrese používá podmíněnou věrohodnost založenou na $\mathrm{P}(Y|X)$ (je to známo jako **diskriminační učení**).
- LDA používá úplnou věrohodnost založenou na $\mathrm{P}(X, Y)$ (to je známo jako **generativní učení**).
- Nehledě na tyto rozdíly jsou výsledky v praxi často velmi podobné.

Poznámka: logistická regrese může také prokládat kvadratické hranice jako QDA, a to tak, že se do modelu explicitně zahrnou kvadratické členy.

Shrnutí

- Logistická regrese je pro klasifikaci velmi oblíbená, zejména při $K = 2$.
- LDA je užitečná, je-li n malé nebo jsou-li třídy dobře odděleny, a jsou-li rozumné předpoklady o Gaussiánu. Také při $K > 2$.
- Naivní Bayesův klasifikátor je užitečný, je-li p velmi velké.
- V odst. 4.5 lze nalézt materiál k porovnání logistické regrese, LDA a KNN.

Lectue Contents

Připomeňme si lineární model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- Nehledě na svou jednoduchost má lineární model zřetelné výhody co do své **interpretovatelnosti** a často vykazuje dobré **výsledky v předpovědích**.
- Proč se zabývat alternativami k nejmenším čtvercům?
  - **Přesnost předpovědi:** Zejména při $p > n$, k regulaci rozptylu.
  - **Interpretovatelnost modelu:** Odstraněním nepodstatných vlastností — to jest tím, že položíme odpovídající odhady koeficientů rovny nule — můžeme získat model, který se snáze interpretuje. Uvedeme některé přístupy k automatickému provádění **volby vlastností.**

Tři třídy metod

- **Výběr podmnožiny.** Identifikujeme podmnožinu z těch p prediktorů, o níž soudíme, že má vztah k odpovědi. Pak proložíme nejmenšími čtverci model tou redukovanou množinou proměnných.
- **Smršťování.** Proložíme model zahrnující všech p prediktorů, ale odhadnuté koeficienty srazíme směrem k nule vzhledem k odhadům nejmenších čtverců. Toto smrštění (známé také jako **regularizace**) má efekt ve snížení rozptylu a může také provádět výběr proměnných.
- **Dimenzionální redukce.** Promítneme těch p prediktorů na M-rozměrný podprostor, kde $M < p$. Toho se dosáhne tak, že vypočítáme M různých **lineárních kombinací**, neboli **projekcí** těch proměnných. Pak se těchto M projekcí použije jako prediktory k proložení lineárního regresního modelu nejmenšími čtverci.

Algoritmy pro model s výběrem nejlepší podmnožiny a postupný výběr modelu

Výběr nejlepší podmnožiny

1. Označme $\mathcal{M}_0$ **nulový model**, který neobsahuje žádné prediktory. Tento model pro každé pozorování prostě předpovídá střední hodnotu vzorku.
2. Pro $k = 1, 2, \ldots, p$:
   (a) Prolož všech $\binom{p}{k}$ modelů, které obsahují přesně k prediktorů.
   (b) Vyber z těchto $\binom{p}{k}$ modelů ten nejlepší a označ jej $\mathcal{M}_k$. **Nejlepší** se zde definuje jako mající nejmenší RSS nebo ekvivalentně největší R^2.
3. Vyber jediný nejlepší model z modelů $\mathcal{M}_0, \ldots, \mathcal{M}_p$ na základě chyby křížové validace, C_p (AIC), BIC nebo upraveného R^2.

Postupný výběr

- Výběr nejlepší podmnožiny se z výpočtových důvodů nedá použít pro velmi velká p. **Proč ne?**
- Když p je velké, může výběr nejlepší podmnožiny také trpět statistickými problémy: čím větší prostor pro vyhledávání, tím větší je šance najít modely, které na tréninkových datech vypadají dobře, i když na budoucích datech nemusejí mít žádnou vypovídací hodnotu.
- Enormní prostor pro vyhledávání tudíž může vést k **přeurčení** a vysokému rozptylu odhadů koeficientů.
- Z obou těchto důvodů jsou atraktivními alternativami k výběru nejlepší podmnožiny metody **postupného výběru**, které zkoumají mnohem omezenější soubor modelů.

Postupná dopředná selekce

- Postupná dopředná selekce začíná modelem, který neobsahuje žádné prediktory, a pak k modelu prediktory přidává jeden po druhém, dokud v modelu nejsou všechny prediktory.
- Konkrétně se v každém kroku k modelu přidává proměnná, která prokládané aproximaci dává největší **dodatečné** zlepšení.

Postupná dopředná selekce

1. Označme $\mathcal{M}_0$ **nulový** model, který neobsahuje žádné prediktory.
2. Pro $k = 0, \ldots, p-1$:
   - 2.1 Uvažujme všech $p-k$ modelů, které rozšiřují prediktory v $\mathcal{M}_k$ o jeden prediktor navíc.
   - 2.2 Vyberme **nejlepší** z těchto $p-k$ modelů a nazvěme jej $\mathcal{M}_{k+1}$. **Nejlepší** zde znamená, že model má nejmenší RSS nebo největší R^2.
3. Vybereme jediný nejlepší model z modelů $\mathcal{M}_0, \ldots, \mathcal{M}_p$ na základě chyby předpovědi křížové validace, C_p (AIC), BIC nebo upraveného R^2.

Postupná zpětná eliminace

- Podobně jako postupná dopředná selekce představuje **postupná zpětná eliminace** efektivní alternativu k výběru nejlepší podmnožiny.
- Avšak na rozdíl od postupné dopředné selekce začíná úplným modelem nejmenších čtverců obsahujícím všech p prediktorů a pak iterativně odstraňuje jeden po druhém nejméně užitečné prediktory.

Postupná zpětná eliminace

1. Označme $\mathcal{M}_p$ **úplný** model, který obsahuje všech p prediktorů.
2. Pro $k = p, p-1, \ldots, 1$:
   - 2.1 Uvažujme všech k modelů, které obsahují všechny prediktory z $\mathcal{M}_k$ kromě jednoho, takže mají celkem $k-1$ prediktorů.
   - 2.2 Vybereme z těchto k modelů ten **nejlepší** a označíme jej $\mathcal{M}_{k-1}$. **Nejlepším** je zde míněn model s nejmenším RSS nebo největším R^2.
3. Vybereme jediný nejlepší model z modelů $\mathcal{M}_0, \ldots, \mathcal{M}_p$ na základě chyby předpovědi křížové validace, C_p (AIC), BIC nebo upraveného R^2.

Výběr optimálního modelu

- Model obsahující všechny prediktory bude vždy mít nejmenší RSS a největší R^2, neboť tyto veličiny se vztahují k trénovací chybě.
- Přejeme si zvolit model s nízkou testovací chybou, ne model s nízkou trénovací chybou. Připomínáme, že trénovací chyba je obvykle špatným odhadem testovací chyby.
- V důsledku toho nejsou RSS a R^2 vhodné pro výběr nejlepšího modelu z kolekce modelů s různými počty prediktorů.

