

# Úvod do statistického učení

## Matematické metody pro ITS (11MAMY)

---

Jan Přikryl

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

7. přednáška 11MAMY

čtvrtek 31. března 2022

verze: 2022-03-30 18:55

Ústav aplikované matematiky

ČVUT v Praze, Fakulta dopravní

Údaje o mzdách

Jak IBM vybudovalo Watsona, svůj superpočítač hrající hru Jeopardy od Dawna Kawamoto, DailyFinance, 8.2.2011



**Učení se z vlastních chyb:** Podle Davida Ferrucciho (Watson DeepQA technology, IBM Research) je Watsonův software určen nejen ke zvládnutí práce s přirozeným jazykem.

*„Jeho strojové učení počítači umožňuje, aby se stával chytřejším tak, jak se pokouší odpovídat na otázky – a aby se učil z toho, jak zjišťuje, zda odpověděl správně nebo špatně.“*





# FiveThirtyEight

Nate Silver's Political Calculus

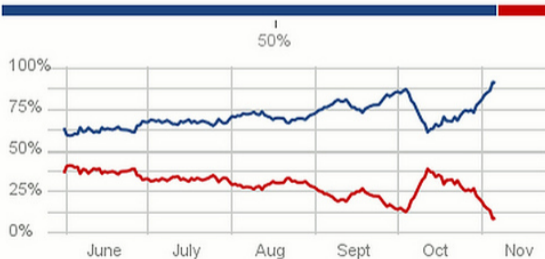
90.9%

+13.5 since Oct. 30

Chance of  
Winning

9.1%

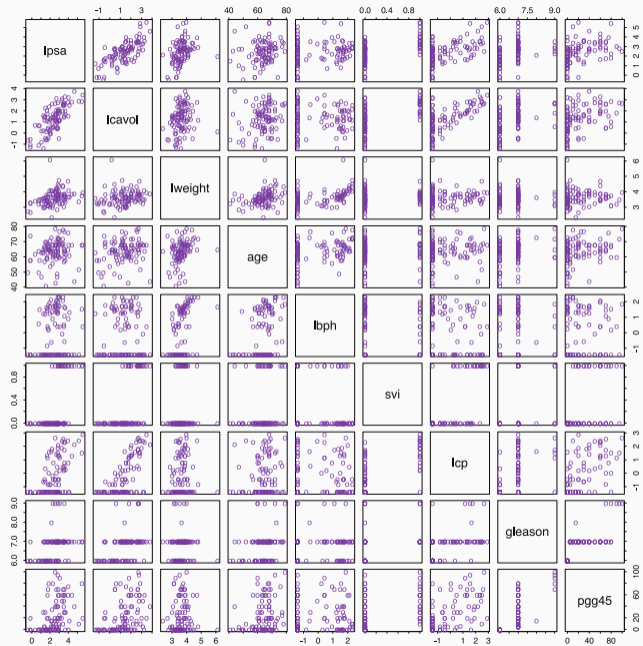
-13.5 since Oct. 30



Click to **LOOK INSIDE!**

*the signal and the noise and the noise and the noise and the noise why so many predictions fail – but some don't the signal and the noise and the noise and the noise nate silver noise and the noise*

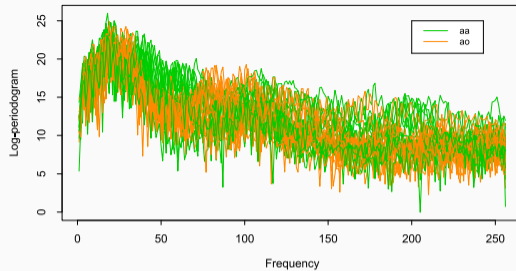
- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci nevyžádaných e-mailů.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.



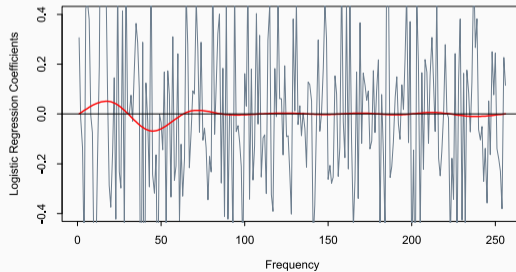
- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci nevyžádaných e-mailů.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.



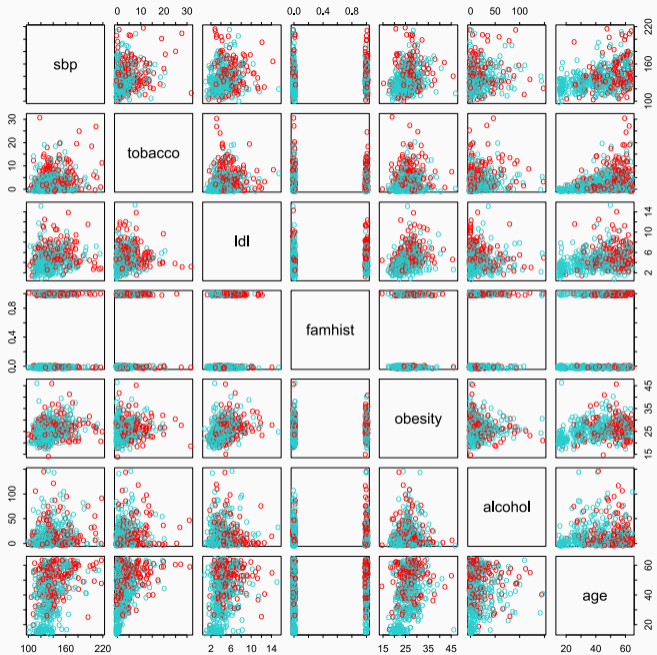
### Phoneme Examples



### Phoneme Classification: Raw and Restricted Logistic Regression



- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci nevyžádaných e-mailů.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.



- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- **Zdokonalování systému pro detekci nevyžádaných e-mailů.**
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.

- data ze 4601 emailových zpráv zaslaných jistému člověku (jménem George, HP labs, před rokem 2000). Každá zpráva je označena jako **spam** nebo **email**.
- cíl: vytvořit z těchto dat systém pro detekci spamu.
- vstupní proměnné: relativní frekvence 57 slov a interpunkčních znamének, které se v těchto emailových správách vyskytují nejčastěji.

|       | george | you  | hp   | free | !    | edu  | remove |
|-------|--------|------|------|------|------|------|--------|
| spam  | 0.00   | 2.26 | 0.02 | 0.52 | 0.51 | 0.01 | 0.28   |
| email | 1.27   | 1.27 | 0.90 | 0.07 | 0.11 | 0.29 | 0.01   |

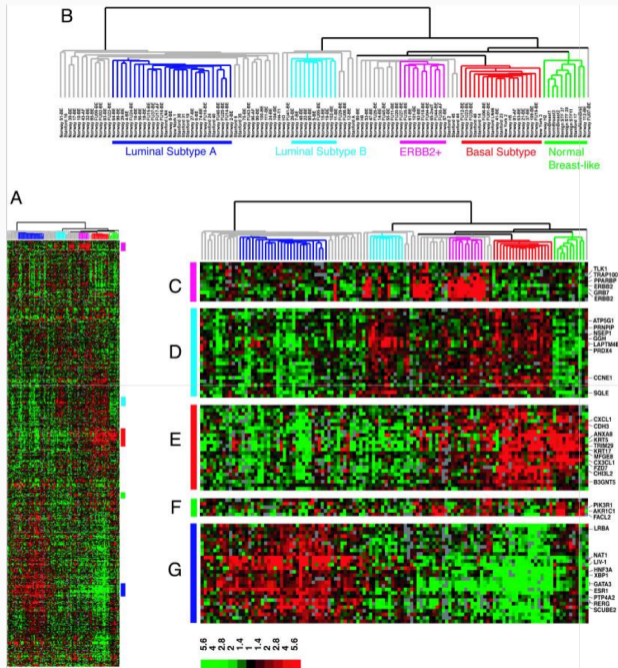
*Průměrný procentní podíl slov nebo znaků ve zprávě, která se rovnají slovům nebo znakům uvedeným v tabulce. Vybrali jsme slova a znaky vykazující největší rozdíl mezi spamem a emailem.*

- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci nevyžádaných e-mailů.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.

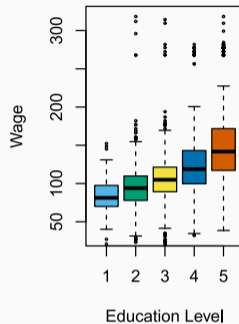
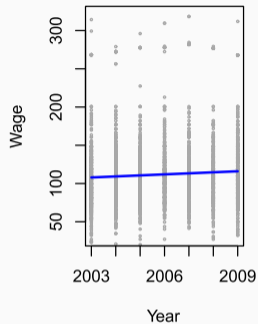
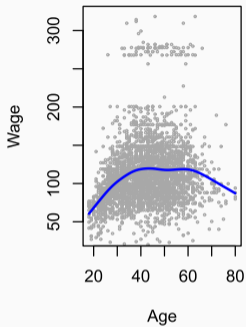


- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci nevyžádaných e-mailů.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.





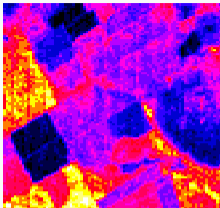
- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci nevyžádaných e-mailů.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.



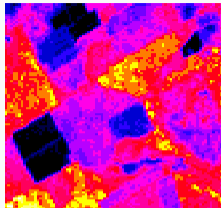
Datová studie příjmu pro muže ve střední atlantické oblasti USA z roku 2009.

- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci nevyžádaných e-mailů.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.

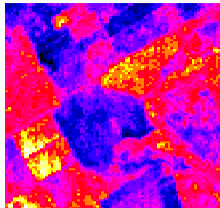
Spectral Band 1



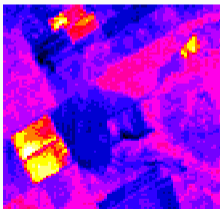
Spectral Band 2



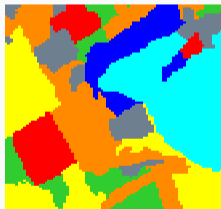
Spectral Band 3



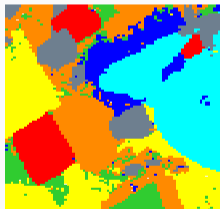
Spectral Band 4



Land Usage



Predicted Land Usage



$Využití \in \{\text{červená zemina, bavlna, strniště, směs, šedá zemina, vlhká šedá zemina}\}$

## Výchozí situace:

- Výstupní měření  $Y$  (také závisle proměnná, odpověď nebo cílová hodnota).
- Vektor  $x$  obsahující  $p$  měření prediktoru  $X$  (také vstupy, regresory, vlastnosti, charakteristiky nebo nezávisle proměnné).
- V *regresní úloze* je  $Y$  kvantitativní (např. cena, hodnota krevního tlaku).
- V *klasifikační úloze* nabývá  $Y$  hodnot z konečné neuspořádané množiny (přežil/zemřel, číslice 0 až 9, třída rakoviny ve vzorku tkáně).
- Máme trénovací data  $(x_1, y_1), \dots, (x_N, y_N)$ . Jsou to výsledky pozorování (příklady, případy) těchto měření.

