Klasifikace

Matematické metody pro ITS (11MAMY)

Jan Přikryl

s využitím díla G. James et al., *An Introduction to Statistical Learning*

9. přednáška 11MAMY

čtvrtek 7. dubna 2022

verze: 2022-04-07 14:08

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

Klasifikace

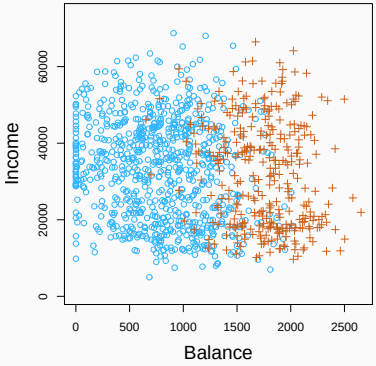
- Kvalitativní proměnné nabývají hodnot z neuspořádané množiny $\mathcal{C}$, například

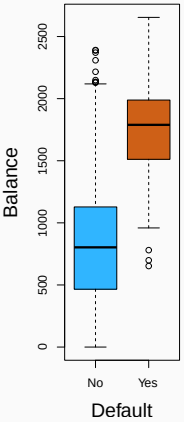
$$\texttt{eye color} \in \{\texttt{brown}, \texttt{blue}, \texttt{green}\}$$
$$\texttt{email} \in \{\texttt{spam}, \texttt{ham}\}.$$

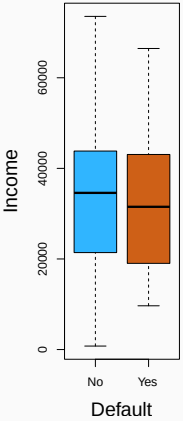
- Pro daný vektor charakteristik X a kvalitativní odpověď Y nabývající hodnot z množiny $\mathcal{C}$ spočívá klasifikační úloha ve vytváření funkce $C(X)$, která jako vstup bere vektor charakteristik X a předpovídá hodnotu Y, tj. $C(X) \in \mathcal{C}$.
- Často nás spíše zajímají odhady **pravděpodobností**,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

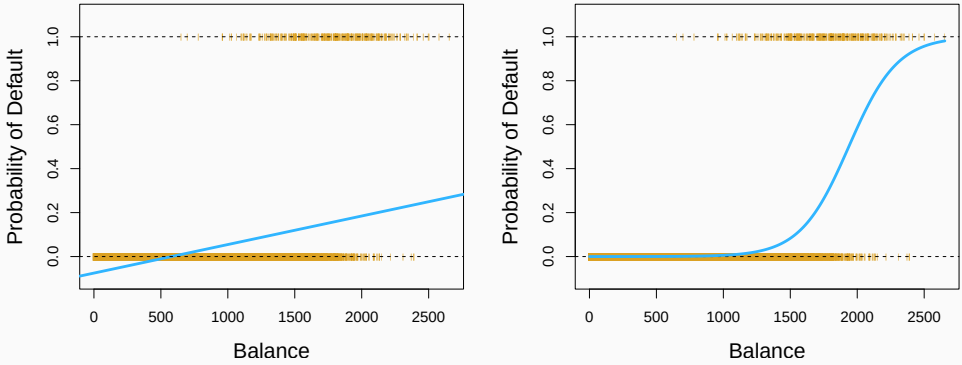
Předpokládejme, že pro klasifikační úlohu `Default` kódujeme

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

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 $E[Y|X = x] = \mathrm{P}(Y = 1|X = x)$, 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.

Lineární regrese — pokračování

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**.

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

Logistická regrese používá výraz

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

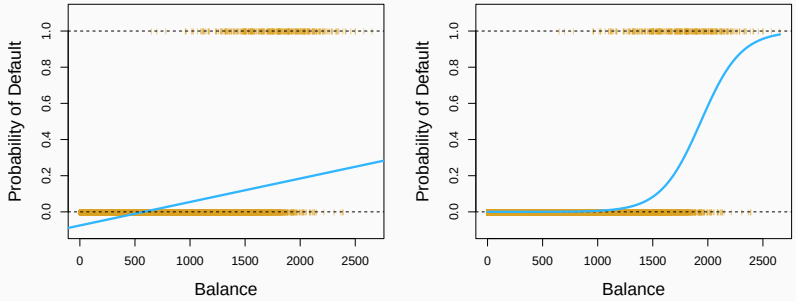
(e $\approx 2{,}71828$ je matematická konstanta — Eulerovo číslo). Je snadné vidět, že bez ohledu na hodnoty β_0, β_1 nebo X bude $p(X)$ nabývat hodnot mezi 0 a 1.

