Kybernetika a umělá inteligence

---

2. Strojové učení

Gerstnerova laboratoř
katedra kybernetiky
České vysoké učení technické v Praze

Daniel Novák
Poděkování: Filip Železný

Shrnutí minulé přednášky

- Rozhodování s neurčitostí, zavedení ztrátové funkce vede na minimalizaci Bayesovského rizika, ilustativní příklad na sdružené pravděpodobnosti - p.Novák a příprava večeře
- Klasifikaci je speciální případ rozhodování s neurčitostí
- V případě ztrátové funkce $l_{01}(d, s)$ Bayesovská klasifikace: klasifikuj do třídy s, pro kterou platí $\arg\max_s P(s|\vec{x})$.
- Bayesovský klasifikátor má nejmenší riziko (klasifikační chybu) mezi všemi klasifikátory.
- Klasifikace je založena na znalosti $P(s|\vec{x})$.
- Často neznáme $P(s|\vec{x})$ nebo $P(\vec{x}, s)$, pouze změřená i.i.d data. Je velmi težké odhadnout $P(\vec{x}, s)$ s rostoucí dimenzí příznaků $\vec{x}$.
- Prokletí dimenzionality není případem, kdy $\vec{x}$ jsou statisticky nezávislé.
- Často známe tvar pravděpodobnosti, $P(\vec{x}|s)$, z dat musíme odhadnout neznámý parametr s. Tedy s je funkcí x [zavedeme metodu maximalní věrohodnosti].
- Pro klasifikaci můžeme použít i neparametrické metody - metoda nejmenšího souseda.
- pojmy generalizace, přeučení, výběr parametrů klasifikátoru

Statistické rozhodování: shrnutí

- **Zadány**:
  - Množina možných **stavů**: $\mathcal{S}$
  - Množina možných **rozhodnutí**: $\mathcal{D}$
  - **Ztrátová funkce**: zobrazení $l : \mathcal{D} \times \mathcal{S} \to \Re$ (reálná čísla)
  - Množina možných hodnot **příznaku** $\mathcal{X}$
  - Pravděpodobnostní rozložení příznaku za daného stavu $P(x|s)$, $x \in \mathcal{X}, s \in \mathcal{S}$.
- **Definujeme**:
  - **Strategie**: zobrazení $\delta : \mathcal{X} \to \mathcal{D}$
  - **Riziko strategie** δ při stavu $s \in \mathcal{S}$: $R(\delta, s) = \sum_x l(s, \delta(x))P(x|s)$
- **MiniMaxová úloha**:
  - Dále zadána: množina přípustných strategií Δ.
  - Úloha: nalézt optimální strategii $\delta^* = \arg\min_{\delta \in \Delta} max_{s \in \mathcal{S}} R(\delta, s)$
- **Bayesovská úloha**:
  - Dále zadáno: pravděpodobnostní rozdělení stavů $P(s)$, $s \in \mathcal{S}$.
  - Dále definujeme: **střední riziko strategie** δ: $r(\delta) = \sum_s R(\delta, s)P(s)$
  - Úloha: nalézt optimální strategii $\delta^* = \arg\min_{\delta \in \Delta} r(\delta)$
  - Řešení: $\delta^*(x) = \arg\min_d \sum_s l(d, s)P(s|x)$

Známá forma pravděpodobnostního rozdělení

- Předpokládáme, že $P(\vec{x}|s)$ má známou formu a z trénovacích dat je třeba odhadnout jen jeho parametry. Tedy: $P(\vec{x}|s) = \phi(\vec{x}, s, \theta_1, \theta_2, \dots)$, kde ϕ je známá funkce a θ_i jsou neznámé parametry.
- Př.: normální hustota pravděpodobnosti, $\phi(\vec{x}, s, \theta_1, \theta_2, \dots) = N(x, \mu, \sigma)$, $\theta_1 = \mu$, $\theta_2 = \sigma$:

