

Kybernetika a umělá inteligence

2. Pravděpodobnostní rozhodování a klasifikace

Ing. Daniel Novák, Ph.D.
Katedra kybernetiky
ČVUT v Praze, FEL



Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

KYBERNETIKA A UMĚLÁ INTELIGENCE

1. Pravděpodobnostní rozhodování a klasifikace



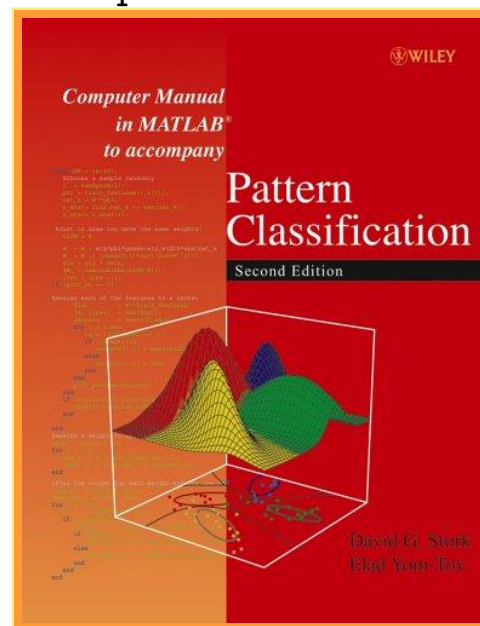
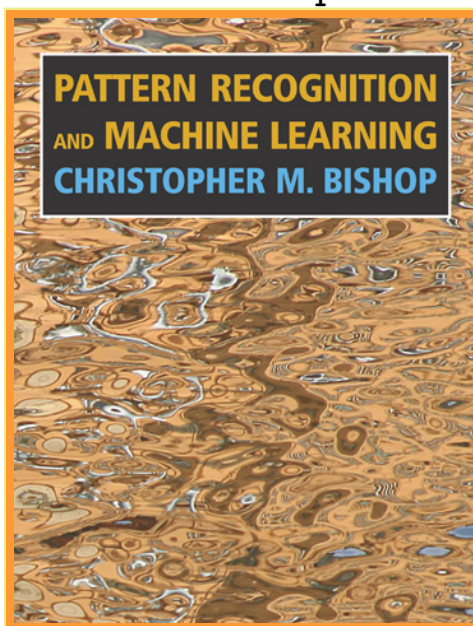
**Gerstnerova laboratoř
katedra kybernetiky
fakulta elektrotechnická
ČVUT v Praze**



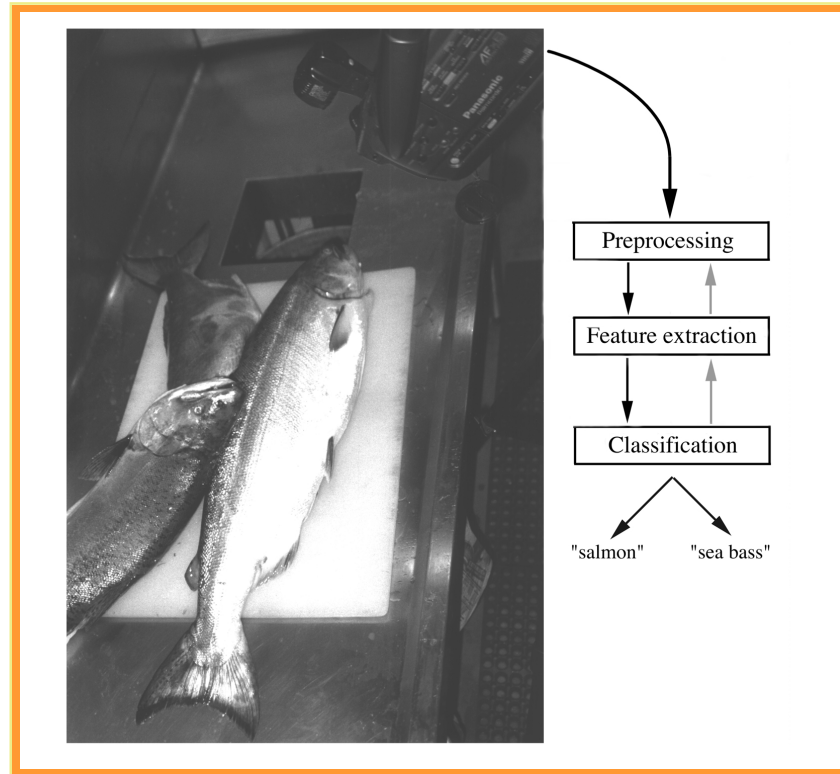
Daniel Novák
Poděkování: Filip Železný

Literatura, dema

- Duda, Hart, Stork: Pattern Classification <http://www.crc.ricoh.com/~stork/DHS.html>
- Ch. Bishop, Pattern Recognition and Machine Learning <http://research.microsoft.com/en-us/um/people/cmbishop/prml/>
- Kotek, Vysoký, Zdráhal: Kybernetika 1990
- Classification toolbox
<http://stuff.mit.edu/afs/sipb.mit.edu/user/arolfe/matlab/>
- Statistical Pattern Recognition Toolbox
<http://cmp.felk.cvut.cz/cmp/software/stprtool/>