Na základě trénovacích dat bychom chtěli:

- Přesně předpovídat dosud nezhlédnuté testovací případy.
- Pochopit, které vstupní hodnoty ovlivňují výstupní hodnotu a jak.
- Posoudit kvalitu našich předpovědí a závěrů.

- Je důležité chápat myšlenky rozličných technik, abychom věděli, kdy a jak je používat.
- Člověk musí nejprve pochopit jednodušší metody, aby mohl zvládnout ty náročnější.
- Je důležité přesně posoudit výkonnost metody, abychom věděli jak dobře nebo jak špatně funguje — navíc jsou jednodušší metody často stejně dobré jako ty fajnovější!
- Toto je vzrušující oblast výzkumu, která má aplikace ve vědě, průmyslu a financích.
- Statistické učení je zásadní součástí tréninku moderního *experta v oboru zpracování dat*.



- Žádná výstupní proměnná, pouze sada prediktorů (vlastností) naměřená na sadě vzorků.
- Cíl je mlhavější – nalézt skupiny vzorků, které se chovají podobně, nalézt proměnné, které se chovají podobně, nalézt lineární kombinace vlastností s největší variací.
- Obtížně se pozná, jak dobře si vedete.
- Odlišné od učení s učitelem, ale může být užitečné jako preprocesor pro učení s učitelem.

# NETFLIX

Cena Netflixu

- Trénovací data tvořilo hodnocení necelých 18 000 filmů cca 480 000 zákazníky Netflixu v počtu, jednotlivá hodnocení jsou 1 až 5.
- Trénovací data jsou velmi řídká – asi v 98 % chybí.
- Cílem je předpovědět hodnocení pro soubor jednoho milionu dvojic **zákazník-film**, jež se v trénovacích datech nevyskytují.
- Původní algoritmus *Cinemat* Netflixu dosahoval na testovacích datech úplné střední chyby (RMSE) 0,9525.
- Soutěž započala v říjnu 2006. První tým, který dosáhne zlepšení výsledku o 10 % (tj.  $RMSE < 0,8572$ ) vyhrává 1 milion dolarů.
- Je to úloha s učitelem nebo bez učitele?

# Netflix Prize

**COMPLETED**
[Home](#)   [Rules](#)   [Leaderboard](#)   [Update](#)

## Leaderboard

 Showing Test Score. [Click here to show quiz score](#)

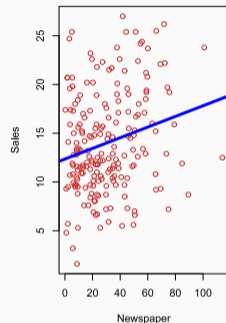
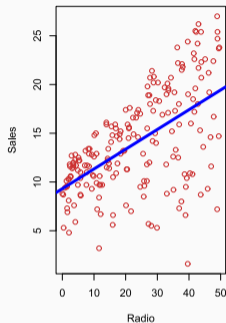
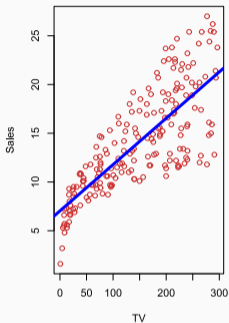
 Display top  leaders.

| Rank  | Team Name   | Best Test Score | % Improvement | Best Submit Time    |
|---|---|-----------------|---------------|---------------------|
| Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos |   |                 |               |                     |
| 1   | <a href="#">BellKor's Pragmatic Chaos</a>           | 0.8567          | 10.06         | 2009-07-26 18:18:28 |
| 2   | <a href="#">The Ensemble</a>                        | 0.8567          | 10.06         | 2009-07-26 18:38:22 |
| 3   | <a href="#">Grand Prize Team</a>                    | 0.8582          | 9.90          | 2009-07-10 21:24:40 |
| 4   | <a href="#">Opera Solutions and Vandelay United</a> | 0.8588          | 9.84          | 2009-07-10 01:12:31 |
| 5   | <a href="#">Vandelay Industries !</a>               | 0.8591          | 9.81          | 2009-07-10 00:32:20 |
| 6   | <a href="#">PragmaticTheory</a>                     | 0.8594          | 9.77          | 2009-06-24 12:06:56 |
| 7   | <a href="#">BellKor in BigChaos</a>                 | 0.8601          | 9.70          | 2009-05-13 08:14:09 |
| 8   | <a href="#">Dace</a>                                | 0.8612          | 9.59          | 2009-07-24 17:18:43 |
| 9   | <a href="#">Feeds2</a>                              | 0.8622          | 9.48          | 2009-07-12 13:11:51 |
| 10  | <a href="#">BigChaos</a>                            | 0.8623          | 9.47          | 2009-04-07 12:33:59 |
| 11  | <a href="#">Opera Solutions</a>                     | 0.8623          | 9.47          | 2009-07-24 00:34:07 |
| 12  | <a href="#">BellKor</a>                             | 0.8624          | 9.46          | 2009-07-26 17:19:11 |

Vyhrál tým s názvem [BellKor's Pragmatic Chaos](#), který velmi těsně porazil [The Ensemble](#).