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

- *Ústřední limitní věta*: Součet mnoha náhodnodných proměnných s jakkýmikoliv rozděleními se řídí rozdělením, které se blíží normálnímu.
- Uvažujeme-li jediný **reálný skalární** příznak, předpoklad *normálního rozdělení* zní, že pro každou třídu s je podmíněná hustota x:

$$p(x|s) = N(x, \mu_s, \sigma_s)$$

- Parametry rozdělení se často zapisují v podmínkové části:

$$p(x|s, \mu_s, \sigma_s) = N(x, \mu_s, \sigma_s)$$

Normální rozdělení

- Též "**Gaussovské**"

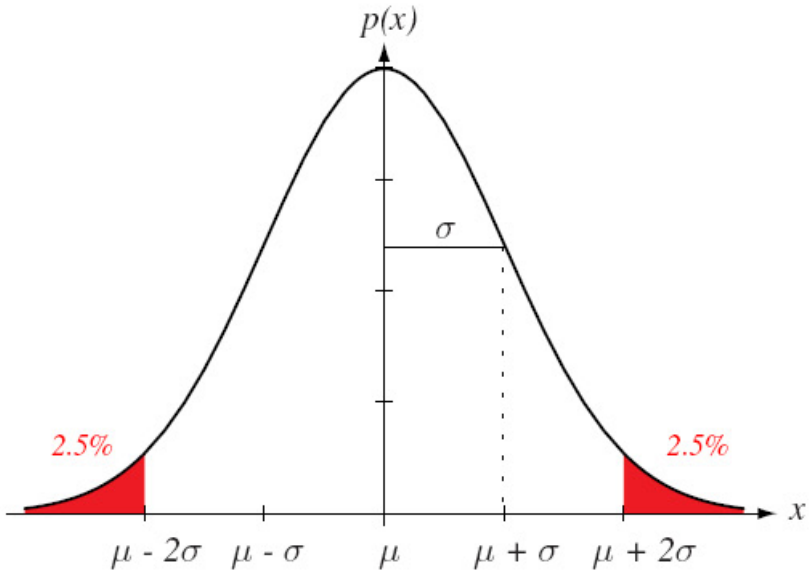


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Diskriminační funkce

- Za předpokladu normálního rozdělení postupujeme při Bayesovské klasifikace následovně:

$$\arg\max_s p(s|x,\mu_s,\sigma_s) = \arg\max_s \frac{p(x|s,\mu_s,\sigma_s)p(s)}{p(x)} = \arg\max_s p(x|s,\mu_s,\sigma_s)p(s)$$

$$= \arg\max_s \frac{1}{\sigma_s\sqrt{2\pi}} \exp\frac{-(x-\mu_s)^2}{2\sigma_s^2} \cdot p(s) = \arg\max_s \ln\left(\frac{1}{\sigma_s\sqrt{2\pi}} \exp\frac{-(x-\mu_s)^2}{2\sigma_s^2} \cdot p(s)\right)$$

$$= \arg\max_s \left(-\frac{1}{2}\ln\sigma_s^2 \underbrace{-\frac{1}{2}\ln 2\pi}_{\text{lze vypustit}} + \frac{-(x-\mu_s)^2}{2\sigma_s^2} + \ln p(s)\right)$$

$$= \arg\max_s \left(-\frac{1}{2}\ln\sigma_s^2 - \frac{1}{2\sigma_s^2}\left(x^2 - 2x\mu_s + \mu_s^2\right) + \ln p(s)\right) = \arg\max_s a_s x^2 + b_s x + c_s$$