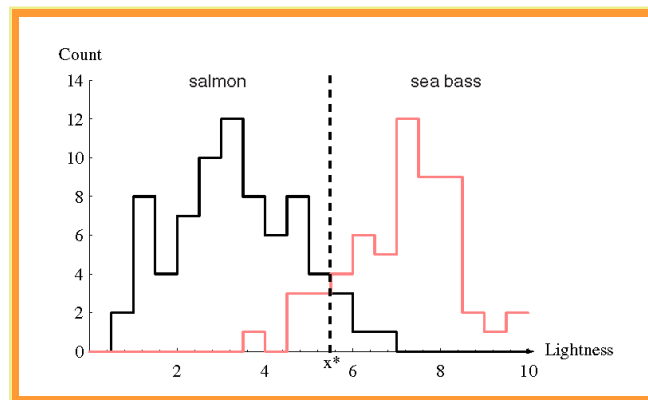
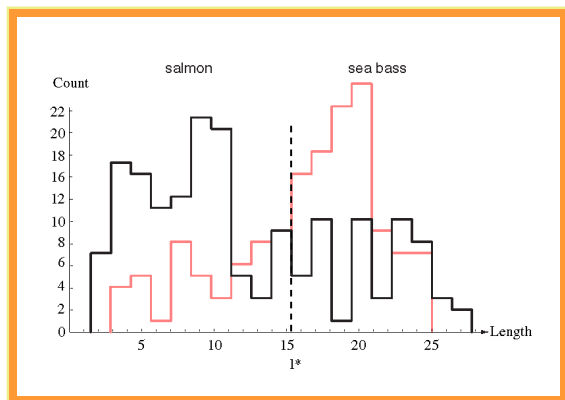


Motivační příklad I [Duda, Hart, Stork: Pattern Classification]

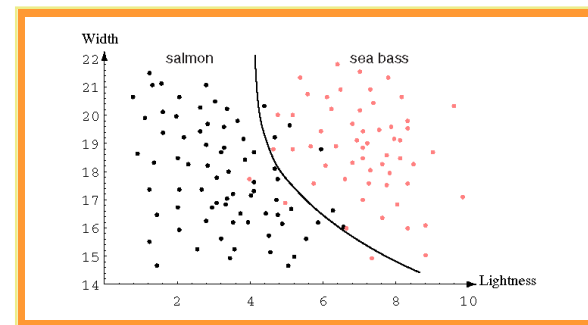
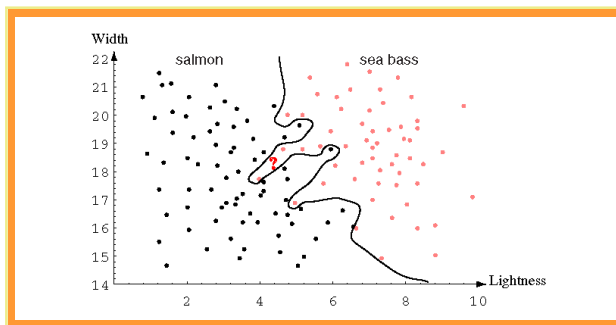
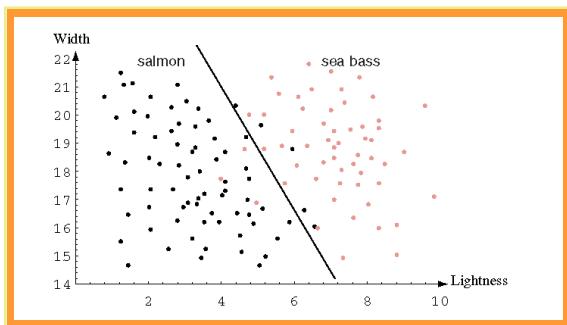


- Továrna na zpracování ryb
- – 2 třídy - Rozeznáváme lososa a okouna pomocí kamery
 - Příznaky - měříme délku, šířku, lesk, průměr oka
- ÚLOHA: JAK KLASIFIKOVAT ???

Motivační příklad II



- Odhadneme rozložení šířky a osvětlení pomocí histogramu
- Klasifikace není bezchybová - histogramy se překrývají
- Zlepšení klasifikace - kombinace příznaků



- Lineární klasifikátor, nejbližší sousedé, kvadratický klasifikátor
- Přeučenost (over-fitting), generalizace, minimalizace chyby

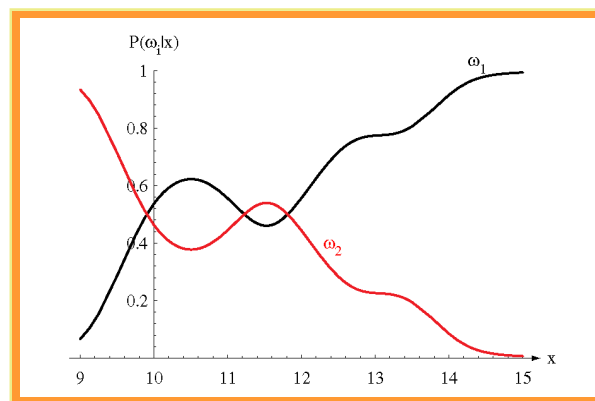
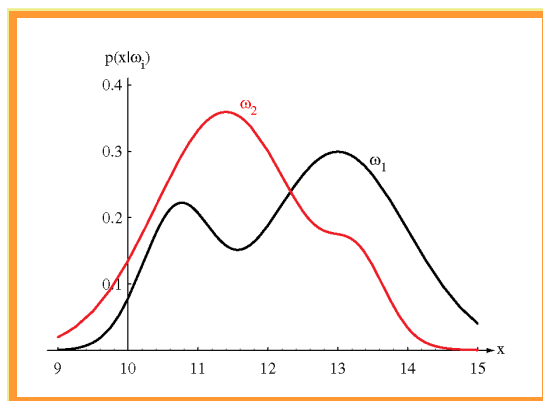
Známe rozložení pravděpodobnosti?