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.

Maximální věrohodnost

K odhadu parametrů používáme maximální věrohodnost:

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

Tato **věrohodnost** udává pravděpodobnost pozorovaných nul a jedniček v datech. Hodnoty β_0 a β_1 vybíráme tak, abychom maximalizovali věrohodnost pozorovaných dat.

Většina statistických balíčků umí proložit lineární logistické regresní modely pomocí maximální věrohodnosti. V R používáme funkci `glm`, v Matlabu obdobné `fitglm`.

	Koef.	SE	z-statistika	p-hodnota
Regr. konst.	-10,6513	0,3612	-29,5	$< 0{,}0001$
`balance`	0,0055	0,0002	24,9	$< 0{,}0001$

Tvorba předpovědí

Jaké je naše odhadovaná pravděpodobnost nesplácení (`default`) při zůstatku (`balance`) $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10{,}6513+0{,}0055\times 1000}}{1+e^{-10{,}6513+0{,}0055\times 1000}} = 0{,}006$$

Při zůstatku $2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10{,}6513+0{,}0055\times 2000}}{1+e^{-10{,}6513+0{,}0055\times 2000}} = 0{,}586$$

Tvorba předpovědí ještě jednou

Udělejme to znovu a použijme `student` jako prediktor:

	Koef.	SE	z-statistika	p-hodnota
Regr. konst.	-3,5041	0,0707	-49,55	$< 0{,}0001$
`student [Yes]`	0,4049	0,1150	3,52	0,0004

$$\hat{\mathrm{P}}(\texttt{default} = \texttt{Yes} | \texttt{student} = \texttt{Yes}) = \frac{\mathrm{e}^{-3{,}5041+0{,}4049\times 1}}{1+\mathrm{e}^{-3{,}5041+0{,}4049\times 1}} = 0{,}0431,$$

$$\hat{\mathrm{P}}(\texttt{default} = \texttt{Yes} | \texttt{student} = \texttt{No}) = \frac{\mathrm{e}^{-3{,}5041+0{,}4049\times 0}}{1+\mathrm{e}^{-3{,}5041+0{,}4049\times 0}} = 0{,}0292.$$

Logistická regrese s několika proměnnými

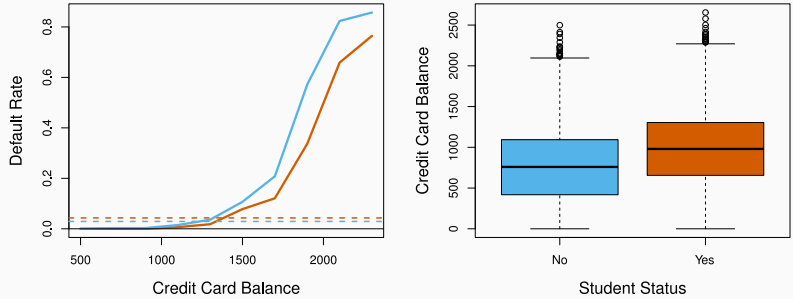
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Koef.	SE	z-statistika	p-hodnota
Regr. konst.	-10,8690	0,4923	-22,08	$< 0{,}0001$
`balance`	0,0057	0,0002	24,74	$< 0{,}0001$
`income`	0,0030	0,0082	0,37	0,7115
`student [Yes]`	-0,6468	0,2362	-2,74	0,0062

Proč je koeficient u proměnné `student` záporný, když předtím byl kladný?

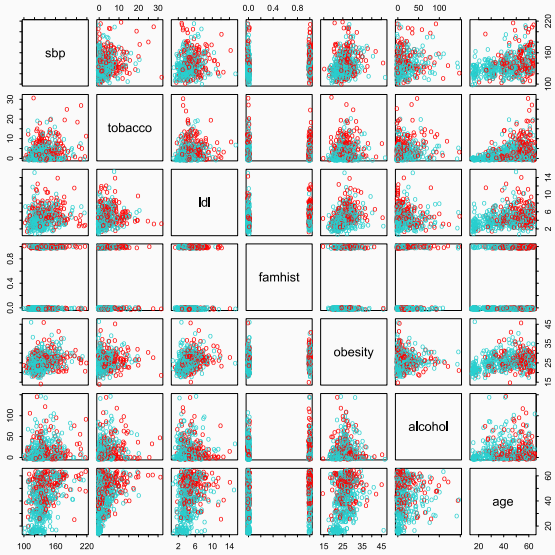
Protiklady



- Studenti mají tendenci mít vyšší zůstatky než nestudenti, takže jejich marginální míra nesplácení je vyšší než u nestudentů.
- Ale pro každou úroveň zůstatku studenti nesplácejí méně často než nestudenti.
- Toto lze odhalit logistickou regresí s více proměnnými.