Odhadování testovací chyby: dva přístupy

Testovací chybu můžeme odhadnout

- *nepřímo* tak, že provedeme **úpravu trénovací chyby**, která vezme v úvahu zkreslení působené přeurčováním. To vede na Mallowovo C_p, BIC či upravené R^2.
- *přímo* buď použitím přístupu s validačním souborem nebo použitím křížové validace.

Hřebenová regrese a metoda Lasso

- Metody výběru podmnožiny používají nejmenší čtverce k prokládání lineárního modelu, který obsahuje podmnožinu prediktorů.
- Jako alternativu můžeme proložit model obsahující všech p prediktorů pomocí techniky, která **omezuje** nebo **regularizuje** odhady koeficientů, nebo ekvivalentně, která **smršťuje** odhady koeficientů směrem k nule.
- Není možná bezprostředně zřejmé, proč by takové omezení mělo prokládanou aproximaci vylepšit, ale ukazuje se, že smrštění odhadů koeficientů může významně snížit jejich rozptyl.

- Připomínáme, že metoda nejmenších čtverců odhaduje $\beta_0, \beta_1, \ldots, \beta_p$ jako hodnoty, které minimalizují

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

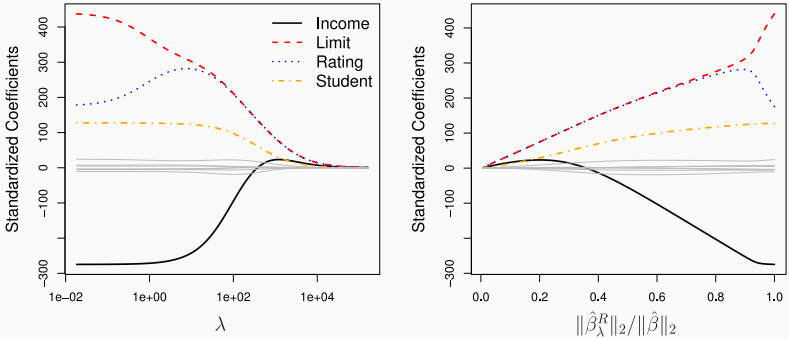
- Naproti tomu odhady koeficientů hřebenové regrese $\hat{\beta}^R$ jsou hodnoty, které minimalizují

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

kde $\lambda \geq 0$ je **ladicí parametr**, který je třeba stanovit odděleně.

- Jako u nejmenších čtverců hledá hřebenová regrese odhady koeficientů, které prokládají data dobře, a to tak, že dělá RSS malé.
- Nicméně druhý člen, $\lambda \sum_j \beta_j^2$, nazývaný **smršťovací penalta** je malý, jsou-li $\beta_1, \ldots, \beta_p$ blízké nule, a tak má efekt smršťování odhadů β_j směrem k nule.
- Ladicí parametr λ slouží k řízení relativního vlivu těchto dvou členů na odhady regresních koeficientů.
- Volba dobré hodnoty λ je kritická; používá se k tomu křížová validace.

Příklad s kreditními daty



Podrobnosti k předchozímu obrázku

- Na levém panelu odpovídá každá křivka odhadu koeficientu hřebenové regrese pro jednu z deseti proměnných, znázorněnému jako funkce λ.
- Pravý panel zobrazuje stejné odhady hřebenových koeficientů jako levý panel, ale místo toho, abychom na ose x vynášeli λ, vynášíme tam nyní $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$, kde $\hat{\beta}$ označuje vektor odhadů koeficientů nejmenších čtverců.
- Označení $\|\beta\|_2$ znamená ℓ_2 normu vektoru (vyslovuje se to "el dva"), která je definována jako $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$.

- Standardní odhady koeficientů metody nejmenších čtverců jsou **měřítkově invariantní:** vynásobení X_j konstantou c vede prostě k přeškálování odhadů koeficientů nejmenších čtverců faktorem $1/c$. Jinými slovy, bez ohledu na to, jak je škálován j-tý prediktor, zůstane $\hat{\beta}_j X_j$ beze změny.
- Naproti tomu se odhady koeficientů hřebenové regrese mohou **podstatně** změnit, vynásobíme-li daný prediktor konstantou, a to díky členu se součtem druhých mocnin koeficient v penalizační části cílové funkce hřebenové regrese.
- Je tudíž nejlepší používat hřebenovou regresi po **normalizaci prediktorů** pomocí vzorce

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

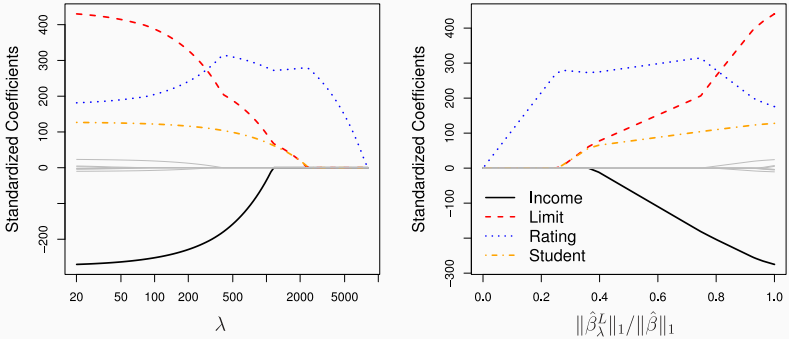
- Hřebenová regrese má jednu zřejmou nevýhodu: na rozdíl od výběru podmnožiny, který bude obecně vybírat modely zahrnující pouze nějakou podmnožinu proměnných, hřebenová regrese do konečného modelu zahrne všech p prediktorů.
- Metoda **Lasso** je poměrně nedávnou alternativou k hřebenové regresi, která tuto nevýhodu překonává. Koeficienty metody Lasso, $\hat{\beta}_\lambda^L$, minimalizují veličinu

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

- Ve statistickém žargonu používá metoda Lasso ℓ_1 (vyslovováno jako "el jedna") penaltu namísto ℓ_2 penalty. Přitom ℓ_1 norma vektoru koeficientů β je dána vztahem $\|\beta\|_1 = \sum|\beta_j|$.

- Stejně jako hřebenová regrese smršťuje metoda Lasso odhady koeficientů směrem k nule.
- Avšak v případě metody Lasso má ℓ_1 penalta ten efekt, že nutí některé z odhadů koeficientů, aby byly přesně nulové, je-li ladicí parametr λ dostatečně velký.
- Tudíž velmi podobně jako při výběru nejlepší podmnožiny provádí metoda Lasso **výběr proměnných**.
- Říkáme, že Lasso nám dává **řídké** modely — to jest modely, které zahrnují pouze nějakou podmnožinu proměnných.
- Stejně jako u hřebenové regrese je volba dobré hodnoty λ pro metodu Lasso kritická; metodou volby je opět křížová validace.