- Ano - Bayesovská klasifikace
- Známe apriorní rozložení třídy $P(s_j)$ a podmíněnou pravděpodobnost $p(x|s_j)$
- Platí $p(s_j, x) = p(x|s_j)p(s_j) = p(s_j|x)p(x)$
- Bayesův vztah

$$p(s_j|x) = \frac{p(x|s_j)p(s_j)}{p(x)}$$

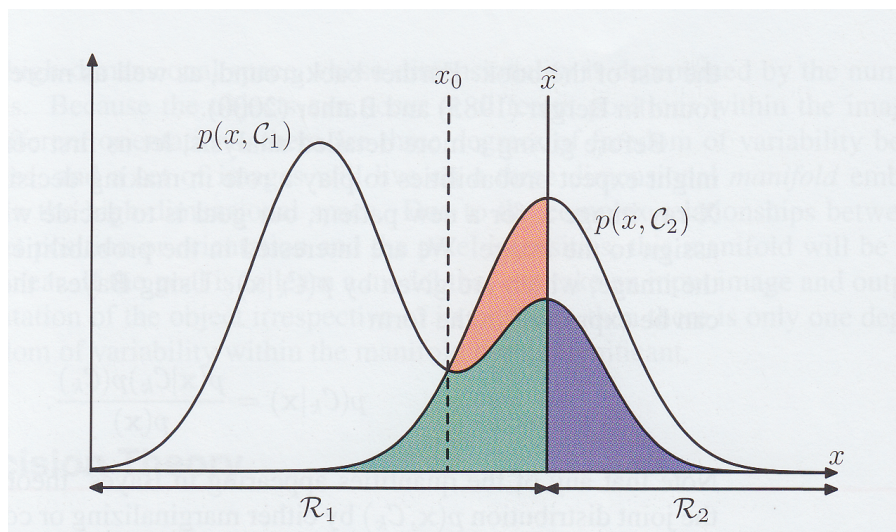
posterior \propto *likelihood* \times *prior*

- Klasifikace $\arg \max_j p(s_j|x)$
- $p(x|s_1) = \frac{1}{3}, p(x|s_1) = \frac{2}{3}$, (v obrázcích níže $s_i = \omega_i$)



Minimalizace chyby - maximum a posteriori pravděpodobnosti [Bishop]

- ilustrativní příklad bez odvozování
- $p(s_1, x) = p(x|s_1)p(s_1), p(s_2, x) = p(x|s_2)p(s_2)$, v obrázku níže ($s_i = C_i$)
- chyba klasifikace: $p(\text{chyba}) = p(x \in R_1, C_2) + p(x \in R_2, C_1)$
- chyba $p(x \in R_1, C_2)$ - červená a zelená oblast - klasifikuji objekty ze třídy C_2 jako C_1
- chyba $p(x \in R_2, C_1)$ - modrá oblast - klasifikuji objekty ze třídy C_1 jako C_2
- minimalizace chyby klasifikace - obě pravděpodobnosti se překrývají v bodě x_0 (zmizí červená plocha)



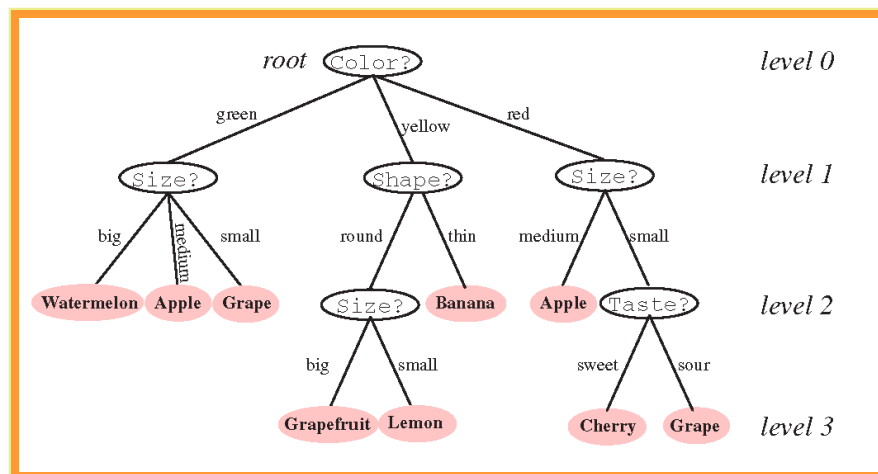
Historická poznámka - Thomas Bayes

- Thomas Bayes - v roce 1736 vydává *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst*
- Řešení tzv. inverzní pravděpodobnosti, např. bílé a černé kuličky pomocí Bayesovského vzorce



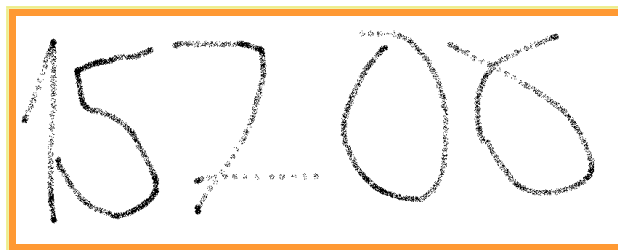
Rozložení pravděpodobnosti je neznámé

- Trénovací a testovací data
- Způsoby klasifikace - tisíce přístupu, již zmíněné, dále rozhodovací stromy, atd.