- Strojové učení vzniklo jako podobor umělé inteligence.
- Statistické učení vzniklo jako podoblast statistiky.
- *Je zde značný překryv* – oba obory se zaměřují na úlohy učení s učitelem a bez učitele:
  - Strojové učení klade větší důraz na aplikace ve *velkém měřítku* a na *přesnost předpovědi*.
  - Statistické učení klade důraz na *modely* a jejich interpretovatelnost a na *přesnost* a *nejistotu*.
- Ale rozdíly se stále více a více stírají a je zde velký podíl vzájemného obohacování.
- Strojové učení má navrch v *marketingu*!

# Úvod do metod statistického učení



Vidíme grafy **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:

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

V modelu jsou **Prodeje** *odpověď* nebo *cílová hodnota*. Odpověď obvykle značíme  $y$ .

**TV** je *charakteristika, vlastnost, vstup* nebo *prediktor*, označíme ho  $x_1$ .

Podobně označíme **Rozhlas** jako  $x_2$  a tak dále.

Můžeme potom odkazovat na *vstupní vektor* souhrnně jako na

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

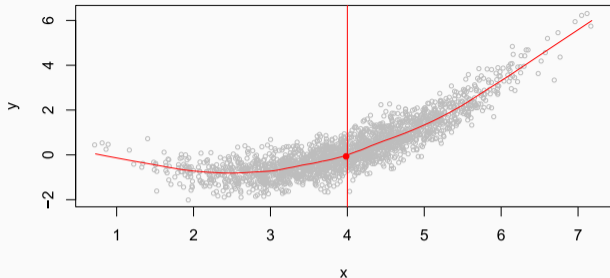
Náš model nyní zapíšeme jako

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

kde  $\epsilon$  zachycuje chyby měření a jiné nepřesnosti.



- S dobrým  $f$  můžeme dělat předpovědi  $y$  v nových bodech  $x = x^*$ .
- Můžeme přijít na to, které složky  $x = (x_1, x_2, \dots, x_p)^T$  jsou pro pochopení  $y$  důležité a které jsou irelevantní. Tak např. **Stáří** a **Rokyvzdělávání** 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  $x = 4$ ?

Pro  $x = 4$  můžeme mít naměřeno mnoho různých hodnot  $y$ . Dobrá hodnota funkce  $f$  je

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

$E[Y|X = 4]$  značí *očekávanou hodnotu* (průměr) hodnot náhodné veličiny  $Y$  pro dané  $X = 4$ .

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

- Je také definována pro vektor  $\mathbf{x}$ , např.

$$f(\mathbf{x}) = f(x_1, x_2, x_3) = E[Y|X_1 = x_1, X_2 = x_2, X_3 = x_3]$$

- 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  $f(x)$  zcela přesně, stejně bychom dělali chyby v předpovídání, neboť v každém bodě  $X = x$  typicky existuje rozložení pravděpodobnosti výskytu možných hodnot  $Y$ .

- Pro každý odhad  $\hat{f}(x)$  funkce  $f(x)$  máme

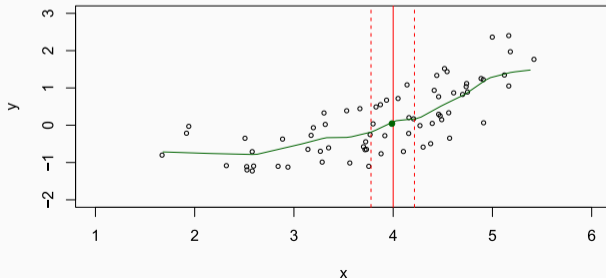
$$E[(Y - \hat{f}(x))^2|X = x] = \underbrace{(f(x) - \hat{f}(x))^2}_{\text{redukovatelné}} + \underbrace{\text{var}(\epsilon)}_{\text{neredukovatelné}}$$

Typicky máme málo bodů pro  $X = 4$  změřeno přesně (pokud vůbec nějaké) ...

- Takže  $E[Y|X = x]$  nemůžeme spočítat!
- Zmírněme definici a položme

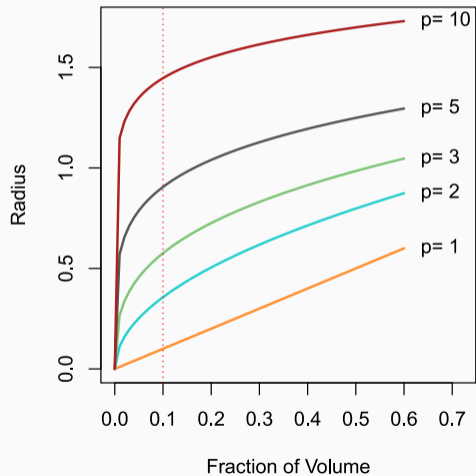
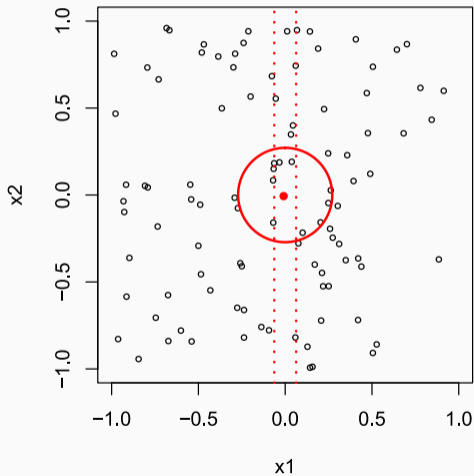
$$\hat{f}(x) = \text{average}(Y|X \in \mathcal{R}(x))$$

kde  $\mathcal{R}(x)$  je nějaké *okolí* bodu  $x$  a average je průměr.