Příklad: soubor kreditních dat



Závěry

- Hřebenová regrese ani metoda Lasso nepřevládnou univerzálně jedna nad druhou.
- Obecně se dá předpokládat, že Lasso bude pracovat lépe, bude-li odpověď funkcí pouze poměrně malého počtu prediktorů.
- Avšak počet prediktorů, které mají vliv na odpověď, není u reálných souborů dat nikdy znám **a priori**.
- K rozhodnutí o tom, který přístup je pro daný soubor dat lepší, se dá použít některá technika typu křížové validace.

- Metody, které jsme v této kapitole dosud probírali, spočívaly v prokládání lineárních regresních modelů, pomocí nejmenších čtverců nebo přístupu se smršťováním, při použití původních prediktorů $X_1, X_2, \ldots, X_p$.
- Budeme se nyní zabývat skupinou přístupů, které **transformují** prediktory a pak nejmenšími čtverci prokládají model užívající transformované proměnné. Tyto postupy budeme nazývat metodami **dimenzionální redukce**.

- Nechť $Z_1, Z_2, \ldots, Z_M$ představují $M < p$ **lineárních kombinací** našich původních p prediktorů. To jest

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j$$

pro nějaké konstanty $\phi_{m1}, \ldots, \phi_{mp}$.

- Prokládáme pak lineární regresní model,

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n,$$

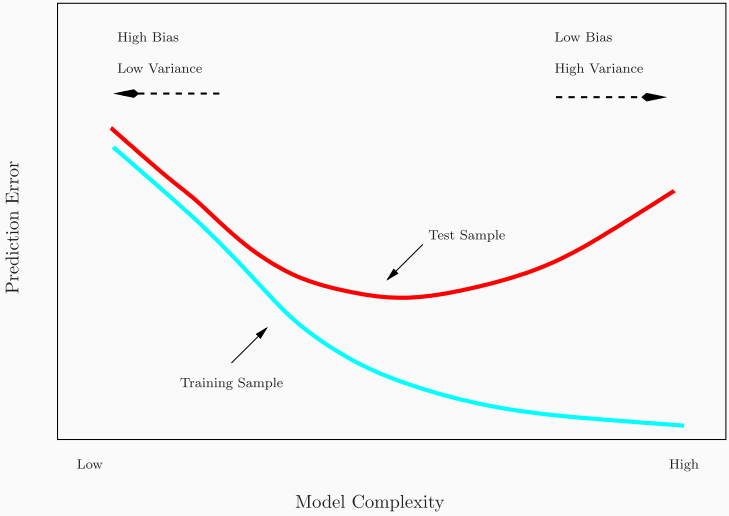
prostřednictvím obvyklých nejmenších čtverců.

- Poznamenáváme, že regresní koeficienty v modelu jsou nyní $\theta_0, \theta_1, \ldots, \theta_M$. Jsou-li konstanty $\phi_{m1}, \ldots, \phi_{mp}$ vybrány vhodně, pak postup dimenzionální redukce může často překonat regresi obvyklými nejmenšími čtverci.

Připomeňte si rozdíl mezi **trénovací chybou** a **testovací chybou**:

- **Testovací chyba** je průměrná chyba, která vzniká při použití metody statistického učení k predikci odpovědi na novém pozorování, takovém, které se nepoužilo při tréninku metody.
- Naproti tomu **trénovací chyba** se dá snadno vypočítat tak, že metodu statistického učení aplikujeme na pozorování použitá k jejímu tréninku.
- Ale míra trénovací chyby je často zcela odlišná od míry testovací chyby a především může ta první z nich **dramaticky podhodnocovat** tu druhou.

Efektivita: trénovací soubor versus testovací soubor



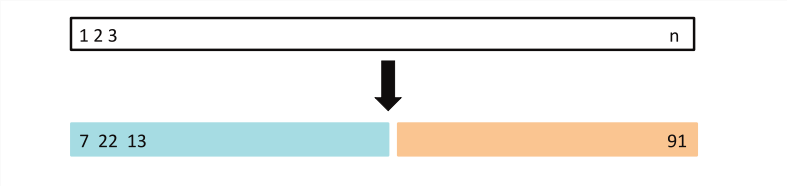
Více o odhadech chyby předpovědi

- Nejlepší řešení: velký k tomu určený soubor dat. Často není k dispozici.
- Některé metody provádějí **matematickou úpravu** míry trénovací chyby s cílem odhadnout míru testovací chyby. Patří k nim **Cp statistika**, **AIC** a **BIC**.
- Zabývejme se nyní třídou metod, které odhadují testovací chybu tak, že **odloží stranou** z procesu prokládání podmnožinu trénovacích pozorování a pak použijí metodu statistického učení na predikci těchto odložených pozorování.

Přístup používající validační soubor

- Zde náhodně rozdělíme dostupný soubor vzorků na dvě části: **tréninkovou sadu** a **validační** neboli **odloženou sadu**.
- Model se proloží na tréninkové sadě a proložený model se použije k předpovědi odpovědí na pozorování ve validační sadě.
- Výsledná chyba na validační sadě nám dává odhad testovací chyby. Ta se obvykle posuzuje pomocí MSE v případě kvantitativní odpovědí a pomocí míry chybné klasifikace v případě kvalitativní (diskrétní) odpovědi.

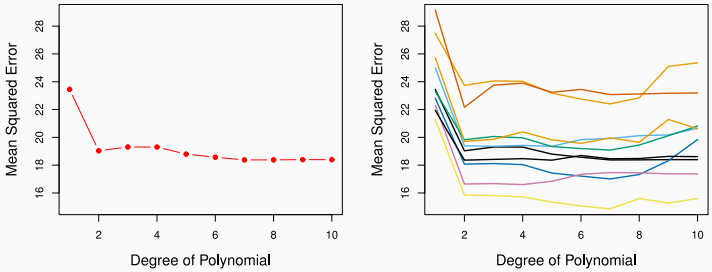
Proces validace



Náhodné rozdělení na dvě poloviny: levá část je trénovací sada, pravá část je validační sada.

Příklad: údaje o automobilech

- Chceme porovnat lineární členy s členy vyšších řádů v polynomech užitých v lineární regresi.
- Náhodně rozdělíme 392 pozorování na dvě sady, trénovací sadu se 196 datovými body a validační sadu obsahující zbylých 196 pozorování.



Levý panel ukazuje jediné rozdělení; pravý panel ukazuje více různých rozdělení.