Formalizace - Rozhodování za neurčitosti

- Důležitou schopností inteligentních systémů je schopnost
 - vybrat co nejlepší rozhodnutí
 - za **nejistých** podmínek (s neurčitostí).
- **Příklad:** Jet z A do B tramvají, nebo metrem?
 - Tramvaj: rychlejší cesta dle jízdního řádu, ale velmi nejisté dodržení.
 - Metro: delší cesta, ale téměř jisté dodržení.
- **Příklad:** kam směřovat dopis s tímto PSČ?



- 15700? 15706? 15200? 15206?
- Jak se **optimálně rozhodnout**?
- Oba příklady lze formalizovat stejným rámcem.
- Zavedení ztrátové funkce - příklad s rakovinou či HIV !

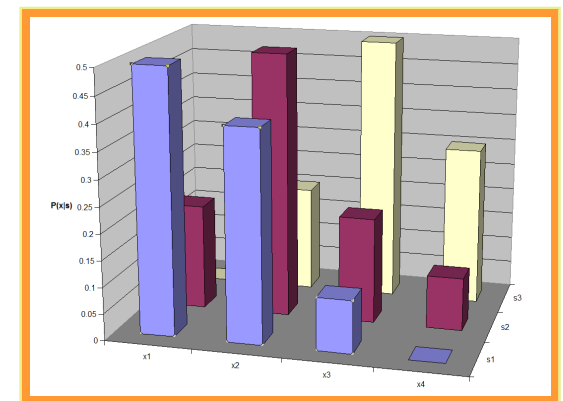
- *Paní Nováková se vrací z práce. Co uvaří pan Novák k večeři?*
- Napadly ho 3 možnosti  **rozhodnutí** (d - **decision**):
 - *nic* ... **neudělat nic** \Rightarrow žádná práce, ale zhorší náladu pí. Novákové.
 - *pizza* ... **ohřát mraženou pizzu** \Rightarrow není pracné, ale neohromí.
 - *n.h.* ... **nadívaná holoubata** \Rightarrow udělá jí radost, ale velmi pracné.
- P. Novák číselně zhodnotí míru nepříjemnosti způsobenou jednotlivými rozhodnutími. Ta závisí na tom, s jakou náladou přijde pí. Nováková domů, což je  **neznámý stav**. Rozlišme tyto možnosti:
 - *dobrá* ... pí. Nováková má **dobrou** náladu.
 - *průměrná* ... pí. Nováková má **průměrnou** náladu.
 - *špatná* ... pí. Nováková má **špatnou** náladu.
- Pro každou z 9 možných situací (3 možná rozhodnutí \times 3 možné stavy) je nepříjemnost dána  **ztrátovou funkcí** $l(d, s)$ (l - **loss**):

$l(s, d)$	$d = nic$	$d = pizza$	$d = n.h.$
$s = dobrá$	0	2	4
$s = průměrná$	5	3	5
$s = špatná$	10	9	6

Příklad (pokračování)

- Neznámý stav - náladu pí. Novákové - zkusí p. Novák odhadnout experimentem: sdělí jí, že ztratil její oblíbený časopis a sleduje její reakci.
- Předpokládá 4 možné reakce:
 - *mírná* ... nic se neděje, časopis najdeme.
 - *podrážděná* ... proč nedáváš věci na své místo?
 - *nasupená* ... proč já si toho Nováka brala?
 - *hrozivá* ... rezignované mlčení
- Reakce je přímo pozorovatelný 🖱️ **příznak** (zde nálady).
- Ze zkušenosti p. Novák dokáže odhadnout pravděpodobnost každé dvojice (reakce x , nálada s), tj. $P(x, s)$.


$P(x, s)$	$x =$ <i>mírná</i>	$x =$ <i>podrážděná</i>	$x =$ <i>nasupená</i>	$x =$ <i>hrozivá</i>
$s = \textit{dobrá}$	0.35	0.28	0.07	0.00
$s = \textit{průměrná}$	0.04	0.10	0.04	0.02
$s = \textit{špatná}$	0.00	0.02	0.05	0.03



Rozhodovací strategie

- **Rozhodovací strategie:** pravidlo pro výběr rozhodnutí na základě pozorovaného příznaku.
- Tj. funkce $d = \delta(x)$.
- Příklady možných strategií p. Nováka:

$\delta(x)$	$x = \text{mírná}$	$x = \text{podrážděná}$	$x = \text{nasupená}$	$x = \text{hrozivá}$
$\delta_1(x) =$	<i>nic</i>	<i>nic</i>	<i>pizza</i>	<i>n.h.</i>
$\delta_2(x) =$	<i>nic</i>	<i>pizza</i>	<i>n.h.</i>	<i>n.h.</i>
$\delta_3(x) =$	<i>n.h.</i>	<i>n.h.</i>	<i>n.h.</i>	<i>n.h.</i>
$\delta_4(x) =$	<i>nic</i>	<i>nic</i>	<i>nic</i>	<i>nic</i>

