Vytěžování dat, přednáška 7: Klasifikace

Filip Železný

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Učení s učitelem

- Pravděpodobnostní rozdělení P_{XY} na $X \times Y$
  - Předpokládáme, že Y je konečná množina
- Data: D je multimnožina

$$D = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_2)\} \ (m \in N)$$

  prvky vybrány náhodně a navzájem nezávisle z P_{XY}
- Vzor se bude používat na tomtéž X, Y a P_{XY}

Obvykle hledáme vzory aproximující (pro zadané $x \in X$)

- podmíněnou pravděpodobnost $P_{Y|X}(y|x) = P_{XY}(x, y)/P_X(x)$
- nebo nejpravděpodobnější hodnotu $\arg\max_{y \in Y} P_{Y|X}(y|x)$
- tyto vzory nazýváme *klasifikátory*

Klasifikace s jedním příznakem

X	Y
Vysoké příjmy	**Splácí úvěr**
ano	ano
ano	ne
ne	ano
ano	ano
ne	ne
ne	ne
ne	ano
ano	ano
ano	ano
ano	ano
ne	ne

Kontingenční tabulka

		Y (splácí úvěr)		
		ano	ne	$\sum$
X (vysoké příjmy)	ano	5	1	6
	ne	2	3	5
	$\sum$	7	4	11

Nový žadatel o úvěr má vysoké příjmy.

- Bude splácet úvěr?
- S jakou pravděpodobností?

Klasifikace dle aposteriorní pravděpodobnosti

- Klient má vysoké příjmy. Jak bude splácet úvěr?
- Tedy z $x =$ vysoké urči nejpravděpodobnější hodnotu y (třídu).
- Hledáme y' vyhovující

$$y' = \arg\max_y P_{Y|X}(y|\text{vysoké})$$

- Řešení je $y' =$ splácí

$$P_{Y|X}(y'|\text{vysoké}) \approx 2/3$$

- S pravděpodobností $1 - P_{Y|X}(y'|\text{vysoké})$ klasifikujeme chybně.
- Klasifikací $y' =$ splácí tedy minimalizujeme chybu.

Ztrátová funkce

- Každá chyba klasifikace je jinak drahá.
- Např. příliš optimististické hodnocení klienta stojí víc než příliš skeptické.
- Klasifikace dle aposteriorní pravděpodobnosti toto nerespektuje. Pro zohlednění odlišných ztrát pro různé chyby potřebujeme pojem ztrátové funkce.
- **Ztrátová funkce** $L(y, y')$ zachycuje ztrátu pro každou kombinaci
  - y - skutečná třída
  - y' - třída, do které klasifikujeme
- Pro náš příklad $L(y, y')$ např.:

y↓ y'→	splácí	problémy	nesplácí
splácí	0	1	2
problémy	5	0	1
nesplácí	10	5	0

Klasifikátor a Riziko

- Hledáme klasifikátor, tj. funkci $f : X \to Y$ minimalizující *střední hodnotu ztráty*:

$$R(f) = \sum_{x,y} L(y, f(x)) P_{XY}(x, y)$$

- Střední hodnotu ztráty pro f se nazývá *riziko* klasifikátoru f.
- Řešením $f^* = \arg\min_f R(f)$ je funkce

$$f^*(x) = \arg\min_{y'} r(x, y')$$

kde

$$r(x, y') = \sum_{y} L(y, y') P_{Y|X}(y|x)$$

je *riziko* klasifikace y' při příznaku x

Klasifikace jako minimalizace rizika

- Jak klasifikovat vysokopříjmového klienta?

$P_{Y|X}$

$x \downarrow y \rightarrow$	splácí	problémy	nesplácí
vysoké	**2/3**	**1/3**	**0/3**
střední	2/6	2/6	2/6
nízké	0/2	1/2	1/2

L

$y \downarrow y' \rightarrow$	splácí	problémy	nesplácí
splácí	**0**	1	2
problémy	**5**	0	1
nesplácí	**10**	5	0

- Klasifikace $y' =$ splácí

skut. třída	**ztráta**	**s pravděp.**
splácí	0	2/3
problémy	5	1/3
nesplácí	10	0

- Riziko při této klasifikaci:

$$0 \cdot 2/3 + 5 \cdot 1/3 + 10 \cdot 0 = 5/3$$

Klasifikace jako minimalizace rizika

- Jak klasifikovat vysokopříjmového klienta?