kde

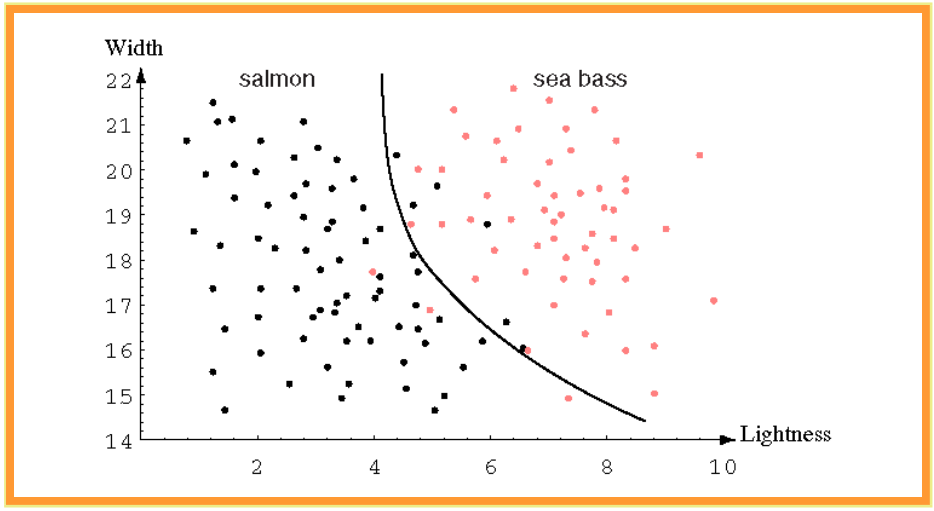
$$a_s = -\tfrac{1}{2\sigma_s^2} \qquad b_s = \tfrac{\mu_s}{\sigma_s^2} \qquad c_s = -\tfrac{1}{2}\ln\sigma_s^2 - \tfrac{\mu_s^2}{2\sigma_s^2} + \ln p(s)$$

- Tímto je definována kvadratická **diskriminační funkce pro každé** $s \in S$,

$$\boxed{g_s(x) = a_s x^2 + b_s x + c_s}$$

Použití: pro zadané x, **klasifikuj do** $\max_s g_s(x)$.

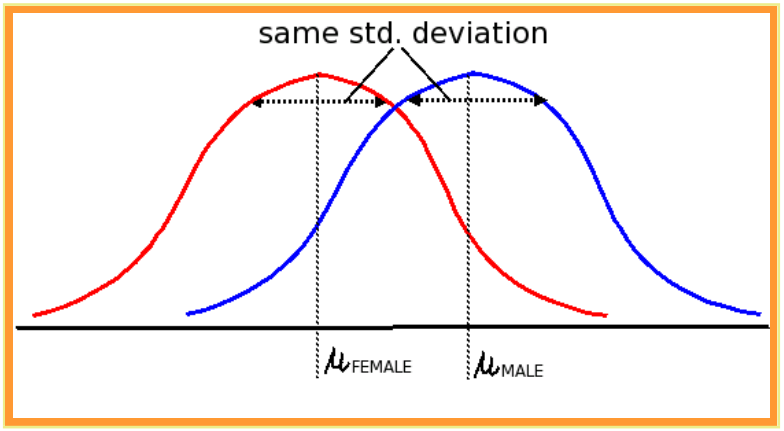
Kvadratická diskriminace



- (Duda, Hart, Stork: Pattern Classification)
- Dvě třídy: černá ($s1$) a červená ($s2$).
- Černá křivka vede body, v nichž $g_{s1}(\vec{x}) = g_{s2}(\vec{x})$. Tato křivka je hranicí rozhodnutí mezi třídami $s1$ a $s2$.

Normální rozdělení, stejné σ pro všechny třídy

- **Jednoduchý případ:** stejné směrodatné odchylky σ. Příklad: $s = \{\textit{muž}, \textit{žena}\}$, $x =$ výška.



- Protože $\forall s\ \sigma_s = \sigma$, je možné další zjednodušení

$$\max_s P\left(s|x, \mu_s, \sigma\right) = \max_s \left(\underbrace{-\frac{x^2}{2\sigma^2}}_{\textit{lze vypustit}} + \frac{1}{2\sigma^2}\left(2x\mu_s - \mu_s^2\right) + \ln P(s) \right) = \max_s \left(b_s \cdot x + c_s\right)$$

kde $b_s = \frac{\mu_s}{\sigma^2}$ and $c_s = -\frac{\mu_s^2}{2\sigma^2} + \ln P(s)$.