Nevýhody přístupu s validačním souborem

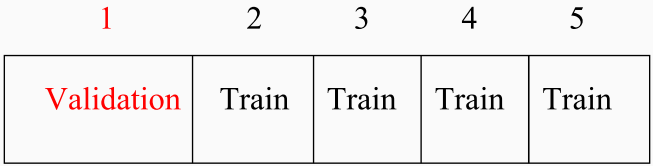
- Validační odhad testovací chyby může být vysoce proměnlivý, závisí totiž přesně na tom, která pozorování se zahrnou do tréninkové sady a která jsou zahrnuta do validační sady.
- U validačního přístupu se k proložení modelu používá pouze podmnožina pozorování — ta, která jsou zahrnuta do tréninkové sady a ne ta ve validační sadě.
- To napovídá, že chyba na validační sadě může mít tendenci **nadhodnocovat** testovací chybu pro model proložený celým souborem dat. **Proč?**

K-násobná křížová validace

- **Široce používaný přístup** k odhadování testovací chyby.
- Odhady se dají použít k výběru nejlepšího modelu a k získání představy o testovací chybě finálního zvoleného modelu.
- Myšlenka zde je náhodně rozdělit data do K stejně velkých částí. Vynecháme část k, proložíme model zbylými $K - 1$ částmi (kombinovaně) a pak získáme předpovědi pro odloženou k-tou část.
- Toto se provádí po řadě pro každou část $k = 1, 2, \ldots, K$ a pak se výsledky zkombinují.

K-násobná křížová validace podrobně

Rozděl data do K zhruba stejně velkých částí (zde je $K = 5$).

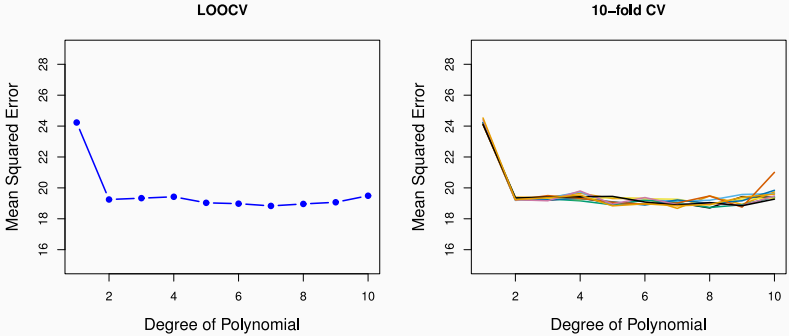


- Nechť těch K částí jsou $C_1, C_2, \ldots, C_K$, kde C_k označuje indexy pozorování v části k. V části k je n_k pozorování; pokud n je násobkem K, je $n_k = n/K$.
- Vypočítejte

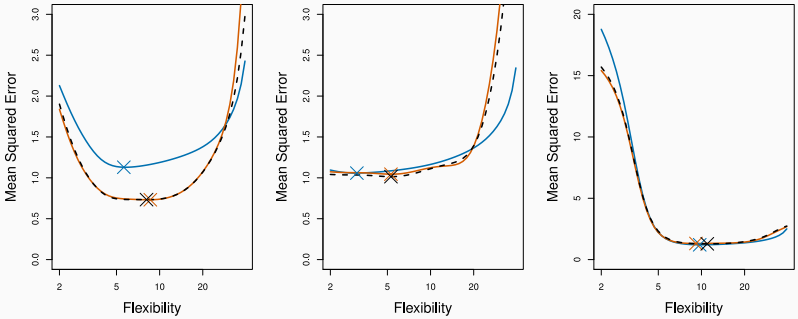
$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k,$$

  kde $\mathrm{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$ a $\hat{y}_i$ je aproximace pro pozorování i získaná z dat s odloženou částí k.
- Položíme-li $K = n$, je výsledkem n-násobná validace neboli **křížová validace s vynecháním jednoho** (LOOCV, leave-one out cross-validation).
- Lepší volba je ale $K = 5$ nebo 10 (průměr LOOCV má vysoký rozptyl kvůi korelaci odhadů).

Data o automobilech zkoumaná znovu



Skutečná a odhadnutá testovací MSE pro simulovaná data



Další problémy s křížovou validací

- Jelikož každá trénovací sada je pouze $(K-1)/K$-krát tak velká jako původní trénovací sada, odhady chyby předpovědi budou typicky zkresleny směrem nahoru. **Proč?**
- Toto zkreslení se minimalizuje při $K = n$ (LOOCV), ale tento odhad má vysoký rozptyl, jek jsme uvedli dříve.
- $K = 5$ nebo 10 dává dobrý kompromis pro vyvážení tohoto vztahu zkreslení a rozptylu.

Křížová validace: správně a chybně

- Uvažujme jednoduchý klasifikátor aplikovaný na nějaká dvoutřídní data:
  1. Začínáme s 5000 prediktory a 50 vzorky a najdeme těch 100 prediktorů, které mají největší korelaci se štítky tříd.
  2. Pak použijeme nějaký klasifikátor, jako je třeba logistická regrese, pouze na těchto 100 prediktorů.

  Jak odhadneme účinnost tohoto klasifikátoru na testovacím souboru?

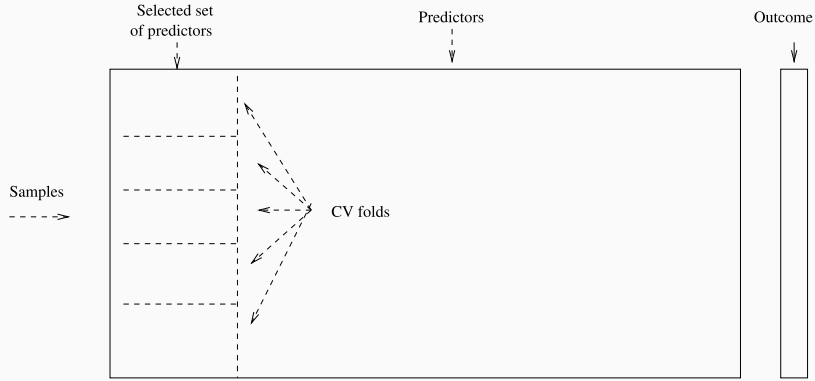
  Můžeme použít křížovou validaci v kroku 2 a zapomenout na krok 1?

- To by ignorovalo skutečnost, že naše procedura v kroku 1 **již viděla štítky trénovacích dat** a zužitkovala je. To je forma tréninku a musí to být do validačního procesu zahrnuto.
- Je snadné nasimulovat realistická data se štítky tříd nezávisejícími na výstupu, takže skutečná testovací chyba je 50 %, ale odhad chyby křížovou validací, který ignoruje krok 1, bude nula! **Zkuste to udělat sami.**
- Viděli jsme tuto chybu dělat v mnoha článcích z genomiky ve vysoce profilovaných časopisech.