$P_{Y|X}$

$x \downarrow y \rightarrow$	splácí	problémy	nesplácí
vysoké	**2/3**	**1/3**	**0/3**
střední	2/6	2/6	2/6
nízké	0/2	1/2	1/2

L

$y \downarrow y' \rightarrow$	splácí	**problémy**	nesplácí
splácí	0	**1**	2
problémy	5	**0**	1
nesplácí	10	**5**	0

- Klasifikace $y' =$ problémy

skut. třída	**ztráta**	**s pravděp.**
splácí	1	2/3
problémy	0	1/3
nesplácí	5	0

- Riziko při této klasifikaci:

$$1 \cdot 2/3 + 0 \cdot 1/3 + 5 \cdot 0 = 2/3$$

Klasifikace jako minimalizace rizika

- Jak klasifikovat vysokopříjmového klienta?

$P_{Y|X}$

$x \downarrow y \rightarrow$	splácí	problémy	nesplácí
vysoké	**2/3**	**1/3**	**0/3**
střední	2/6	2/6	2/6
nízké	0/2	1/2	1/2

L

$y \downarrow y' \rightarrow$	splácí	problémy	nesplácí
splácí	0	2	**2**
problémy	5	1	**1**
nesplácí	10	0	**0**

- Klasifikace $y' =$ nesplácí

skut. třída	**ztráta**	**s pravděp.**
splácí	2	2/3
problémy	1	1/3
nesplácí	0	0

- Riziko při této klasifikaci:

$$2 \cdot 2/3 + 1 \cdot 1/3 + 0 \cdot 0 = 4/3$$

Klasifikace jako minimalizace rizika

- Klasifikujeme do

$$y' = \arg\min_y r(\text{vysoké}, y) = \text{problémy}$$

- Pozor, jiný výsledek než dle aposteriorní pravděpodobnosti

$$y' = \arg\max_y P_{Y|X}(y|\text{vysoké}) = \text{splácí}$$

- Při jaké ztrátové funkci $L(y, y')$ by výsledky vyšly stejně?

y↓ y'→	splácí	problémy	nesplácí
splácí	0	1	1
problémy	1	0	1
nesplácí	1	1	0

- Tzv. L_{01} ztrátová funkce. Je-li použita, je $r(x, y')$ pravděpodobnost, že klasifikace instance s příznakem x do třídy y' je chybná.

Klasifikace s několika příznaky

- Zatím jsme klasifikovali pouze dle jediného příznaku (x - výše příjmů)
- O klientech toho víme obvykle více.

Příjmy (p)	Rok narození (n)	Úvěr (y)
vysoké	1969	splácí
nízké	1974	nesplácí
střední	1940	problémy
nízké	1985	problémy
...	...	...

- $x \equiv (p, n)$
- Třírozměrná kontingenční tabulka
  - p vs. n vs. y

Klasifikace s několika příznaky

- Na principech klasifikace se nic nemění. Např. jak klasifikovat nízkopříjmového klienta narozeného v r. 1985?
  - Maximalizací aposteriorní pravděpodobnosti

$$f(x) = y' = \arg\max_y P_{Y|P,N}(y|\text{nízké}, 1974)$$

  - Minimalizací rizika

$$f(x) = y' = \arg\min_y r((\text{nízké}, 1974), y)$$

- Z kontingenční tabulky

$$P_{Y|P,N}(y'|\text{nízké}, 1974) \approx \frac{\text{počet klientů s } y' = y, p = \text{nízké}, n = 1974}{\text{počet klientů s } p = \text{nízké}, n = 1974}$$

Prokletí rozměrnosti

$$P_{Y|P,N}(y'|\text{nízké}, 1974) \approx \frac{\text{počet klientů s } y' = y, p = \text{nízké}, n = 1974}{\text{počet klientů s } p = \text{nízké}, n = 1974}$$

- Čím více příznaků, tím větší nebezpečí výsledku "0/0"!
- Kolik dat potřebujeme, aby odhady dobře konvergovaly k pravděpodobnostem?
- Kontingenční tabulka musí být 'dostatečně zaplněna'.
- V předchozím příkladě (jediný příznak)

p↓ y→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	3
střední	2	2	2	6
nízké	0	1	1	2
$\sum$	4	4	3	11

v průměru 11/9 případů na kolonku tabulky.