- Nyní je diskriminační funkce **lineární**:

$$g_s(x) = b_s x + c_s$$

Mnoharozměrný případ [bonusová látka]

- Mnoharozměrný případ ($\vec{x}$ je n-rozměrný reálný vektor, $\vec{x} \in \Re^n$)

$$N(\vec{x}, \vec{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{\Sigma})}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^t |\mathbf{\Sigma}| (\vec{x}-\vec{\mu})\right]$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{2,1} & \dots & \sigma_{n,1} \\ \sigma_{1,2} & \sigma_{2,2} & \dots & \sigma_{n,2} \\ \vdots & \vdots & & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \dots & \sigma_{n,n} \end{bmatrix} \dots \textbf{kovariační} \text{ matice:} \quad \begin{aligned} \sigma_{i,j} &= \overline{(x_i - \mu_i)(x_j - \mu_j)} \\ \sigma_{i,i} &= \sigma_i^2 \end{aligned}$$

- Předpoklad normálního rozdělení: $p(\vec{x}|s, \vec{\mu}, \mathbf{\Sigma}) = N(\vec{x}, \vec{\mu}_s, \mathbf{\Sigma}_s)$ pro každou třídu s.
- **Kvadratická** diskriminační funkce $\boxed{g_s(x) = \vec{x}^t \boldsymbol{A}_s \vec{x} + \vec{b}_s^t \vec{x} + c_s}$ kde

$$\boldsymbol{A}_s = -\tfrac{1}{2}\mathbf{\Sigma}_s^{-1} \qquad \vec{b}_s = \mathbf{\Sigma}_s^{-1}\mu_s \qquad c_s = -\tfrac{1}{2}\mu_s^t \mathbf{\Sigma}_s^{-1}\mu_s - \tfrac{1}{2}\ln\det(\mathbf{\Sigma}_s) + \ln P(s)$$

- Zvláštní případ:$\forall s\ \Sigma_s = \Sigma$: **Lineární** diskriminační funkce $\boxed{g_s(x) = \vec{b}_s^t \vec{x} + c_s}$ kde

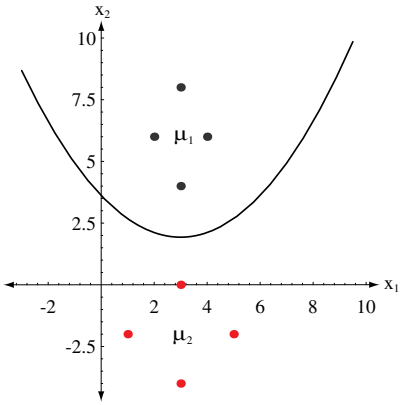
$$\vec{b}_s = \mathbf{\Sigma}_s^{-1}\mu_s \qquad c_s = -\tfrac{1}{2}\mu_s^t \mathbf{\Sigma}_s^{-1}\mu_s + \ln P(s)$$

Příklad hranice

- hranice pro dvourozměrný případ

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \Sigma_1^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} a \quad \Sigma_2^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