Chybný a správný způsob

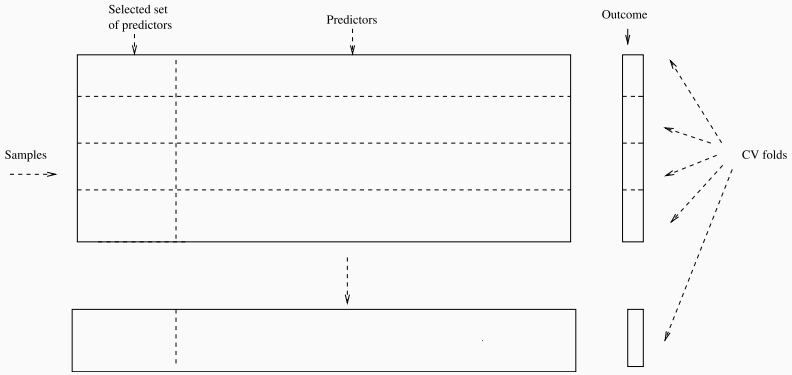
- **Chybně:** Použijeme křížovou validaci v kroku 2.
- **Správně:** Použijeme křížovou validaci v krocích 1 a 2.

Chybný způsob



Texty: Vzorky, Vybraná sada prediktorů, Složky křížové validace, Prediktory, Výsledky

Správný způsob



Texty: Vzorky, Vybraná sada prediktorů, Prediktory, Výsledky, Složky křížové validace

Lectue Contents

Nesupervizované versus supervizované učení:

- Většina tohoto kurzu je zaměřena na metody učení s učitelem (*supervizovaného učení*), jako je regrese a klasifikace.
- V takové situaci pozorujeme jak soubor vlastností $X_1, X_2, \ldots, X_p$ každého objektu, tak rovněž odpověď nebo odezvu Y. Cílem pak je předpovídat Y pomocí $X_1, X_2, \ldots, X_p$.
- Zde se místo toho soustředíme na *nesupervizované učení* (učení bez učitele), kde pozorujeme pouze vlastnosti $X_1, X_2, \ldots, X_p$. Nezajímá nás předpovídání, protože nemáme přidruženou proměnnou odpovědi Y.

Cíle učení bez učitele

- Cílem je objevit zajímavé věci o měřeních: existuje nějaký informativní způsob, jak vizualizovat daná data? Můžeme mezi proměnnými nebo mezi pozorováními odhalit nějaké podskupiny?
- Připomeneme si dvě metody:
  - *analýzu hlavních komponent*, nástroj používaný pro vizualizaci dat nebo předběžné zpracování dat před tím, než použijeme supervizované postupy, a
  - *shlukování*, širokou třídu metod k objevování neznámých podskupin v datech.

Obtížný úkol učení bez učitele

- Učení bez učitele je subjektivnější než učení s učitelem, protože zde analýza nemá jednoduchý cíl jako je předpověď odpovědi.
- Ale techniky učení bez učitele mají rostoucí význam v řadě oborů:
  - podskupiny pacientek s rakovinou prsu seskupené na základě měření jejich genové exprese,
  - skupiny kupujících charakterizované historií jejich prohlížení zboží a nákupů,
  - filmy seskupené podle hodnocení uděleného jejich diváky.

Další výhoda

- Je často snazší získat *neoznačená data* — z laboratorního přístroje nebo počítače — než *označená data*, která mohou vyžadovat lidský zásah.
- Tak je například obtížné automaticky posoudit celkové vyznění recenze filmu: je příznivá nebo ne?

- Analýza hlavních komponent (PCA, Principal Component Analysis) poskytuje nízkorozměrnou reprezentaci souboru dat. Stanovuje posloupnost lineárních kombinací proměnných, které mají maximální rozptyl a jsou navzájem nekorelovány.
- Kromě toho, že poskytuje odvozené proměnné k použití v úlohách supervizovaného učení, slouží PCA také jako nástroj pro vizualizaci dat.

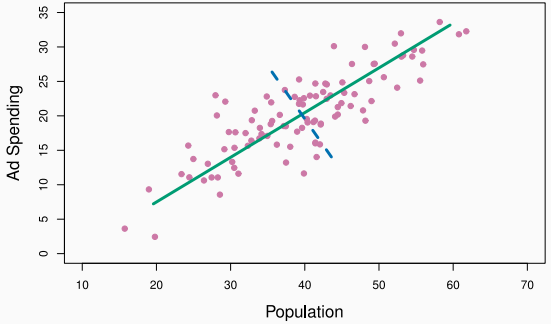
Analýza hlavních komponent: podrobnosti

- *První hlavní komponenta* souboru vlastností $X_1, X_2, \ldots, X_p$ je normalizovaná lineární kombinace těchto vlastností

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p,$$

  která má největší rozptyl. Slovem *normalizovaná* rozumíme, že $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.
- Na prvky $\phi_{11}, \ldots, \phi_{p1}$ odkazujeme jako na **zátěže** první hlavní komponenty; dohromady zátěže vytvářejí **vektor zátěže** dané hlavní komponenty, $\phi_1 = (\phi_{11}\ \phi_{21}\ \ldots\ \phi_{p1})^T$.
- Omezujeme zátěže tak, že jejich součet druhých mocnin je roven jedné, neboť pokud bychom jinak připustili, aby tyto prvky byly v absolutní hodnotě libovolně velké, mohlo by to vést k libovolně velkému rozptylu.

PCA: příklad



Velikost populace (pop) a náklady na reklamu (ad) pro 100 různých měst jsou zobrazeny jako purpurové kroužky. Zelená plná přímka označuje směr první hlavní komponenty, modrá čárkovaná přímka označuje směr druhé hlavní komponenty.

- Vektor zátěže ϕ_1 se složkami $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ definuje v prostoru vlastností směr, ve kterém se data nejvíce mění.
- Jestliže promítneme n bodů dat $x_1, \ldots, x_n$ na tento směr, jsou hodnoty projekcí samotná skóre $z_{11}, \ldots, z_{n1}$ hlavní komponenty.

Další hlavní komponenty

- Druhá hlavní komponenta je lineární kombinace $X_1, \ldots, X_p$, která má maximální rozptyl mezi všemi lineárními kombinacemi, jež jsou *nekorelované* se Z_1.
- Skóre $z_{12}, z_{22}, \ldots, z_{n2}$ druhé hlavní komponenty mají tvar

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \cdots + \phi_{p2} x_{ip},$$

kde ϕ_2 je vektor zátěže druhé hlavní komponenty o složkách $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$.

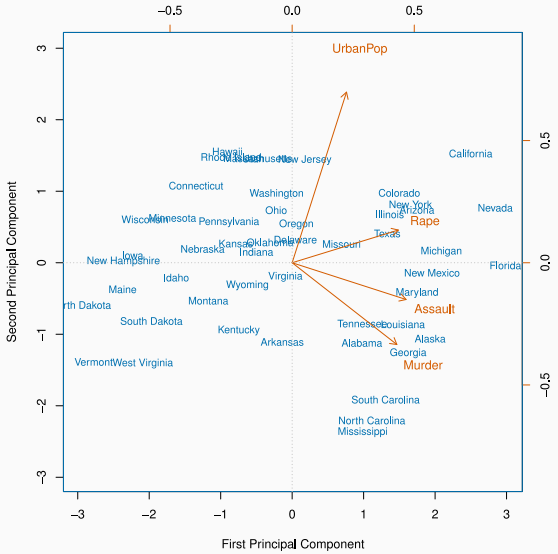
Další hlavní komponenty: pokračování

- Ukazuje se, že omezení na Z_2, aby bylo nekorelované se Z_1, je ekvivalentní omezení směru ϕ_2 tak, aby byl ortogonální (kolmý) ke směru ϕ_1. A tak dále.
- Směry hlavních komponent $\phi_1, \phi_2, \phi_3, \ldots$ jsou uspořádaná posloupnost pravých singulárních vektorů matice X a rozptyly složek jsou $\frac{1}{n}$ násobky kvadrátů singulárních čísel. Je zde nejvýše $\min(n-1, p)$ hlavních komponent.