Prokletí rozměrnosti

- Předpokládejme, že $11/9$ je dostatečný poměr. Kolik případů (m) potřebujeme pro jeho zachování se dvěma příznaky p a n?
- Přepodkládejme 100 možných roků narození. Kontingenční tabulka má $100 \cdot 3 \cdot 3 = 900$ kolonek.

$$\frac{m}{900} = \frac{11}{9}$$

  tedy nyní již potřebujeme $11 \cdot 900/9 = 1100$ dat.
- Po přidání dalšího příznaku, např. roku ukončení studia už potřebujeme $11 \cdot 90000/9 = 110000$ dat!
- Obecně pro odhady z kontingenční tabulky (tzv. neparametrické odhady) roste potřebný počet dat **exponenciálně** s počtem příznaků.
  - "Prokletí rozměrnosti"

Podmíněně nezávislé příznaky

- Situace se zjednoduší, jsou-li výše příjmů a rok narození **podmíněně nezávislé**, tj. platí

$$P_{P,N|\,Y}(p, n|y) = P_{P|\,Y}(p|y) \cdot P_{N|\,Y}(n|y)$$

pro každou z hodnot $y \in \{\text{splácí}, \text{problémy}, \text{nesplácí}\}$

- Využijeme tzv. Bayesova pravidla

$$P_{Y|P,N}(y|p, n) = \frac{P_{P,N|\,Y}(p, n|y)P_Y(y)}{P_{P,N}(p, n)}$$

Podmíněně nezávislé příznaky

- Z Bayesova pravidla platí pro klasifikaci maximalizací aposteriorní pravděpodobnosti

$$\arg\max_y P_{Y|P,N}(y|p,n) = \arg\max_y P_{P,N|Y}(p,n|y)P_Y(y)$$

- Podobně pro klasifikaci minimalizací rizika

$$\arg\min_{y'} \sum_y L(y,y')P_{Y|P,N}(y|p,n)$$

$$= \arg\min_{y'} \sum_y L(y,y')P_{P,N|Y}(p,n|y)P_Y(y)$$

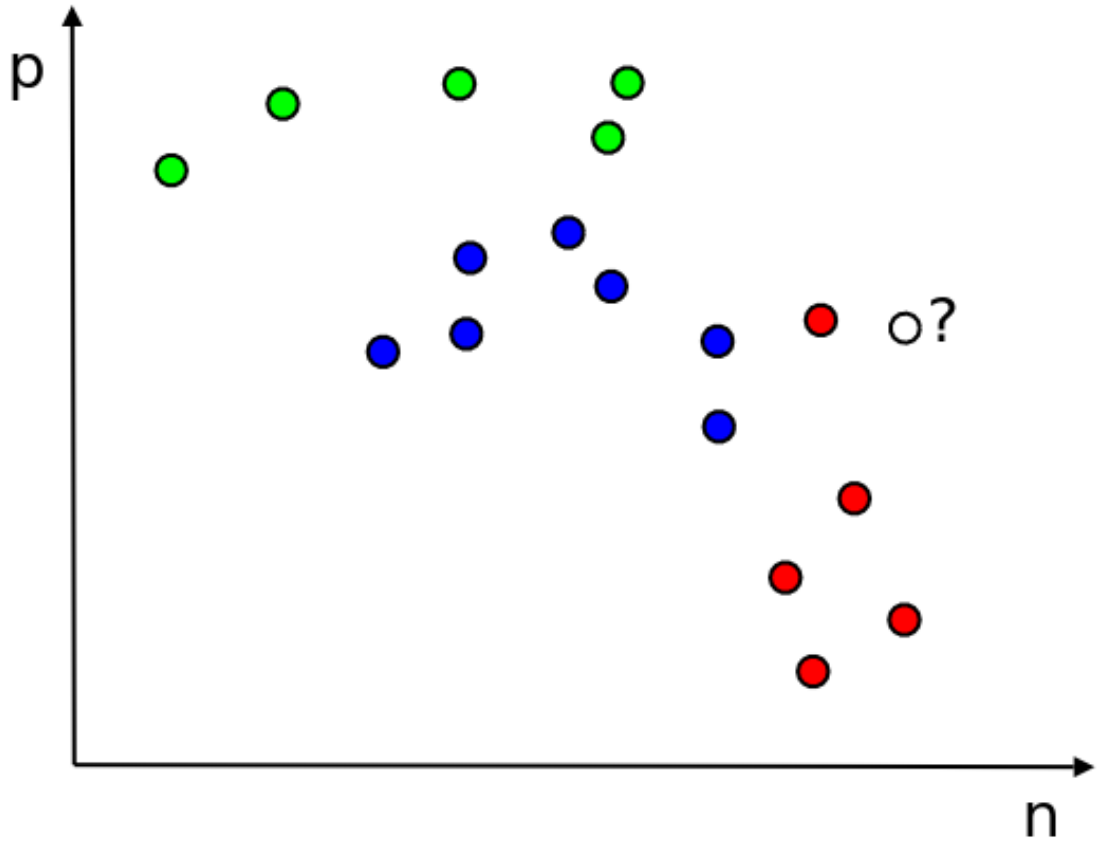
- Proč ‘zmizelo’ $P_{P,N}(p,n)$? Nezávisí na y!
- K oběma typům klasifikace tedy potřebujeme odhady dvou rozdělení: $P_{P,N|Y}$ a P_Y.