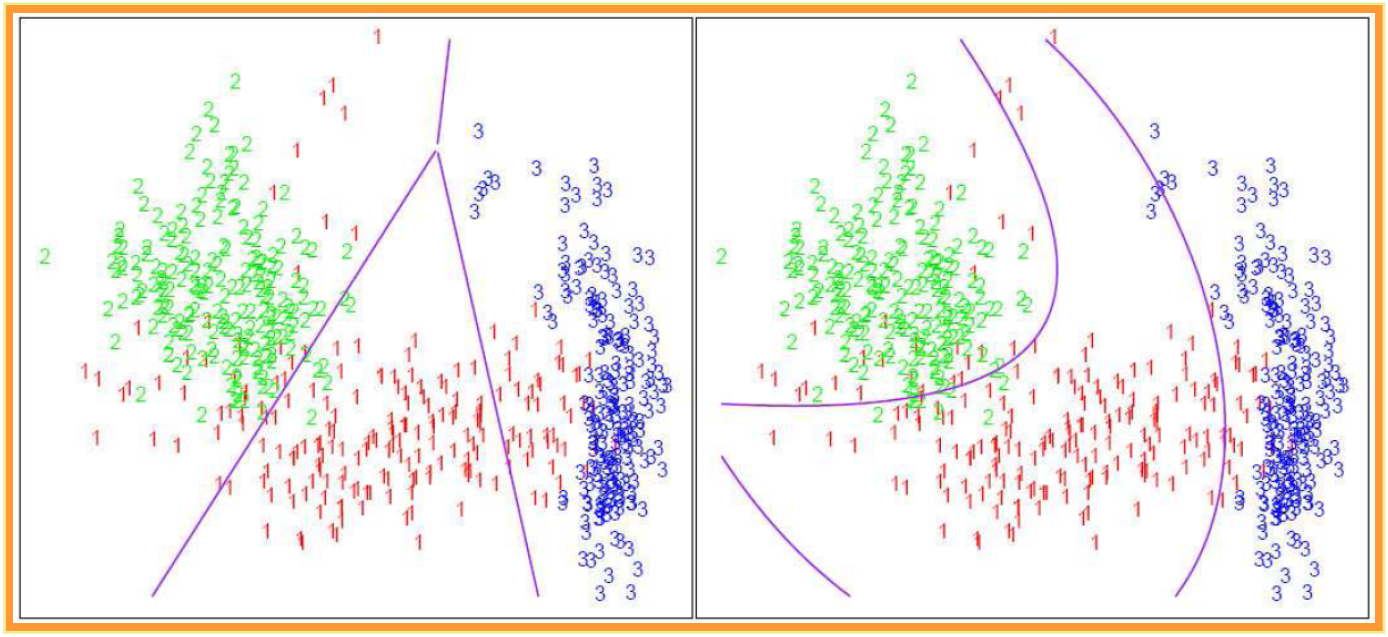
- apriorní pravděpodobnosti $p(s_1) = p(s_2) = 0.5$
- hranici dostaneme, když $g_1(\vec{x}) = g_2(\vec{x}) = -\frac{x_1^2}{4} - \frac{x_2^2}{4} + \frac{3x_1}{2} - x_2 - 3.12 = -x1^2 - \frac{x_2^2}{4} + 6x_1 + 3x_2 - 18.69$
- tedy $x_2 = 0.187x_1^2 - 1.125x_1 + 3.8925$



- **Kvadratická** diskriminační funkce $g_s(x) = \vec{x}^t \boldsymbol{A}_s \vec{x} + \vec{b}_s^t x + c_s$ kde

$$\boldsymbol{A}_s = -\tfrac{1}{2}\boldsymbol{\Sigma}_s^{-1} \qquad \vec{b}_s = \boldsymbol{\Sigma}_s^{-1}\mu_s \qquad c_s = -\tfrac{1}{2}\mu_s^t\boldsymbol{\Sigma}_s^{-1}\mu_s - \tfrac{1}{2}\ln\det(\boldsymbol{\Sigma}_s) + \ln P(s)$$

Lineární vs. Kvadratická diskriminace



- Vlevo: lineární diskriminace v $\Re^2$. Body x, kde $s' = \arg\max_s g_s(\vec{x})$ pro dané s' tvoří konvexní oblasti s hranicemi po částech lineárními.
- Vpravo: kvadratická diskriminace v $\Re^2$. Body x, kde $s' = \arg\max_s g_s(\vec{x})$ pro dané s' tvoří konvexní oblasti s hranicemi po částech kvadratickými.