Příklad: Srdeční choroby v Jižní Africe

- 160 případů IM (infarktu myokardu) a 302 kontrolních osob (všechno muži ve věku 15–64 let), z oblasti Western Cape, Jižní Afrika, počátkem 80. let.
- Celkový výskyt choroby v této oblasti je velmi vysoký: 5,1%.
- Měření na sedmi prediktorech (rizikových faktorech), výsledky zobrazeny v matici bodových grafů.
- Cílem je identifikovat relativní síly a směrování rizikových faktorů.
- Jde tu o část intervenční studie s cílem vést veřejnost ke zdravějšímu stravování.



Matice bodových grafů pro data *srdečních chorob v Jižní Africe*. Odpověď je kódována barevně — případy choroby (IM) jsou červené, kontrolní osoby tyrkysové, `famhist` je binární proměnná, u níž 1 označuje rodinnou historii IM.

```
> heartfit<-glm(chd~., data=heart, family=binomial)
> summary (heartfit)
Call:
  glm(formula=chd~., family=binomial, data=heart)
Coefficients:
            Estimate   Std.Error z value Pr(>|z|)
(Intercept) -4.1295997 0.9641558 -4.283  1.84e-05 ***
sbp          0.0057607 0.0056326  1.023  0.30643
tobacco      0.0795256 0.0262150  3.034  0.00242  **
ldl          0.1847793 0.0574115  3.219  0.00129  **
famhist[Yes] 0.9391855 0.2248691  4.177  2.96e-05 ***
obesity     -0.0345434 0.0291053 -1.187  0.23529
alcohol      0.0006065 0.0044550  0.136  0.89171
age          0.0425412 0.0101749  4.181  2.90e-05 ***
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 596.11 on 461 degrees of freedom
Residual deviance: 483.17 on 454 degrees of freedom
AIC: 499.17
```

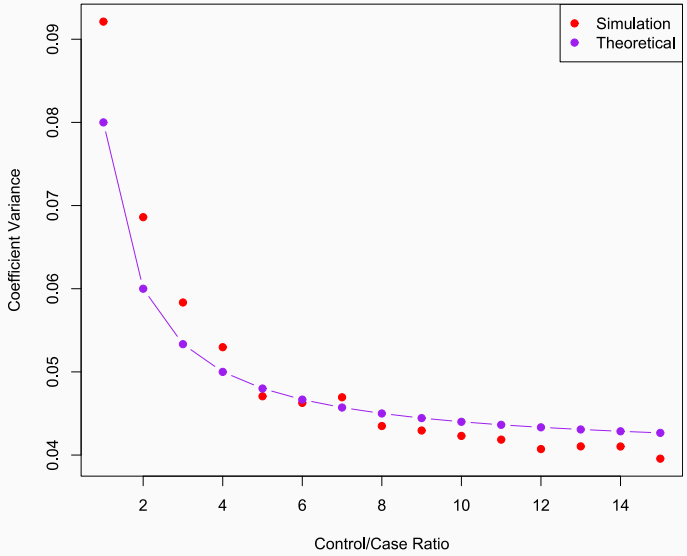
Vzorkování případ-kontrola a logistická regrese

- V datech z Jižní Afriky je 160 případů a 302 kontrol — $\tilde{\pi} = 0{,}35$ jsou případy. Přesto je riziko IM v této oblasti $\pi = 0{,}05$.
- Při vzorkování případ-kontrola můžeme odhadnout regresní parametry β_j přesně (pokud je náš model správný); regresní konstanta β_0 je nesprávná.
- Odhadnutou regresní konstantu můžeme opravit jednoduchou transformací

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log \frac{\pi}{1-\pi} - \log \frac{\tilde{\pi}}{1-\tilde{\pi}}.$$

- Často jsou případy vzácné a bereme je všechny; počet kontrol až do pětinásobku je postačující. Viz následující slajd.