Podmíněně nezávislé příznaky

- K oběma typům klasifikace tedy potřebujeme odhady dvou rozdělení: $P_{P,N|\,Y}$ a P_Y.
- P_Y odhadneme z jednorozměrné kontingenční tabulky
- Z podmíněné nezávislosti plyne $P_{P,N|\,Y} = P_{P|\,Y} \cdot P_{N|\,Y}$
- $P_{P|\,Y}$ odhadneme z dvourozměrné tabulky 3×3
- $P_{N|\,Y}$ odhadneme z dvourozměrné tabulky 100×3.
  - Nejnáročnější na počet dat, pro zachování poměru $11/9$ vyžaduje cca 367 ($<< 1100$).
- Obecně: při podmíněně nezávislých příznacích neroste potřebný počet dat exponenciálně s počtem příznaků.
  - Je určen příznakem s největším oborem hodnot.
- Příznaky obvykle nejsou podmíněně nezávislé!
  - Je-li přesto použita tato metoda, mluvíme o **naivní Bayesovské klasifikaci**.

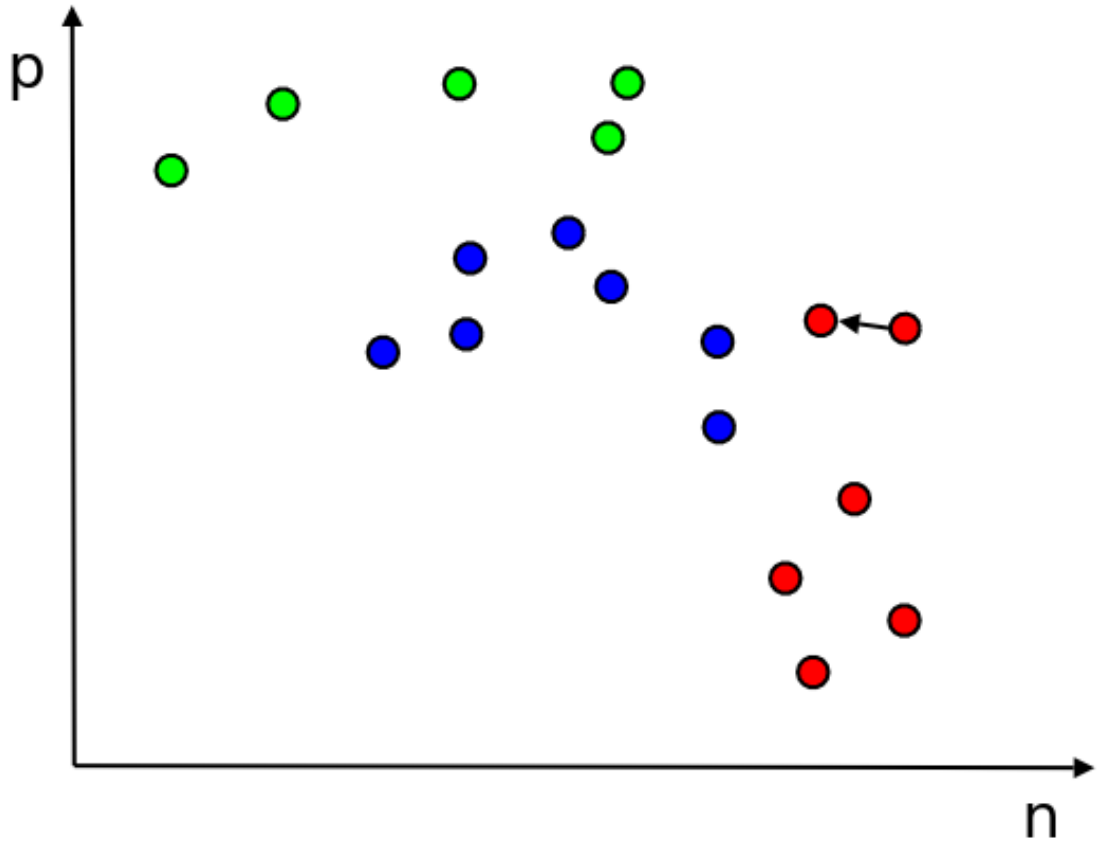
Klasifikace dle nejbližšího souseda

- Metoda, která nevyžaduje odhad pravděpodobností.
- Předpoklad: umíme spočítat podobnost dvou datových instancí (jako např. u shlukování)
- Zde $p, n \in N$, $u \in \{$splácí, problémy, nesplácí$\}$