Carl Friedrich Gauss (1777-1855)

- Matematický genius, podle některých autorů patří mezi tři největší matematiky - Archimedes, Newton, Gauss
- „Matematika je královnou vědy a teorie čísel je královnou matematiky.“
- O jeho schopnostech kolují mnohé historky, např. v 7 letech řešil úlohu součtu čísel od 1 do 100 aritmetickou posloupností
- 50 dvojic $1 + 100 = 101 = 2 + 99 = 101 = 3 + 98 = 101$, tedy $50 * 101 = 5050$

Učení: Maximální věrohodnost

- Předpoklad **normální** hustoty $p(\vec{x}|s)$: jak lze využít v učení? Namísto odhadu neznámé hustoty rozdělení p odhadujeme parametry normálního rozdělení $p(\vec{x}|s, \vec{\mu}, \mathbf{\Sigma})$
- Tedy odhadujeme $\vec{\mu}_s$ a $\mathbf{\Sigma}_s$ pro každou třídu s.
- **Maximální věrohodnost**: ze zadaného vzorku $\vec{x}_1, \vec{x}_2, \ldots \vec{x}_m$ třídy s, najdi

$$(\hat{\vec{\mu}}_s, \hat{\mathbf{\Sigma}}_s) = \arg\max_{\vec{\mu},\mathbf{\Sigma}} p(\vec{x}_1, \vec{x}_2, \ldots |s, \vec{\mu}, \mathbf{\Sigma})$$

$$= \arg\max_{\vec{\mu},\mathbf{\Sigma}} \prod_{i=1}^{m} p(\vec{x}_i|s, \vec{\mu}, \mathbf{\Sigma}) = \arg\max_{\vec{\mu},\mathbf{\Sigma}} \sum_{i=1}^{m} \ln p(\vec{x}_i|s, \vec{\mu}, \mathbf{\Sigma})$$

  tj. maximalizuj věrohodnost toho, že vzorek byl generován ze třídy s s parametry $\vec{\mu}, \mathbf{\Sigma}$.
- Řešení je ve shodě s intuicí:

$$\hat{\vec{\mu}}_s = \frac{1}{m} \sum_{i=1}^{m} \vec{x}_i \qquad\qquad \hat{\mathbf{\Sigma}}_s = \frac{1}{m} \sum_{i=1}^{m} (\vec{x}_i - \hat{\vec{\mu}}_s)(\vec{x}_i - \hat{\vec{\mu}}_s)^t$$

- Tedy: vypočítej střední hodnotu vzorku a průměr m **matic** $(\vec{x}_i - \hat{\vec{\mu}}_s)(\vec{x}_i - \hat{\vec{\mu}}_s)^t$.
- Toto pro všechny třídy s.

Předpoklad 4: lineární forma klasifikátoru

- Předpokládejme binární klasifikační problém $S = \{s_1, s_2\}$,$\vec{x}$ je n-dimenzionální příznak
- Zde postačí jedna diskriminační funkce $g(\vec{x})$: klasifikuj $y = \begin{cases} s_1, & \text{pokud } g(\vec{x}) > 0; \\ s_2, & \text{jinak.} \end{cases}$
- Motivace: za předpokladu normálního rozdělení, pokud $\Sigma_{s_1} = \Sigma_{s_2}$, $g(\vec{x})$ je lineární, tj. $g(\vec{x}) = \vec{b}^t\vec{x} + c$.
- Namísto odhadu $\vec{\mu}, \mathbf{\Sigma}_s$ a následného výpočtu $\vec{b}$ and c můžeme odhadovat $\vec{b}, c$ přímo ze zadaného vzorku $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2) \ldots (\vec{x}_m, y_m)\}$.
- Požadujeme $\left(\vec{b}^t\vec{x}_i + c\right) > 0$ pokud $y_i = s_1$ a $\left(\vec{b}^t\vec{x}_i + c\right) < 0$ jinak.
- Totéž, jako požadovat $\left(\vec{b}^t\vec{z}_i + c\right) > 0$ pro všechna z_i, kde $z_i = x_i$ pokud $y_i = s_1$ a $z_i = -x_i$ jinak.
- Formálně, nechť $z_i^0 = 1\ \forall i$ a $\vec{w} = [\vec{b}, c]$ (přidejme c jako poslední složku $\vec{w}$).
- Nyní můžeme jednoduše psát $g(\vec{z}) = \vec{w}^t\vec{z}$ a požadovat $\vec{w}^t\vec{z}_i > 0$ pro všechna z_i .
- Nechť