- Celkem má k dispozici $3^4 = 81$ možných strategií (3 možná rozhodnutí pro každou ze 4 možných hodnot příznaku).
- Jak definovat, která ze dvou strategií je lepší? Obecně: jak strategie uspořádat dle kvality?
- Definujeme  **Bayesovské riziko strategie** jako střední hodnotu ztráty:

$$r(\delta) = \sum_x \sum_s l(s, \delta(x)) P(x, s)$$

Bayesovské kritérium

■ Příklad:

$P(x, s)$	$x =$ <i>mírná</i>	$x =$ <i>podrážděná</i>	$x =$ <i>nasupená</i>	$x =$ <i>hrozivá</i>
$s =$ <i>dobrá</i>	0.35	0.28	0.07	0.00
$s =$ <i>průměrná</i>	0.04	0.10	0.04	0.02
$s =$ <i>špatná</i>	0.00	0.02	0.05	0.03

$l(s, d)$	$d =$ <i>nic</i>	$d =$ <i>pizza</i>	$d =$ <i>n.h.</i>
$s =$ <i>dobrá</i>	0	2	4
$s =$ <i>průměrná</i>	5	3	5
$s =$ <i>špatná</i>	10	9	6

$\delta(x)$	$x =$ <i>mírná</i>	$x =$ <i>podrážděná</i>	$x =$ <i>nasupená</i>	$x =$ <i>hrozivá</i>
$\delta_1(x) =$	<i>nic</i>	<i>nic</i>	<i>pizza</i>	<i>n.h.</i>

$$r(\delta_1) = 0 \cdot 0.35 +$$

- výpočet první strategie $r(\delta_1)$ pro $s=\text{dobrá}, \forall x$

$$r(\delta_1) = 0 \cdot 0.35 + 0 \cdot 0.28 + 2 \cdot 0.07 + 4 \cdot 0 +$$

- $s=\text{průměrná}, \forall x$

$$+ 5 \cdot 0.04 + 5 \cdot 0.1 + 3 \cdot 0.04 + 5 \cdot 0.02 +$$

- $s=\text{špatná}, \forall x$

$$+ 10 \cdot 0 + 10 \cdot 0.02 + 9 \cdot 0.05 + 6 \cdot 0.03 = \mathbf{1.89}$$

- podobně pro třetí strategii $r(\delta_3) = 4 \cdot 0.7 + 5 \cdot 0.2 + 6 \cdot 0.1 = \mathbf{4.4}$

Bayesovské kritérium

■ Příklad:

$P(x, s)$	$x =$ <i>mírná</i>	$x =$ <i>podrážděná</i>	$x =$ <i>nasupená</i>	$x =$ <i>hroživá</i>
$s =$ <i>dobrá</i>	0.35	0.28	0.07	0.00
$s =$ <i>průměrná</i>	0.04	0.10	0.04	0.02
$s =$ <i>špatná</i>	0.00	0.02	0.05	0.03

$l(s, d)$	$d =$ <i>nic</i>	$d =$ <i>pizza</i>	$d =$ <i>n.h.</i>
$s =$ <i>dobrá</i>	0	2	4
$s =$ <i>průměrná</i>	5	3	5
$s =$ <i>špatná</i>	10	9	6

$\delta(x)$	$x =$ <i>mírná</i>	$x =$ <i>podrážděná</i>	$x =$ <i>nasupená</i>	$x =$ <i>hroživá</i>
$\delta_1(x) =$	<i>nic</i>	<i>nic</i>	<i>pizza</i>	<i>n.h.</i>

$r(\delta_1) = 0 * 0.35 + 0 * 0.28 + 2 * 0.07 + \dots$

- ... a tak dále pro celou tabulku $P(x, s)$
- **Bayesovské kritérium:** ze dvou strategií je lepší ta s nižším rizikem.
- Z Bayesovského hlediska je tedy např. lepší δ_1 lepší než δ_3 ($r(\delta_1) = 1.89$, $r(\delta_3) = 4.4$).

Bayesovsky optimální strategie

-  **Bayesovsky optimální strategie** je ta, která minimalizuje Bayesovské riziko. Tj.

$$\delta^* = \arg \min_{\delta} r(\delta)$$

- Protože $P(x, s) = P(s|x)P(x)$, platí

$$r(\delta) = \sum_x \sum_s l(s, \delta(x))P(x, s) = \sum_x \sum_s l(s, \delta(x))P(s|x)P(x)$$

- Protože $P(x)$ nezávisí na s :

$$r(\delta) = \sum_x P(x) \underbrace{\sum_s l(s, \delta(x))P(s|x)}_{\text{riziko podmíněné pozorováním } x}$$

- Optimální strategii tedy můžeme dostat minimalizací tohoto *podmíněného rizika* zvlášť pro jednotlivá pozorování x :

$$\delta^*(x) = \arg \min_d \sum_s l(s, d)P(s|x)$$

Co když neznáme $P(x, s)$?

- Sdružená pravděpodobnost - potřebujeme znát všechna možná měření
- U reálných problémů nepraktické, někdy nemožné - v případě rozpoznávání písmen, 10 x 10 pixelů, 256 hodnot intensity, 256^{100} měření
- Z definice podmíněné pravděpodobnosti

$$P(x, s) = P(s|x)P(x) = P(x|s)P(s)$$

- **Případ 1:** Známe pouze $P(s|x)$, ale ne $P(x)$ (pravděpodobnosti pozorování).
 - Bayesovsky optimální strategii umíme stále vypočítat - viz předešlou stranu.
 - Dále příklad na cvičeních: roztřídění mincí podle jejich hmotnosti získané měření

Kritérium MiniMax

- **Případ 2:** Známe pouze $P(x|s)$, ale ne $P(s)$ (pravděpodobnosti stavů).