Ilustrace

- Data `USAarrests`: Pro každý z padesáti států ve Spojených státech soubor dat obsahuje počet zatčení na 100 000 obyvatel za každý ze tří zločinů: `Assault`, `Murder` a `Rape` (přepadení, vražda a znásilnění). Zaznamenáváme rovněž hodnoty `UrbanPop` (procento obyvatel každého státu žijících v městských oblastech).
- Vektory skóre hlavních komponent mají délku $n = 50$ a vektory zátěže hlavních komponent mají délku $p = 4$.
- PCA byla provedena po normalizaci každé proměnné tak, aby měla střední hodnotu nula a směrodatnou odchylku jedna.

Data USAarrests: graf PCA



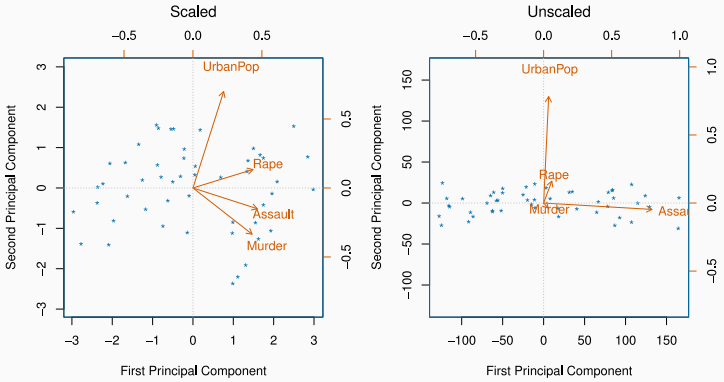
Podrobnosti k obrázku

První dvě hlavní komponenty pro data USAarrests.

- Modré názvy států reprezentují skóre pro první dvě hlavní komponenty.
- Oranžové šipky označují vektory zátěže prvních dvou hlavních komponent (s osami souřadnic nahoře a vpravo). Tak například zátěž pro `Rape` je u první hlavní komponenty 0,54 a u druhé hlavní komponenty je to 0,17 (slovo `Rape` je centrováno kolem bodu $(0{,}54, 0{,}17)$.
- Tento graf je známý jako *biplot*, protože zobrazuje jak skóre, tak zátěže hlavních komponent.

Škálování proměnných hraje roli

- Jsou-li proměnné v odlišných jednotkách, doporučuje se přeškálovat je tak, aby každá měla směrodatnou odchylku rovnu jedné.
- Pokud jsou proměnné ve stejných jednotkách, můžete je škálovat nebo ne.



- *Shlukování* představuje velmi širokou třídu technik pro nalézání *podskupin* neboli *shluků* (také klastrů) v souboru dat.
- Hledáme rozdělení dat do rozdílných skupin takových, že pozorování uvnitř každé skupiny jsou si navzájem zcela podobná.
- Abychom to učinili konkrétním, musíme definovat, co pro dvě nebo více pozorování znamená, že jsou *podobná* nebo *odlišná*.
- Toto je ovšem často úvaha, která je specifická pro daný obor a musí se dělat na základě znalostí o studovaných datech.

PCA versus shlukování

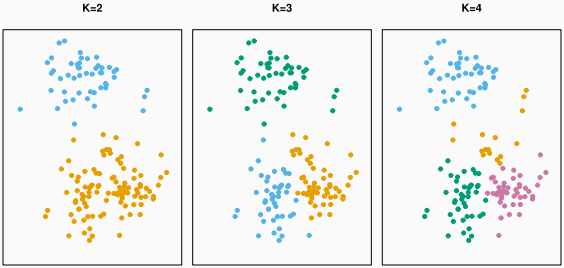
- PCA vyhledává nízkorozměrnou reprezentaci daných pozorování, která vysvětluje značný podíl rozptylu.
- Shlukování vyhledává mezi danými pozorováními homogenní podskupiny.

- Předpokládejme, že máme přístup k velkému počtu měření (např. medián příjmu domácnosti, zaměstnání, vzdálenost od nejbližší městské oblasti atd.) pro velký počet osob.
- Naším cílem je provést *segmentaci trhu* tím, že stanovíme podskupiny osob, které by mohly být vnímavější k určitým způsobům reklamy nebo které by pravděpodobněji zakoupily konkrétní produkt.
- Úloha provést segmentaci trhu vede ke shlukování osob v daném souboru dat.

Dvě metody shlukování

- V *metodě K-průměrů* se snažíme rozdělit pozorování do předem specifikovaného počtu shluků.
- U *hierarchického shlukování* nevíme předem, kolik shluků chceme; ve skutečnosti končíme se stromovitou vizuální reprezentací daných pozorování, tak zvaným *dendrogramem*, který nám umožňuje vidět najednou shluky získané pro každý jejich možný počet, od 1 do n.

Shlukování metodou K-průměrů



Simulovaný soubor dat se 150 pozorováními ve dvourozměrném prostoru. Panely ukazují výsledky použití metody K-průměrů s různými hodnotami K, počtu shluků. Barva každého pozorování označuje shluk, k němuž bylo přiřazeno algoritmem metody K-průměrů. Poznamenáváme, že zde není žádné uspořádání shluků, takže obarvení shluků je libovolné. Tyto štítky shluků nebyly při shlukování použity; jsou to spíše výstupy procedury shlukování.

- Myšlenka za metodou K-průměrů je, že *dobré* shlukování je to, pro něž je *vnitroshluková variabilita* WCV co nejmenší.
- Vnitroshluková variabilita pro shluk C_k je $\mathrm{WCV}(C_k)$. Udává míru, kterou se pozorování uvnitř shluku navzájem liší.
- Chceme tudíž řešit úlohu

$$\underset{C_1,\ldots,C_K}{\text{minimalizuj}} \left\{ \sum_{k=1}^{K} \mathrm{WCV}(C_k) \right\}.$$

- Ve slovech tento vzorec říká, že chceme rozdělit pozorování do K shluků tak, aby celková vnitroshluková variabilita, sečtená přes všech K shluků, byla co nejmenší.