- 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  $E[Y|X = x]$  lokálním průměrováním.

## 10% Neighborhood

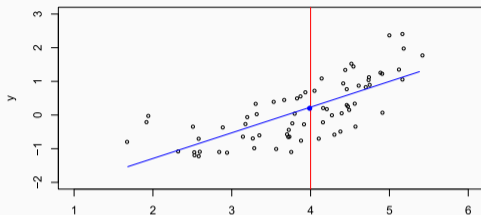


*Lineární* model je důležitý příklad parametrického modelu:

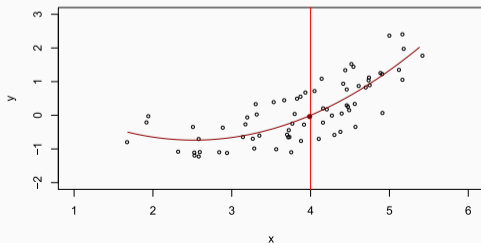
$$f_L(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

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

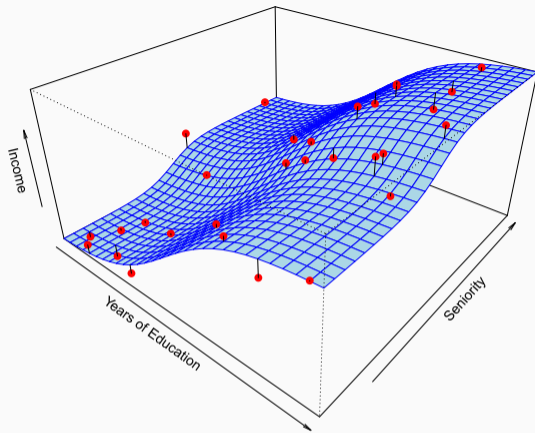
Lineární model  $\hat{f}_L(x) = \hat{\beta}_0 + \hat{\beta}_1 x$  zde dává rozumnou aproximaci:



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



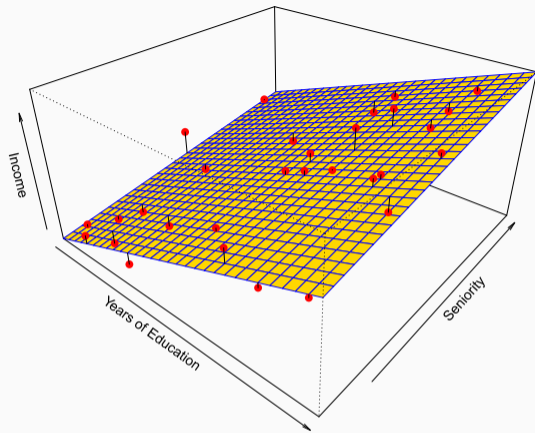




Simulovaný příklad. Červené body jsou simulované hodnoty příjmu z modelu

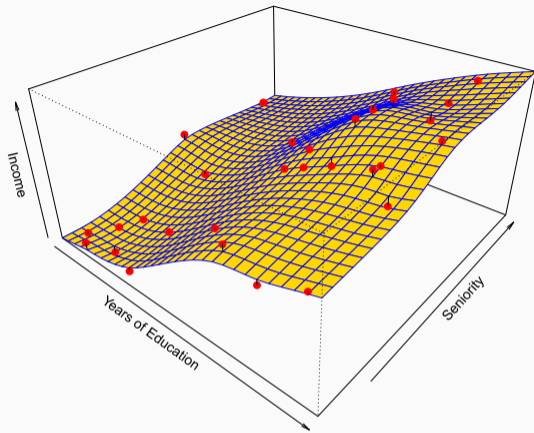
$$\text{income} = f(\text{education}, \text{seniority}) + \epsilon,$$

$f$  je ta modrá plocha.

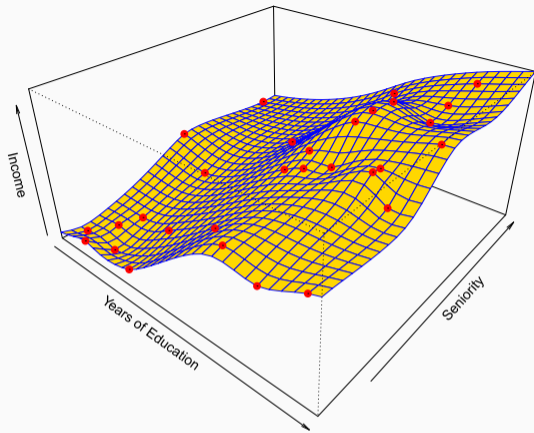


Lineární regresní model proložený nasimulovanými daty:

$$\hat{f}_{\mathcal{L}}(\text{education}, \text{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \text{education} + \hat{\beta}_2 \times \text{seniority}$$