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.

Logistická regrese s více než dvěma třídami

Až dosud jsme probírali logistickou regresi se dvěma třídami. Dá se snadno zobecnit na více než dvě třídy. Jedna verze (používaná v balíčku `glmnet` jazyka R) má symetrický tvar

$$\mathrm{P}(Y = k|X) = \frac{\mathrm{e}^{\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} \mathrm{e}^{\beta_{0\ell}+\beta_{1\ell}X_1+\cdots+\beta_{p\ell}X_p}}$$

Je zde lineární funkce pro **každou** třídu.

(*Matematicky založený* student si může všimnout, že výraz lze zjednodušit a že je zapotřebí pouze $K - 1$ lineárních funkcí podobně jako u dvoutřídní logistické regrese.)

Logistická regrese s více třídami se nazývá také **multinomiální regrese**.

Náš přístup zde je modelovat rozdělení X v každé třídě odděleně a pak použít **Bayesovu větu** k obrácenému pohledu na věc a získat $\mathrm{P}(Y|X)$.

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

Reverend 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**:

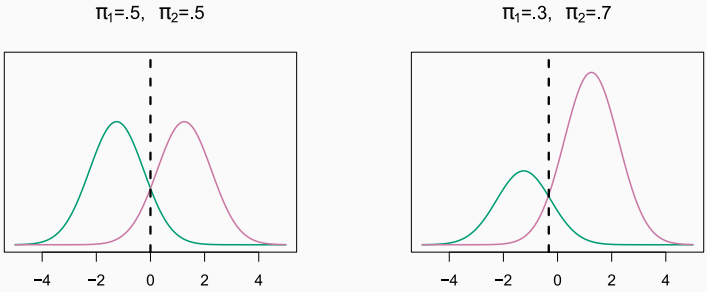
$$\mathrm{P}(Y = k|X = x) = \frac{\mathrm{P}(X = x|Y = k) \cdot \mathrm{P}(Y = k)}{\mathrm{P}(X = x)}$$

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 rozdělení** 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, protože $\pi_1 < \pi_2$ – 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 $p_k(x) = \mathrm{P}(Y = k|X = x)$:

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

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**:

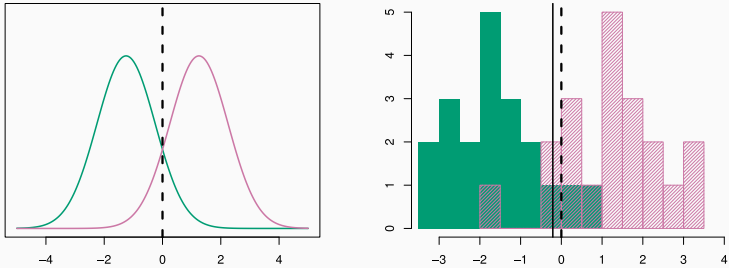
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

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.

Odhadování parametrů

$$\hat{\pi}_k = \frac{n_k}{n}$$

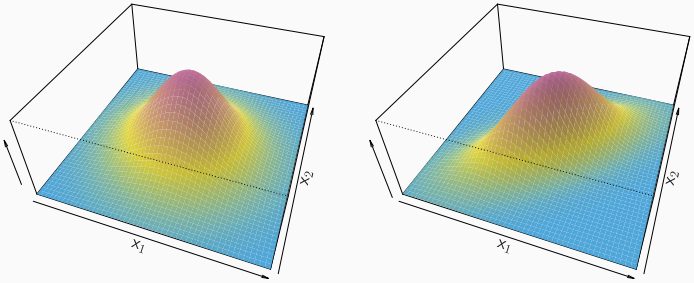
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma}_k^2$$

kde $\hat{\sigma}_k^2 = \frac{1}{n_k-1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$ je obvyklý vzorec pro odhad rozptylu v k-té třídě.