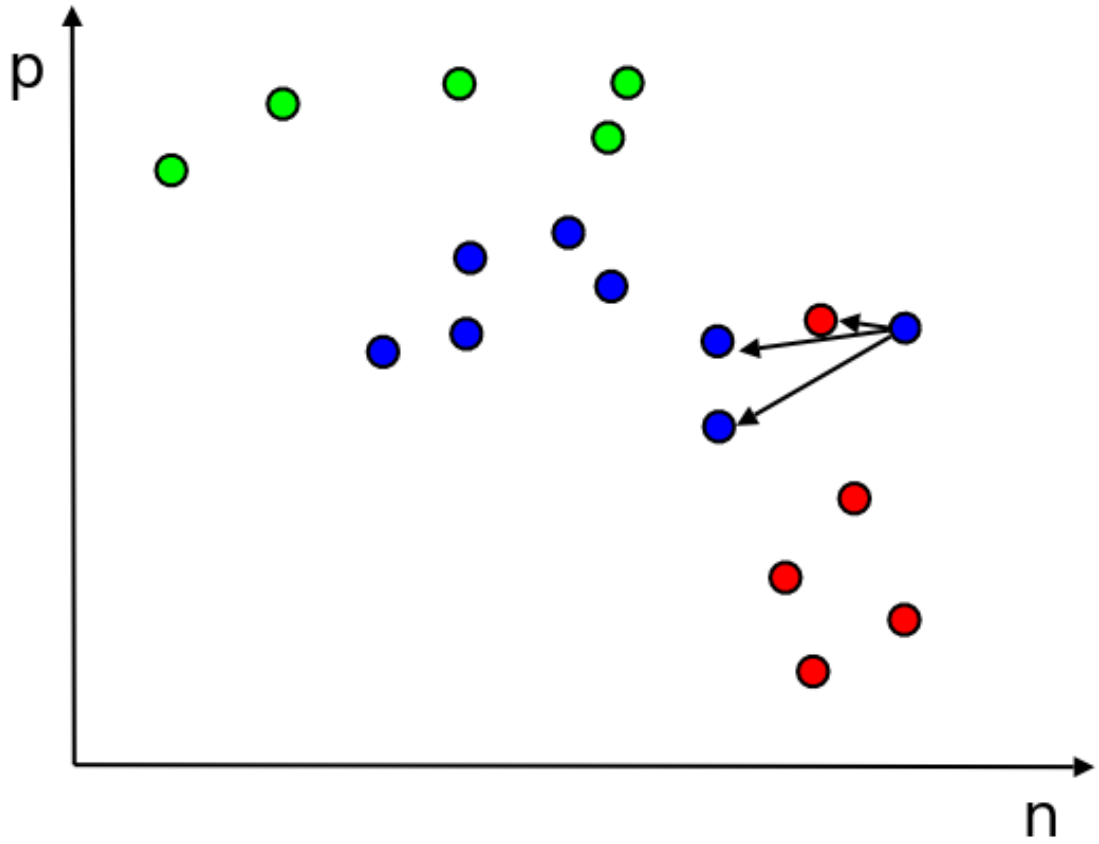
Klasifikace dle nejbližšího souseda

- Zde podobnost např. $(p_1 - p_2)^2 + (n_1 - n_2)^2$
- Zařazujeme do třídy nejpodobnější instance.
- Náchylné k šumu v datech.



Klasifikace dle k nejbližších sousedů

- Zde podobnost např. $(p_1 - p_2)^2 + (n_1 - n_2)^2$
- Zařazujeme do třídy převládající mezi k nejpodobnějšími instancemi.
- S rostoucím k klesá náchylnost k šumu. Nevýhody?



Trénovací a skutečná chyba klasifikátoru

Trénovací chyba