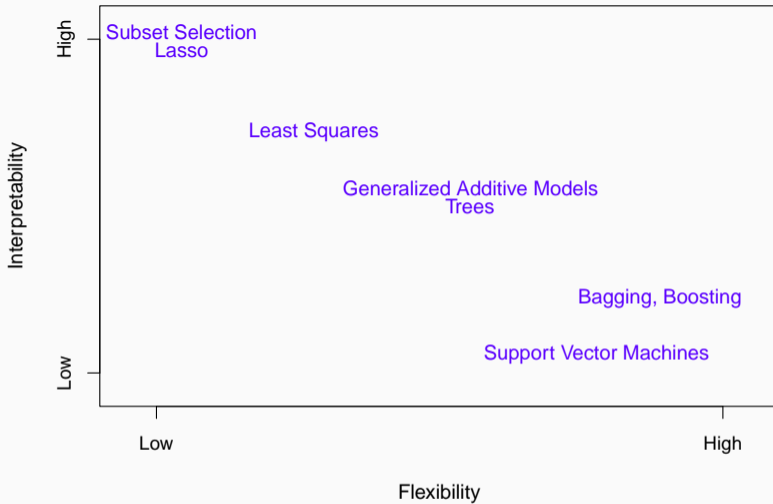


Flexibilnější regresní model  $\hat{f}_S(\text{education}, \text{seniority})$  proložený nasimulovanými daty. Používáme zde postup nazývaný *splajn tenké desky* (thin-plate spline), abychom proložili pružnou plochu. Ovládáme hrbolatost této aproximace (ISLR, kapitola 7).



Ještě flexibilnější regresní model  $\hat{f}_S(\text{education}, \text{seniority})$  pomocí splajnu proložený nasimulovanými daty. Proložený model zde v trénovacích datech nedělá žádné chyby! Známo také jako *přeurčení* nebo *přetrénování* (overfitting).

- Přesnost předpovědi versus interpretovatelnost.
  - Lineární modely se snadno interpretují, splajny tenké desky nikoli.
- Dobrá aproximace versus přeurčení nebo podurčení.
  - Jak poznáme, že aproximace je zrovna ta pravá?
- Úspornost versus černá skříňka.
  - Často dáváme přednost jednoduššímu modelu s méně proměnnými před prediktorem typu černé skříňky, který je zahrnuje všechny.



Překlad názvů metod (zleva doprava, odshora dolů): výběr podmnožiny, Lasso, nejmenší čtverce, zobecněné aditivní modely, stromy, bagging, boosting, metoda podpůrných vektorů.

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

- Mohli bychom vypočítat střední kvadratickou chybu předpovědi přes  $\mathbf{T}_r$ :

$$\text{MSE}_{\mathbf{T}_r} = \text{average}_{i \in \mathbf{T}_r} \left( y_i - \hat{f}(x_i) \right)^2$$

Tento přístup ale může více stranit přeureným modelům!

**Q:** Proč tomu tak je?

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

$$\text{MSE}_{\mathbf{T}_e} = \text{average}_{i \in \mathbf{T}_e} \left( y_i - \hat{f}(x_i) \right)^2$$

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

- Mohli bychom vypočítat střední kvadratickou chybu předpovědi přes  $\mathbf{T}_r$ :

$$\text{MSE}_{\mathbf{T}_r} = \text{average}_{i \in \mathbf{T}_r} \left( y_i - \hat{f}(x_i) \right)^2$$

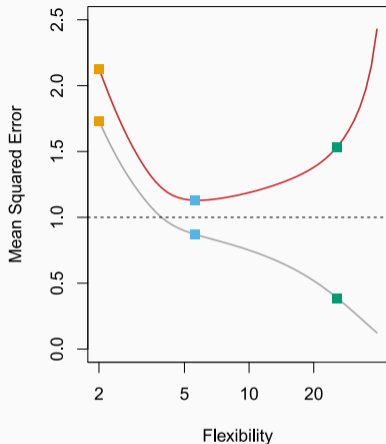
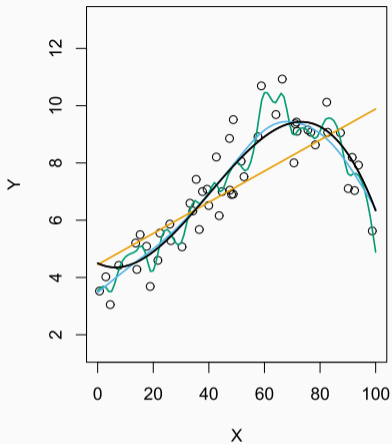
Tento přístup ale může více stranit přeurčeným modelům!

Q: Proč tomu tak je?

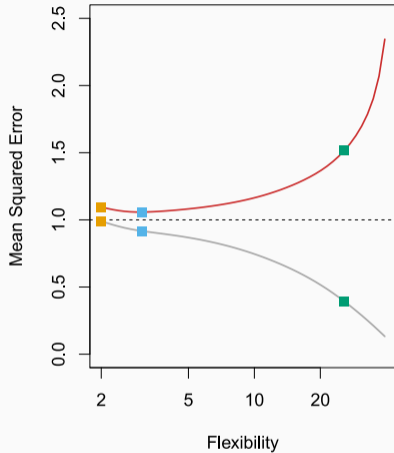
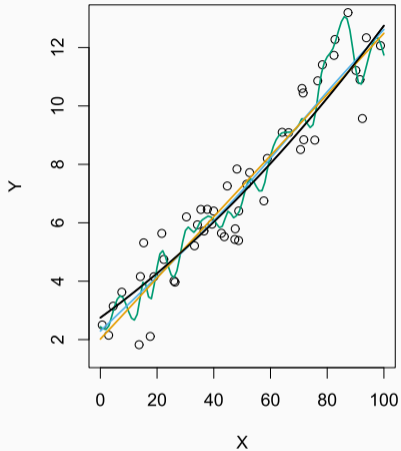
- Místo toho bychom měli, pokud je to možné, vypočítat chybu předpovědi pomocí nových *testovacích* dat  $\mathbf{T}_e = \{x_i, y_i\}_1^M$ :

$$\text{MSE}_{\mathbf{T}_e} = \text{average}_{i \in \mathbf{T}_e} \left( y_i - \hat{f}(x_i) \right)^2$$