Lineární diskriminační analýza při $p > 1$



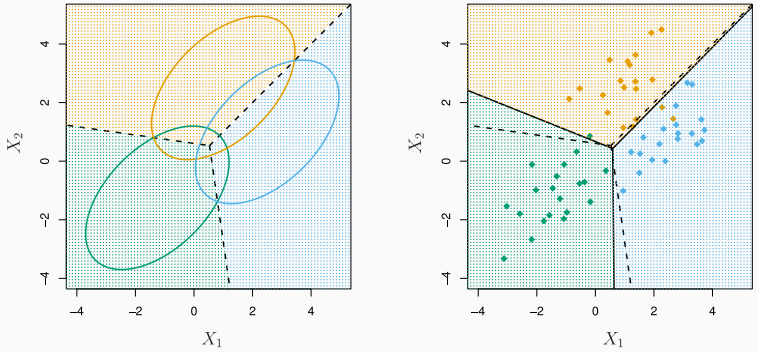
$$\text{Hustota} : f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$\text{Diskriminační funkce: } \delta_k(\mathbf{x}) = \mathbf{x}^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k$$

Nehledě na její složitý tvar je $\delta_k(x)$ lineární funkce:

$$\delta_k(\mathbf{x}) = c_{k0} + c_{k1} x_1 + c_{k2} x_2 + \cdots + c_{kp} x_p.$$

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.

Fisherova data o kosatcích (iris)

4 proměnné

3 odrůdy

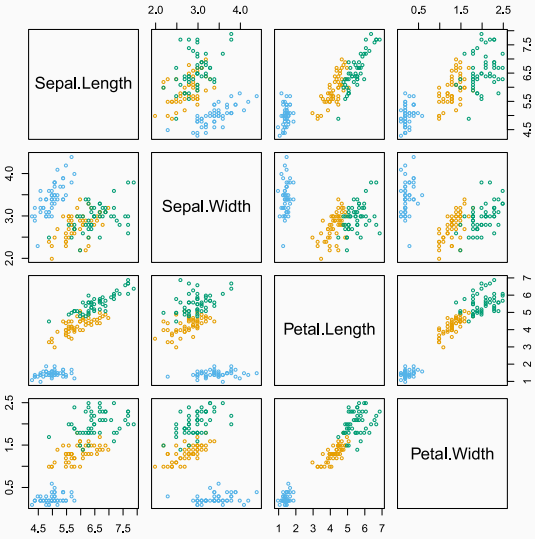
50 vzorků na třídu

- Setosa
- Versicolor
- Virginica

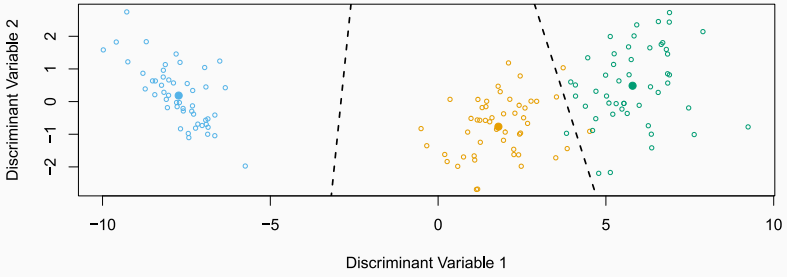
sepal = kališní lístek,

petal = okvětní lístek

LDA klasifikuje všechny až na 3 ze 150 trénovacích vzorků správně.



Fisherův diskriminační graf



Jestliže máme K tříd, lineární diskriminační analýzu lze vidět přesně na $(K-1)$-rozměrném grafu.

Proč? Protože v podstatě klasifikuje na nejbližší centroid, a ty generují $(K-1)$-rozměrnou rovinu.

I pro $K > 3$ můžeme nalézt „nejlepší“ dvourozměrnou rovinu pro vizualizaci diskriminačního pravidla.

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

Máme $(23 + 252)/10000$ chyb — míra chybné klasifikace je 2.75 %!

Některá upozornění:

- Toto je **trénovací** chyba, a možná tu je přeurčení. 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

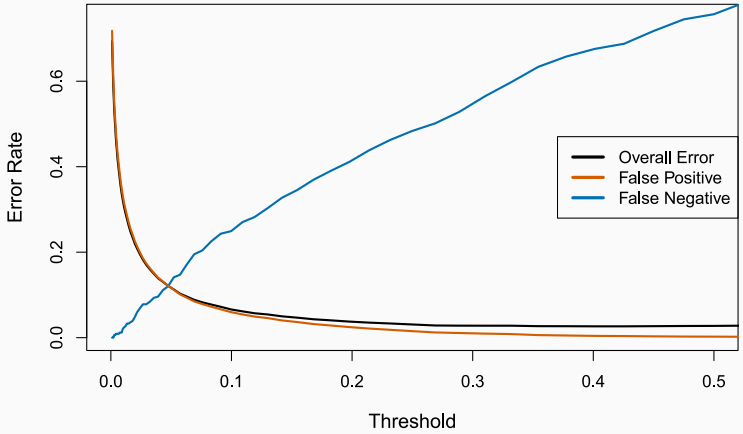
$$\hat{\mathrm{P}}(\texttt{Default} = \texttt{Ano} \,|\, \texttt{Balance}, \texttt{Student}) \geq 0{,}5.$$