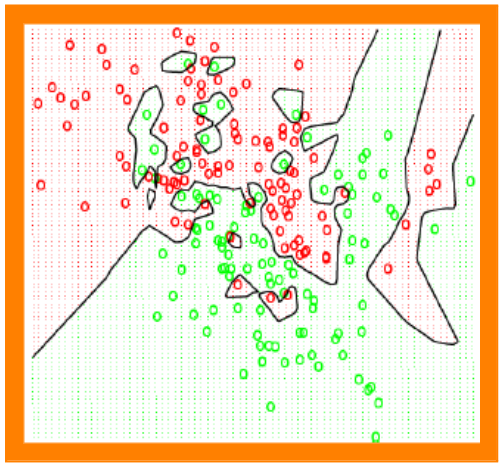
Podíl instancí v datech chybně klasifikovaných klasifikátorem f sestrojeným z těchto dat (tzv. trénovacích dat).

Skutečná chyba

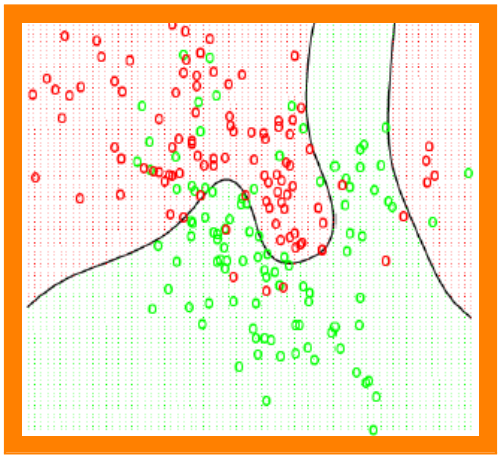
Pravděpodobnost chybné klasifikace f při náhodném výběru $(x \in X, y \in Y)$ dle P_{XY}. Při použití ztrátové funkce L_{01} rovna riziku $R(f)$.

Jaké k je nejlepší?

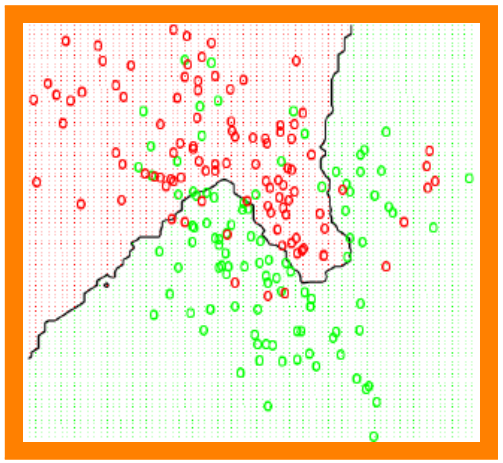
Experiment: generovaná data [Hastie et al.: Elements of Statistical Learning]