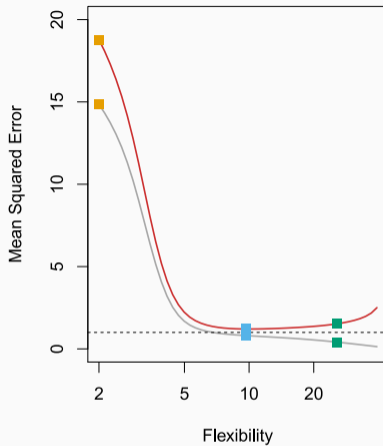
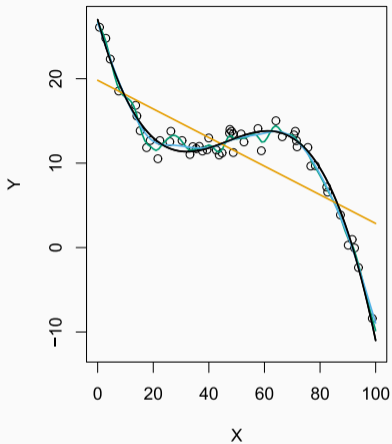




Černá křivka je skutečnost. Červená křivka vpravo je  $MSE_{Te}$ , šedá křivka je  $MSE_{Tr}$ . Oranžová, modrá a zelená křivka (a čtverečky těchto barev) odpovídají aproximacím různé flexibility.



Zde je skutečnost hladší, takže hladší aproximace a lineární model si vedou opravdu dobře.



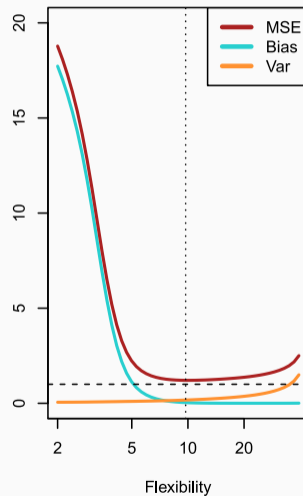
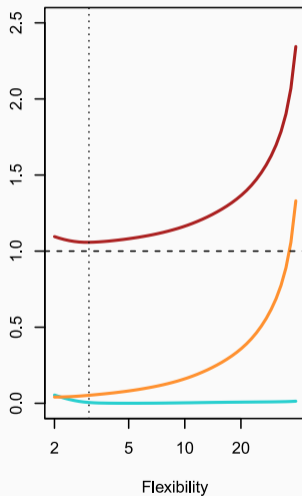
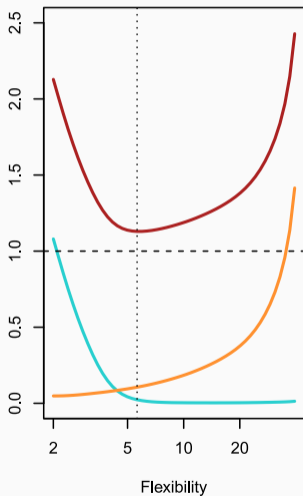
Zde je skutečnost zvlněná a šum je malý, takže si nejlépe vedou flexibilnější aproximace.

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

$$E \left[ y_0 - \hat{f}(x_0) \right]^2 = \text{var}(\hat{f}(x_0)) + \left[ \text{bias}(\hat{f}(x_0)) \right]^2 + \text{var}(\epsilon).$$

Očekávaná hodnota průměruje přes variabilitu v  $y_0$  a rovněž variabilitu v  $\text{Tr}$ , která ovlivňuje  $\hat{f}$ . Poznamenejme, že  $\text{bias}(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0)$ .

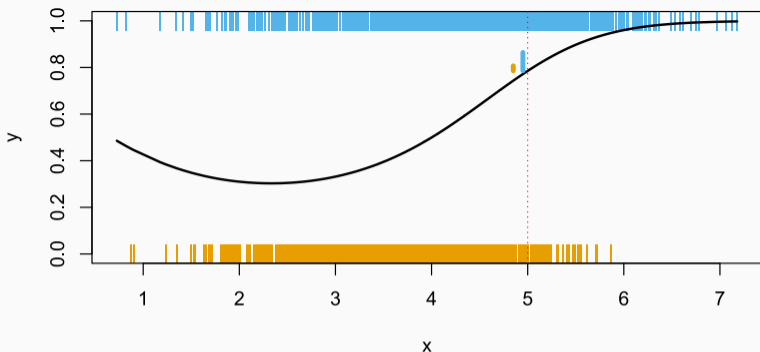
Typické je, že jak roste *flexibilita*  $\hat{f}$ , roste její rozptyl a zkreslení (bias) se snižuje. Takže *volba flexibility založená na střední testovací chybě odpovídá kompromisu mezi zkreslením a rozptylem.*



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

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  $X = (X_1, X_2, \dots, X_p)$ .