– Bayesovsky optimální strategii již **nemůžeme** určit, protože nespočítáme

$$r(\delta) = \sum_s P(s) \sum_x l(s, \delta(x)) P(x|s)$$

Nahradíme-li první sumu *maximem*, definujeme *maximální* riziko, které můžeme spočítat.

$$R(\delta) = \max_s \sum_x l(s, \delta(x)) P(x|s)$$

– Tato veličina maximalizuje mezi všemi riziky *podmíněnými* jednotlivými *stavy*.

-  **MiniMaxové** kritérium: vybírá strategii, která **minimalizuje maximální riziko**
 $R(\delta)$

$$\delta^* = \arg \min_{\delta} R(\delta)$$

Příklad - Kritérium MiniMax

- Příklad výpočtu $R(\delta_1)$
- Co když p. Novák ví, že p. Nováková má *obvykle dobrou náladu*? Obecněji: ví, jak jsou její jednotlivé nálady pravděpodobné, tj. zná rozložení $P(s)$. Např:

	$s = \text{dobrá}$	$s = \text{průměrná}$	$s = \text{špatná}$
$P(s) =$	0.7	0.2	0.1

- Můžeme tedy spočítat podmíněné rozložení $P(x|s)$

$$P(x|s) = \frac{P(x, s)}{P(s)}$$

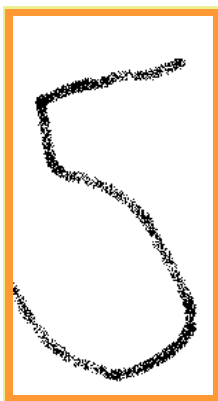
$P(x s)$	$x =$ <i>mírná</i>	$x =$ <i>podrážděná</i>	$x =$ <i>nasupená</i>	$x =$ <i>hrozivá</i>
$s = \text{dobrá}$	0.5	0.4	0.1	0.00
$s = \text{průměrná}$	0.2	0.5	0.2	0.1
$s = \text{špatná}$	0.00	0.2	0.5	0.3

Příklad - Kritérium MiniMax (pokračování)

- Pro $s = \text{dobrá}$: $\sum_x l(s, \delta(x))P(x|\text{dobrá}) =$
 $l(\text{dobrá}, \delta_1(\text{mírná})) \cdot P(\text{mírná}|\text{dobrá}) + l(\text{dobrá}, \delta_1(\text{podrážděná})) \cdot P(\text{podrážděná}|\text{dobrá})$
 $+ l(\text{dobrá}, \delta_1(\text{nasupená})) \cdot P(\text{nasupená}|\text{dobrá}) + l(\text{dobrá}, \delta_1(\text{hrozivá})) \cdot P(\text{hrozivá}|\text{dobrá})$
 $= l(\text{dobrá}, \text{nic}) \cdot 0.5 + l(\text{dobrá}, \text{nic}) \cdot 0.4 + l(\text{dobrá}, \text{pizza}) \cdot 0.1 + l(\text{dobrá}, \text{n.h.}) \cdot 0$
 $= 0 \cdot 0.5 + 0 \cdot 0.4 + 2 \cdot 0.1 + 4 \cdot 0 = 0.2$
- $\sum_x l(s, \delta(x))P(x|\text{průměrná}) = 4.6$
- $\sum_x l(s, \delta(x))P(x|\text{špatná}) = 8.3$
- **Maximální riziko** strategie δ_1 (přes všechny možné stavy) je tedy 8.3.
- Podobně: maximální riziko strategie δ_3 je 7.85.
- Tedy podle MiniMaxu je δ_3 lepší než δ_1 (obráceně než u Bayesovského kritéria!)
- Pro nalezení optimální strategie bychom v aktuálním příkladě museli spočítat max. rizika všech 81 možných strategií.

Příznakové rozpoznávání

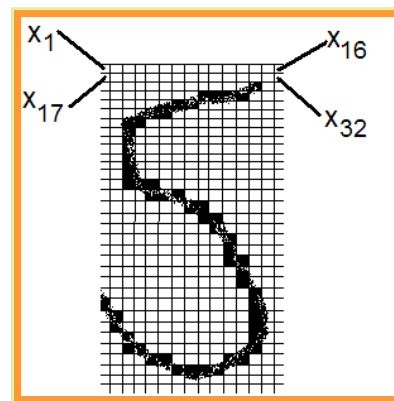
- Systémy pro rozpoznávání. **Příklad** úlohy:



O jakou jde číslici?



Lze formulovat jako
úlohu
statistického
rozhodování



Příznak = vektor hodnot pixelů.