$k = 1$
Nulová trénovací chyba, přesto klasifikace odlišná od optimální→



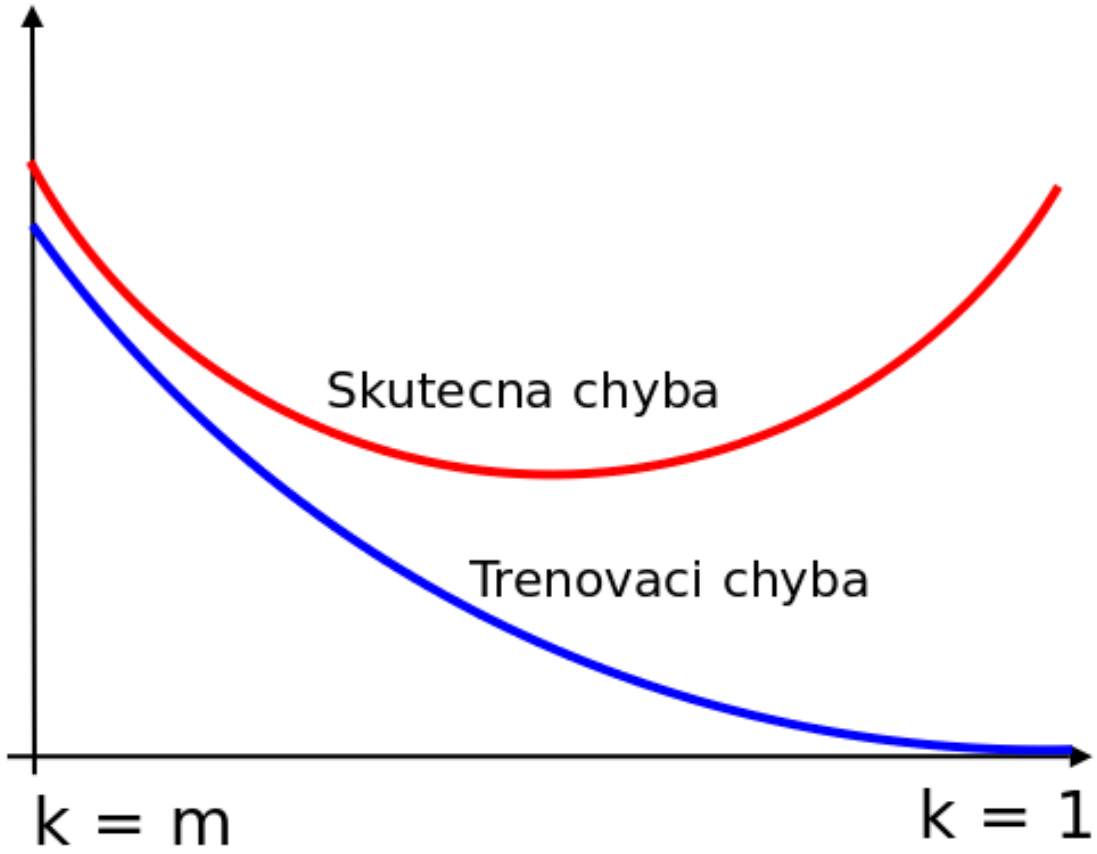
Dle maximální aposteriorní pravděpodobnosti (optimální)



$k = 15$
Malá náchylnost k šumu, přesto klasifikace odlišná od ←optimální

Trénovací vs. skutečná chyba

- Typický průběh chyb pro $k = m$ (počet dat), $m - 1, \ldots 1$
- Trénovací chyba obecně není dobrým odhadem skutečné chyby!



Klasifikace dle etalonů

- Metoda nevyžadující přístup ke všem instancím při klasifikaci.
- Každá třída má **etalon**, tj. instanci s minimální průměrnou vzdáleností ke všem instancím třídy.
  - (vybírána buďto z trénovacích dat, nebo z celého oboru hodnot (p, n))
- Etalony jsou jednoduchým nepravděpodobnostním **modelem** dat.