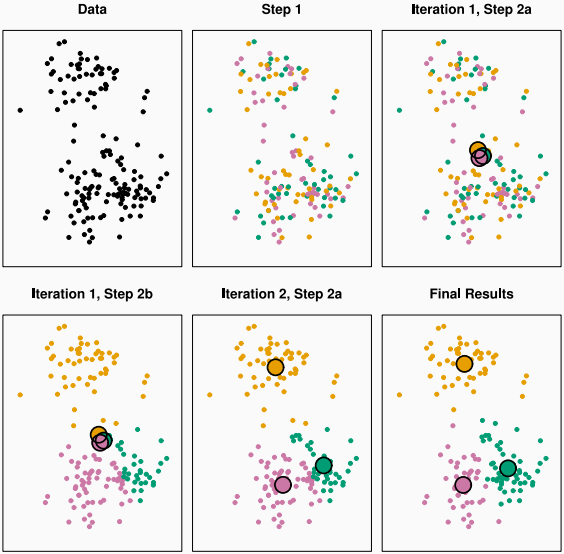
Algoritmus shlukování metodou K-průměrů

1. Každému z pozorování náhodně přiřaď číslo od 1 do K. Tato čísla slouží jako počáteční přiřazení pozorování do shluků.
2. Iteruj do té doby, až se přiřazení do shluků přestane měnit:
   - 2a. Pro každý z K shluků vypočítej *centroid* shluku. Centroid k-tého shluku je vektor p středních hodnot vlastností pro pozorování v k-tém shluku.
   - 2b. Přiřaď každé pozorování do shluku, jehož centroid je nejblíže (kde *nejblíže* je definováno pomocí euklidovské vzdálenosti).

Tento algoritmus zaručeně v každém kroku snižuje hodnotu cílové funkce.

Není však zaručeno, že to bude dávat globální minimum. *Proč ne?*

Příklad


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Podrobnosti k předchozímu obrázku

Postup algoritmu K-průměrů s $K = 3$:

- *Nahoře vlevo:* Znázorněna jsou pozorování.
- *Nahoře střed:* V Kroku 1 algoritmu se každé pozorování náhodně přiřadí některému shluku.
- *Nahoře vpravo:* V Kroku 2a se vypočítají centroidy shluků. Ty jsou znázorněny jako velké obarvené disky. Na začátku jsou centroidy umístěny téměř úplně přes sebe, protože počáteční přiřazení do shluků bylo provedeno náhodně.
- *Dole vlevo:* V Kroku 2b se každé pozorování přiřadí k nejbližšímu centroidu.
- *Dole střed:* Ještě jednou se provede Krok 2a, což vede k novým centroidům shluků,
- *Dole vpravo*: Výsledky získané po 10 iteracích.

Příklad: rozdílné startovací hodnoty

Shlukování metodou K-průměrů provedené šestkrát na datech z předchozího obrázku s $K = 3$. pokaždé s rozdílným náhodným přiřazením pozorování v Kroku 1 algoritmu metody.

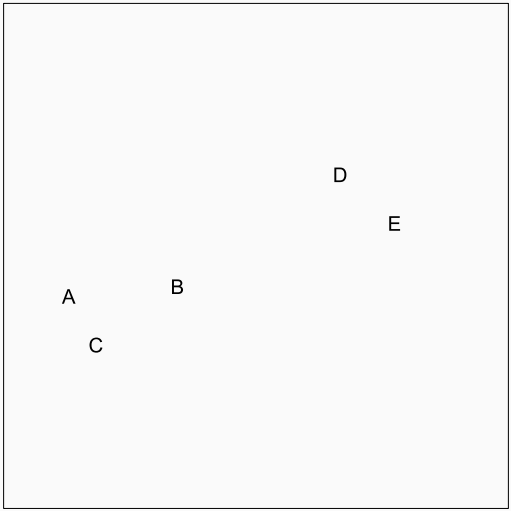
Nad každým grafem je hodnota cílové funkce (4).

Byla získána tři rozdílná lokální optima, z nichž jedno vedlo k menší hodnotě cílové funkce a poskytuje lepší oddělení shluků.

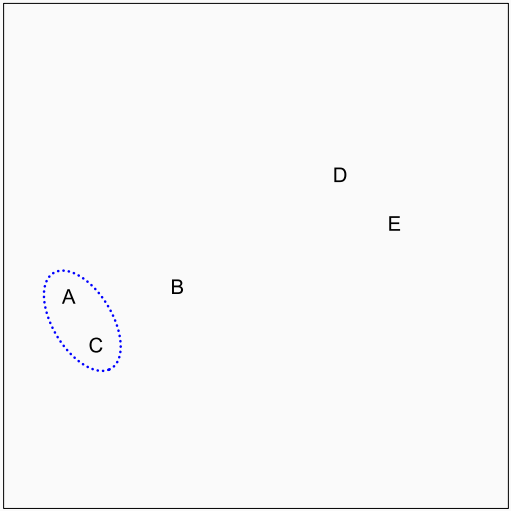
Příklady označené červeně všechny dosáhly téhož nejlepšího řešení s hodnotou cílové funkce 235,8.

- Shlukování metodou K-průměrů od nás vyžaduje, abychom předem specifikovali počet shluků K. To může znamenat nevýhodu.
- *Hierarchické shlukování* je alternativní přístup, který nevyžaduje, abychom se vázali na konkrétní volbu K.
- V bakalářském studiu jsme si popsali shlukování *zdola nahoru* neboli *aglomerativní*. Je to nejběžnější typ hierarchického shlukování a název odkazuje na skutečnost, že se buduje dendrogram počínaje listy a shluky se kombinují směrem ke kmeni.

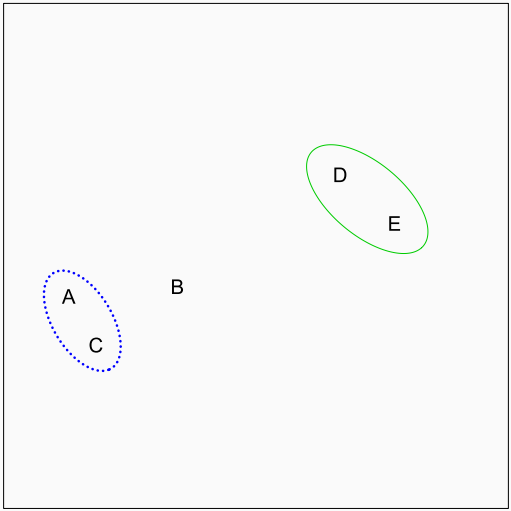
Tvoří se hierarchie způsobem "zdola nahoru" ...



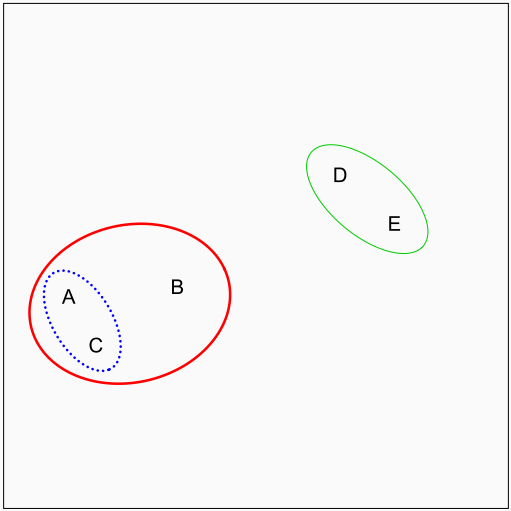
Tvoří se hierarchie způsobem "zdola nahoru" ...