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

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

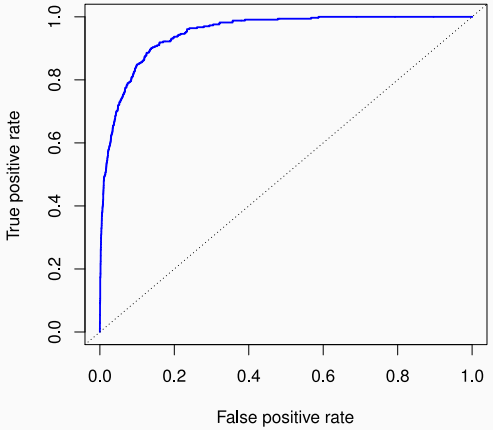
a měníme *práh*.

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.

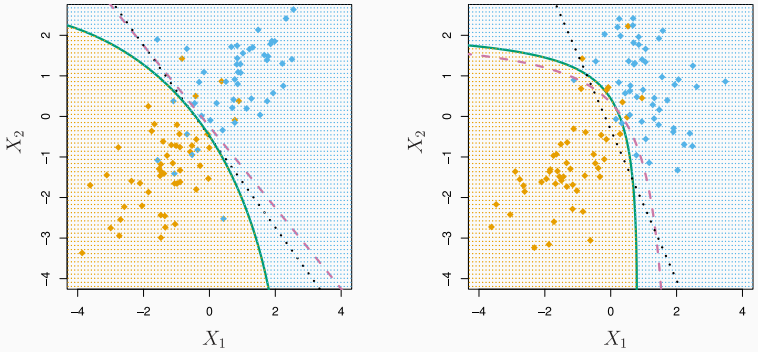
Další způsoby diskriminační analýzy

$$\mathrm{P}(Y = k|X = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\mathbf{x})}$$

Pokud jsou $f_k(\mathbf{x})$ Gaussovské hustoty se stejnou kovarianční maticí $\mathbf{\Sigma}$ v každé třídě, toto vede na lineární diskriminační analýzu. Měněním tvarů $f_k(\mathbf{x})$ dostáváme odlišné klasifikátory.

- S Gaussiány, ale odlišnou $\mathbf{\Sigma}_k$ v každé třídě, dostáváme **kvadratickou diskriminační analýzu**.
- Pro $f_k(\mathbf{x}) = \prod_{j=1}^{p} f_{jk}(x_j)$ (model s podmíněnou nezávislostí) v každé ze tříd dostaneme **naivní Bayesův klasifikátor**. Pro Gaussián toto znamená, že $\mathbf{\Sigma}_k$ jsou diagonální.
- Mnoho dalších způsobů, tím, že pro $f_k(x)$ použijeme specifické modely hustoty, počítaje v to neparametrické přístupy.

Kvadratická diskriminační analýza



$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log\pi_k$$

Protože matice $\mathbf{\Sigma}_k$ jsou rozdílné, kvadratické členy zde hrají roli.

Naivní Bayesův klasifikátor

Předpokládá se, že vlastnosti jsou v každé třídě nezávislé.
Užitečné, je-li p velké, takže metody pro více proměnných jako QDA a dokonce LDA selhávají.

- Gaussovský naivní Bayesüv klasifik8tor předpokládá, že každá matice $\mathbf{\Sigma}_k$ je diagonální:

$$\delta_k(x) \propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] = -\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log \pi_k$$

- Dá se použít pro **smíšené** vektory vlastností (kvalitativní a kvantitativní). Jestliže je X_j kvalitativní, nahraďte $f_{kj}(x_j)$ pravděpodobnostní funkcí (histogramem) přes diskrétní kategorie.

Nehledě na silné předpoklady poskytuje naivní Bayesův klasifikátor často dobré klasifikační výsledky.

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.