- **Příznakové rozpoznávání** číslic: **klasifikace** do jedné ze **tříd** $0 \dots 9$ na základě vektoru hodnot pixelů.
- Klasifikace je speciálním případem statistického rozhodování:
 - Příznakový vektor $\vec{x} = (x_1, x_2, \dots)$: hodnoty pixelů č. 1, 2, \dots
 - Množina stavů \mathcal{S} = množina rozhodnutí $\mathcal{D} = \{0, 1, \dots, 9\}$.
 - Stav = skutečná třída, Rozhodnutí = rozpoznaná třída.
- video - *synth_2LP*

MAP klasifikace

- Častá (i když ne nutná) volba ztrátové funkce při klasifikaci:

$$l_{01}(d, s) = \begin{cases} 0, & d = s \\ 1, & d \neq s \end{cases}$$

- Pokud zvolíme l_{01}
 - jde o tzv. MAP (maximum a posteriori) klasifikaci
 - Bayesovské riziko = střední **chyba** klasifikace.
- Optimální klasifikace při příznaku \vec{x} (Bayesovsky optimální strategie):

$$\begin{aligned} \delta^*(\vec{x}) &= \arg \min_d \sum_s \underbrace{l(d, s)}_{0 \text{ pokud } d=s} P(s|\vec{x}) = \\ &= \arg \min_d [1 - P(d|\vec{x})] = \arg \max_d P(d|\vec{x}) \end{aligned}$$

- Volíme tedy nejpravděpodobnější třídu pro danou hodnotu příznakového vektoru - viz naše klasifikace lososa a okouna

Strojové učení

- Obvykle ale **není známo** rozdělení $P(s|\vec{x})$.
- Je třeba **odhadnout** (“naučit se”) z již klasifikovaných příkladů, tzv. **trénovacích dat**.
- Uvažujme **diskrétní \vec{x}** . **Trénovací data (příklady)**: $(\vec{x}_1, s_1), (\vec{x}_2, s_2), \dots, (\vec{x}_l, s_l)$.
 - Tzv. *i.i.d. multimnožina* (i.i.d = “independent, identically distributed”)
 - Každý příklad (\vec{x}_i, s_i) je náhodně vybrán nezávisle na ostatních příkladech z rozdělení $P(x, s)$.

- Bez jakékoliv znalosti o tomto rozdělení se můžeme uchýlit k tzv. *neparametrickému* odhadu:

$$P(s|\vec{x}) \approx \frac{\text{počet příkladů v nichž } \vec{x}_i = \vec{x} \text{ a } s_i = s}{\text{počet příkladů v nichž } \vec{x}_i = \vec{x}}$$

- Tento přístup naráží na zásadní překážky:
 - Počet příkladů l postačující ke spolehlivému odhadu $P(s|\vec{x})$ roste **exponenciálně** s počtem složek vektoru \vec{x} .
 - tj. např. s rozlišením (počtem pixelů) v rozpoznávaných obrazcích.
 - “prokletí kombinatorické exploze”. Reálné úlohy: jmenovatel často nulový!
 - Bayesovská klasifikace: horní limit kvality klasifikace, v praxi obvykle nedosažitelný.

- Lze též využít vztahu:

$$P(s|\vec{x}) = \frac{P(\vec{x}|s)P(s)}{P(\vec{x})}$$

- **Odhad** $P(\vec{x}|s)$: analogicky jako odhad $P(s|\vec{x})$.
- **Odhad** $P(s)$: jako relativní četnost jednotlivých tříd s v trénovacích datech, tj.





$$P(s) \approx \frac{\text{počet příkladů třídy } s}{l}$$

- Tento přístup sám o sobě neřeší problém množství dat potřebných k odhadu pravděpodobností.
- Ale umožňuje ho “řešit” nepřímo:
 1. Hodnoty $P(s)$ jsou často **známy a není nutno je odhadovat**.
Příklad: při rozpoznávání 1. číslice PSČ je nejčastější číslice 1, např. $P(1) = 0.6$.
Takto je do klasifikace zapojena **apriorní znalost** o pravděpodobnostech tříd.
 $P(s)$... ‘**apriorní pravděpodobnost**’.
 2. Přístup umožňuje formulovat zjednodušenou, tzv. naivní Bayesovskou klasifikaci, v níž nemusíme odhadovat $P(\vec{x}|s)$, ale pouze $P(x(1)|s), P(x(2)|s), \dots$

Předpoklad 1: Nezávislost složek příznaku

- Ve **výjimečném** případě **statistické nezávislosti** jednotlivých příznakových složek $x^{(i)}$ v rámci každé třídy s platí


$$P(\vec{x}|s) = P(x(1)|s) \cdot P(x(2)|s) \cdot \dots$$

- Stačí tedy odhadnout $P(s)$ pro každé s a $P(x^{(i)}|s)$ zvlášť pro každé i a s .
 - Např: $P(x(3)|8) \approx$ podíl případů číslice 8 s rozsvíceným 3. pixelem.
 -  Žádná kombinatorická exploze (pouze jednosložkové pravděpodobnosti).
- V praxi: nezávislost se často předpokládá, i když neplatí, příp. platí přibližně.
 - Potom jde o tzv.  **Naivní Bayesovskou klasifikaci**.
- Nezávislost mezi příznakovými složkami je jen **jedním z možných předpokladů**, jehož splnění vede k zabránění kombinatorické explozi.
- Alternativní předpoklady jsou např.:
 - *Podobné objekty patří do stejné třídy*  klasifikace dle **nejbližších sousedů**.
 - $P(s|\vec{x})$ *náleží do určité omezené třídy rozdělení, hledáme pouze jeho parametry*
 - *Třída je plně určena lineární kombinací složek příznaku*  klasifikace dle **lineárního modelu**.