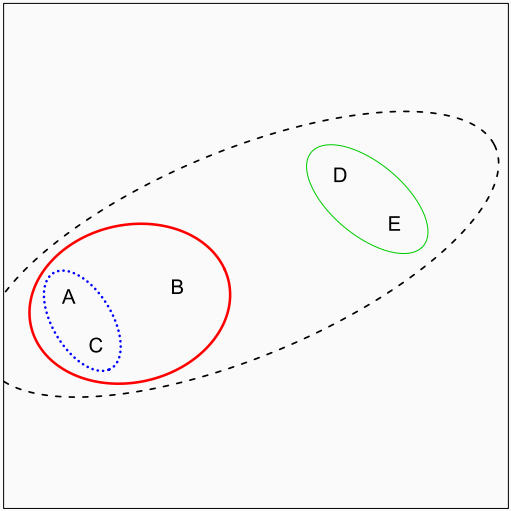
Tvoří se hierarchie způsobem "zdola nahoru" ...



Tvoří se hierarchie způsobem "zdola nahoru" ...

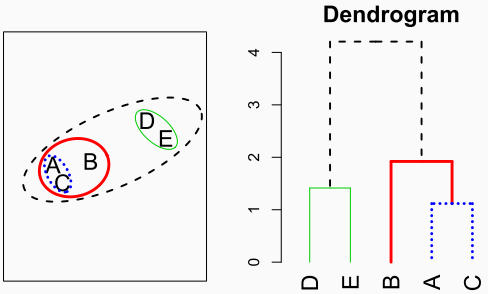


Tvoří se hierarchie způsobem "zdola nahoru" ...

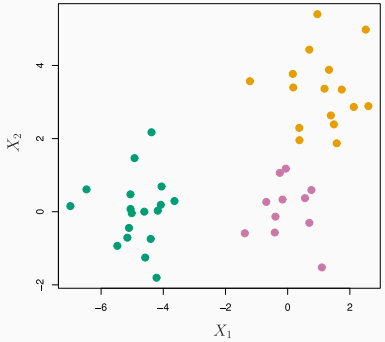


Postup slovy:

- Začni s každým bodem jako vlastním shlukem.
- Urči nejbližší dva shluky a spoj je v jeden.
- Opakuj.
- Konči, když jsou všechny body v jediném shluku.

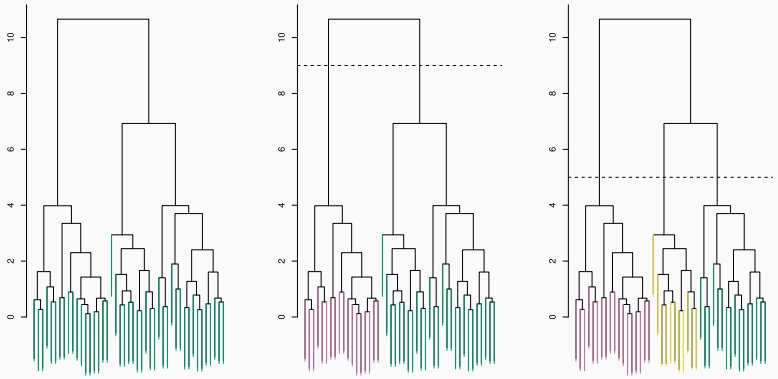


Příklad



45 pozorování vygenerovaných ve dvourozměrném prostoru. Ve skutečnosti jsou zde tři rozličné třídy, označené různými barvami. My však budeme považovat tyto štítky tříd za neznámé a budeme se snažit ta pozorování shluknout, abychom z dat ty třídy odhalili.

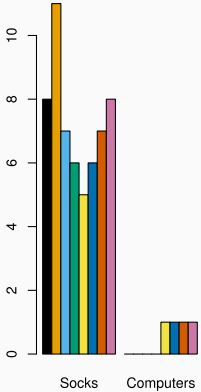
Aplikace hierarchického shlukování

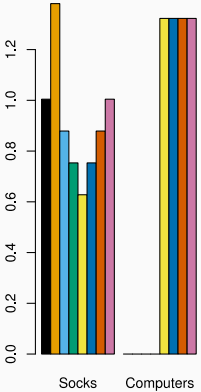


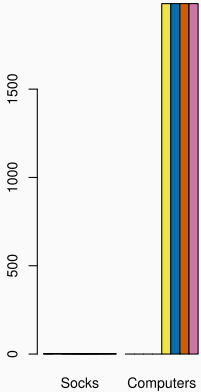
Podrobnosti k předchozímu obrázku

- *Vlevo:* Dendrogram získaný hierarchickým shlukováním dat z předchozího obrázku metodou nejvzdálenějšího souseda a užitím euklidovské vzdálenosti.
- *Střed:* Dendrogram z levého panelu uříznutý ve výšce 9 (označeno čárkovanou přímkou). Tento řez vede na dva rozličné shluky označené rozdílnými barvami.
- *Vpravo:* Dendrogram z levého panelu, tentokrát uříznutý ve výšce 5. Tento řez vede na tři rozličné shluky označené rozdílnými barvami. Poznamenáváme, že ty barvy se nepoužívaly při shlukování, jsou zde prostě použity pro zobrazovací účely.

Škálování proměnných hraje roli







Praktické otázky

- Měla by být pozorování nebo vlastnosti nejprve nějakým způsobem normalizovány? Například by proměnné možná měly být nejdříve vycentrovány tak, aby měly střední hodnotu nula, a přeškálovány tak, aby měly směrodatnou odchylku rovnou jedné.
- V případě hierarchického shlukování:
  - Jaká míra rozdílnosti by měla být použita?
  - Jaký typ vazby by se měl použít?
- Kolik shluků zvolit (jak v metodě K-průměrů, tak při hierarchickém shlukování)? Obtížný problém. Není shoda na metodě. Viz Elements of Statistical Learning, kap. 13 pro více podrobností.

- *Učení bez učitele* je důležité pro pochopení variability a struktury seskupování u neoznačených souborů dat a může být užitečným předběžným procesem pro učení s učitelem.
- Je vnitřně obtížnější než *učení s učitelem,* protože zde není žádný zlatý standard (jako nějaká výstupní proměnná) a žádný jednotlivý cíl (jako přesnost na testovacím souboru).
- Je to aktivní oblast výzkumu s mnoha nástroji vyvinutými v poslední době, jako jsou *samoorganizační mapy, analýza nezávislých komponent* a *spektrální shlukování.*
  Viz *The Elements of Statistical Learning,* kapitola 14.