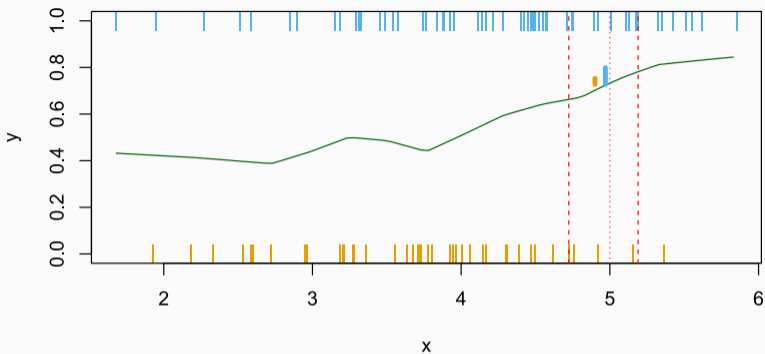


Existuje ideální  $C(X)$ ? Předpokládejme, že  $K$  prvků množiny  $\mathcal{C}$  je očíslováno  $1, 2, \dots, K$ . Položme

$$p_k(x) = P(Y = k | X = x), \quad k = 1, 2, \dots, K.$$

Toto jsou *podmíněné pravděpodobnosti tříd* pro dané  $x$ , viz např. malý sloupcový graf pro  $x = 5$ . *Bayesův optimální klasifikátor* pro  $x$  je pak

$$C(x) = j \quad \text{jestliže} \quad p_j(x) = \max\{p_1(x), p_2(x), \dots, p_K(x)\}.$$



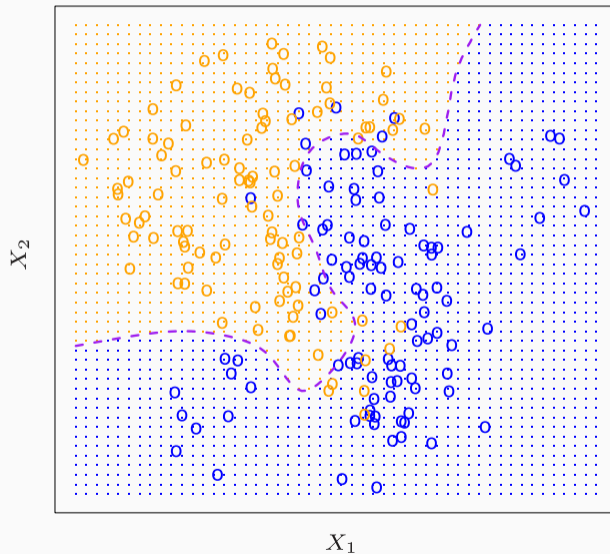
Dá se zde stejně jako dříve použít průměrování přes nejbližší sousedy.  
 A pro rostoucí dimenze se to také hroutí. Avšak dopad na  $\hat{C}(x)$  je menší než na  $\hat{p}_k(x)$ ,  $k = 1, 2, \dots, K$ .

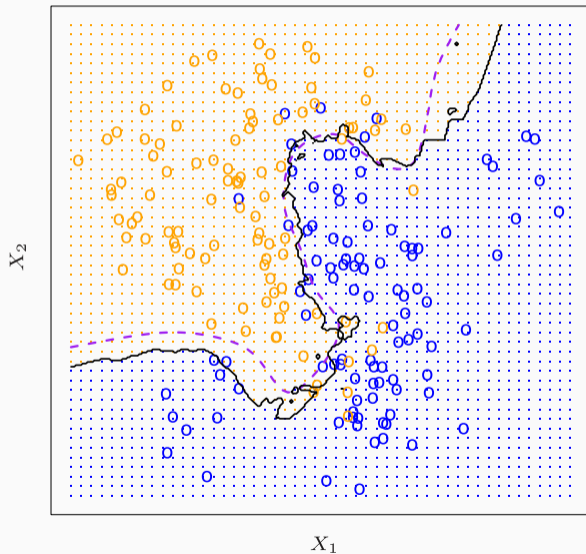


- Výkonnost klasifikátoru  $\hat{C}(x)$  typicky měříme pomocí míry chyby nesprávné klasifikace:

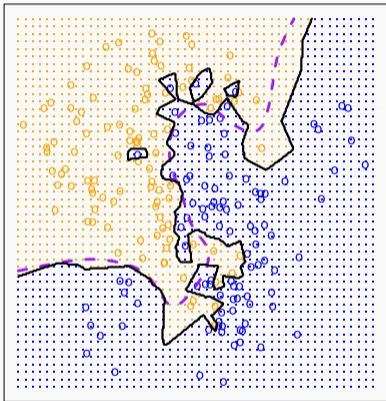
$$\text{Err}_{\text{Te}} = \text{average}_{i \in \text{Te}} I[y_i \neq \hat{C}(x_i)].$$

- Bayesův klasifikátor (užívající skutečné hodnoty  $p_k(x)$ ) má nejmenší chybu (v dané populaci).
- Metoda podpůrných vektorů (SVM) vytváří strukturované modely  $C(x)$ .
- Budeme také vytvářet strukturované modely pro reprezentaci  $p_k(x)$ , např. logistickou regresi, zobecněné aditivní modely.





KNS:  $K = 1$



KNS:  $K = 100$