Předpoklad 2: Podobnost objektů v dané třídě

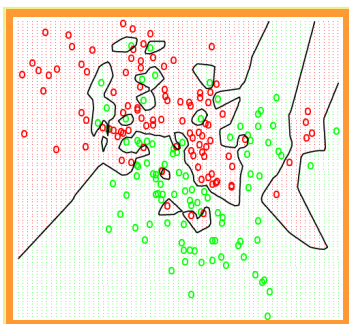
- *Podobnost* chápeme jako malou *vzdálenost* v prostoru příznakových hodnot.
- Funkce měřící vzdálenost dvou příznakových vektorů, tzv. **metrika**: $\rho : X \times X \rightarrow \mathbb{R}^+ \cup \{0\}$ taková, že $\forall x, y, z: \rho(x, x) = 0, \rho(x, y) = \rho(y, x), \rho(x, z) \leq \rho(x, y) + \rho(y, z)$. **Příklad:**
 - **Euklidovská metrika** pro vektory \vec{x}_1, \vec{x}_2 se reálnými složkami $x_1(i)$ resp. $x_2(i)$:

$$\rho_E(\vec{x}_1, \vec{x}_2) = \sqrt{\sum_i (x_1(i) - x_2(i))^2}$$

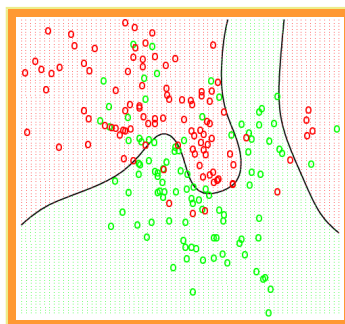
- Jsou-li složky binární (z $\{0, 1\}$), tak $\rho_E(\vec{x}_1, \vec{x}_2)^2$ je počet složek, v nichž se \vec{x}_1 liší od \vec{x}_2 - tzv. **Hammingova metrika**.
-  **Klasifikace dle k nejbližších sousedů** (k -nearest neighbor classification, k -NN).
- Zadáno:
 - $k \in \mathbb{N}$
 - trénovací příklady: $(\vec{x}_1, s_1), (\vec{x}_2, s_2), \dots, (\vec{x}_l, s_l)$
 - metrika $\rho : X \times X \rightarrow \mathbb{R}$
 - neklasifikovaný objekt s příznakem \vec{x} .
- Úloha: klasifikovat \vec{x}
- Postup: z trénovacích příkladů vyber k nejbližších k \vec{x} vzhledem k metrice ρ . Třída, které mezi nimi převládá, budiž třídou \vec{x} .

Flexibilita klasifikace

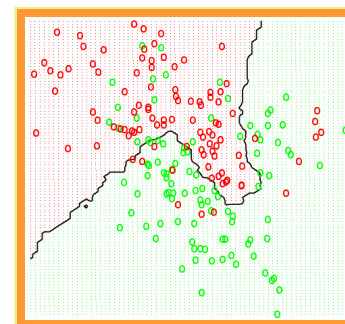
- Jak volit k ? Odpověď není jednoznačná.
- **Příklad:** Uvažujme trénovací data se dvěma třídami (červená/zelená) a šumem (některé s_i chybné). Značky - trénovací data, křivka - hranice klasifikace:



$k = 1$: Dobré přizpůsobení trénovacím datům. Velká citlivost k šumu.




Bayesovská klasifikace: Méně flexibilní než 1-nn, více než 15-nn.



$k = 15$: Špatné přizpůsobení trénovacím datům. Malá citlivost k šumu.

- Vzpomeňte: Bayesovská klasifikace δ^* má nejnižší možné Bayesovské riziko $r(\delta^*)$. Pozn.: Znárodněná Bayesovská vychází z přesných pravděpodobností $P(s|\vec{x})$, které jsou pro klasifikační algoritmus neznámé!
- Pozorování: příliš velká flexibilita (malé k) i příliš malá flexibilita (velké k) vedou ke klasifikátorům značně odlišným od Bayesovského, tedy ke zvyšování středního rizika $r(\delta)$.

Trénovací chyba a riziko

- Bayesovské riziko $r(\delta) =$ limita relativní četnosti nesprávných klasifikací (s rostoucím vzorkem dat).
- Definujme **trénovací chybu** $r_T(\delta)$ jako relativní četnost nesprávně klasifikovaných příkladů na datech, kde je klasifikátor δ učen.
- Je $r_T(\delta)$ dobrým odhadem skutečného středního rizika $r(\delta)$?
- Příklad: 1-nn není dobrý klasifikátor (viz minulou stranu), přestože správně klasifikuje *všechny trénovací příklady*, tj. má trénovací chybu 0.
-  Trénovací chyba tedy není dobrým odhadem středního rizika. Pro jeho odhad je třeba
 - mít k dispozici **trénovací množinu** $(\vec{x}_1, s_1), \dots, (\vec{x}_l, s_l)$ a nezávislou **testovací množinu** $(\vec{x}_{l+1}, s_{l+1}), \dots, (\vec{x}_{l+m}, s_{l+m})$
 - (může vzniknout rozdělením původních trénovacích dat např. v poměru 75% a 25%).
 - klasifikátor sestrojít na základě trénovací množiny
 - riziko tohoto klasifikátoru spočítat na testovací množině.
- Relativní četnost chyb na testovací množině je **nevychýleným** odhadem skutečného středního rizika. (Pozor: nevychýlený neznamená přesný!)