$$J(\vec{w}) = \sum_{\vec{z}_i \in M} -\vec{w}^t\vec{z}_i$$

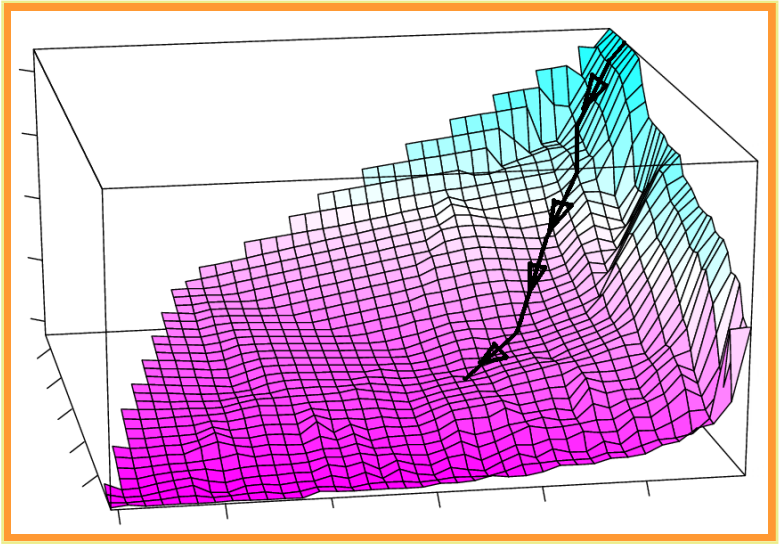
kde M je množina všech **nesprávně klasifikovaných** $\vec{z}_i$.

Perceptron

- $J(\vec{b}, c)$ je vždy nezáporná.
- Pokud $J(\vec{w}) = 0$, tak všechny příklady v D jsou správně klasifikovány a D je **lineárně separabilní**. Chceme najít minimum $J(\vec{w})$.
- $J(\vec{w})$ je po částech lineární. Pro hledání minima lze použít **gradientního algoritmu**.
- Gradientní algoritmus: přibližuj se k minimu pomocí malých diskrétních krorků v $\Re^{n+1}$ ve směru proti gradientu $J(\vec{w})$.

$$\nabla(J(\vec{w})) = \left(\frac{\partial J(\vec{w})}{\partial w_1}, \frac{\partial J(\vec{w})}{\partial w_2}, \dots \frac{\partial J(\vec{w})}{\partial w_{n+1}}\right) = \sum_{z_i \in M} -\vec{z}$$

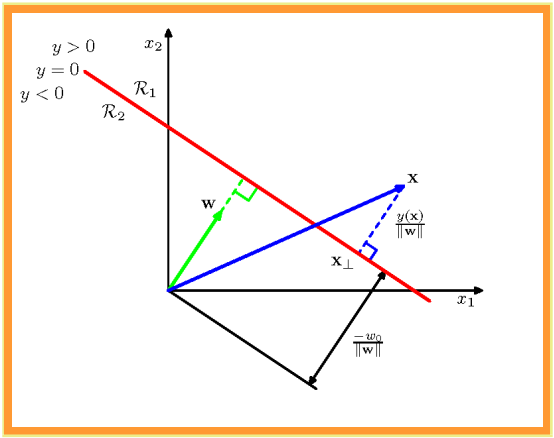
- **Perceptronový gradientní algoritmus**:
  1. $k = 0$. Zvol náhodně $\vec{w}$.
  2. $k \leftarrow k + 1$
  3. $\vec{w} \leftarrow \vec{w} + \eta(k) \sum_{z_i \in M_k} \vec{z}$
  4. pokud $|\sum_{z_i \in M_k} \vec{z}| > \theta$ jdi na 2
  5. vrať $\vec{w}$
- η - **rychlost učení**, θ - mez přijatelné chyby.



Příklad pro $\vec{w} \in \Re^2$, $J(\vec{w})$ na svislé ose.

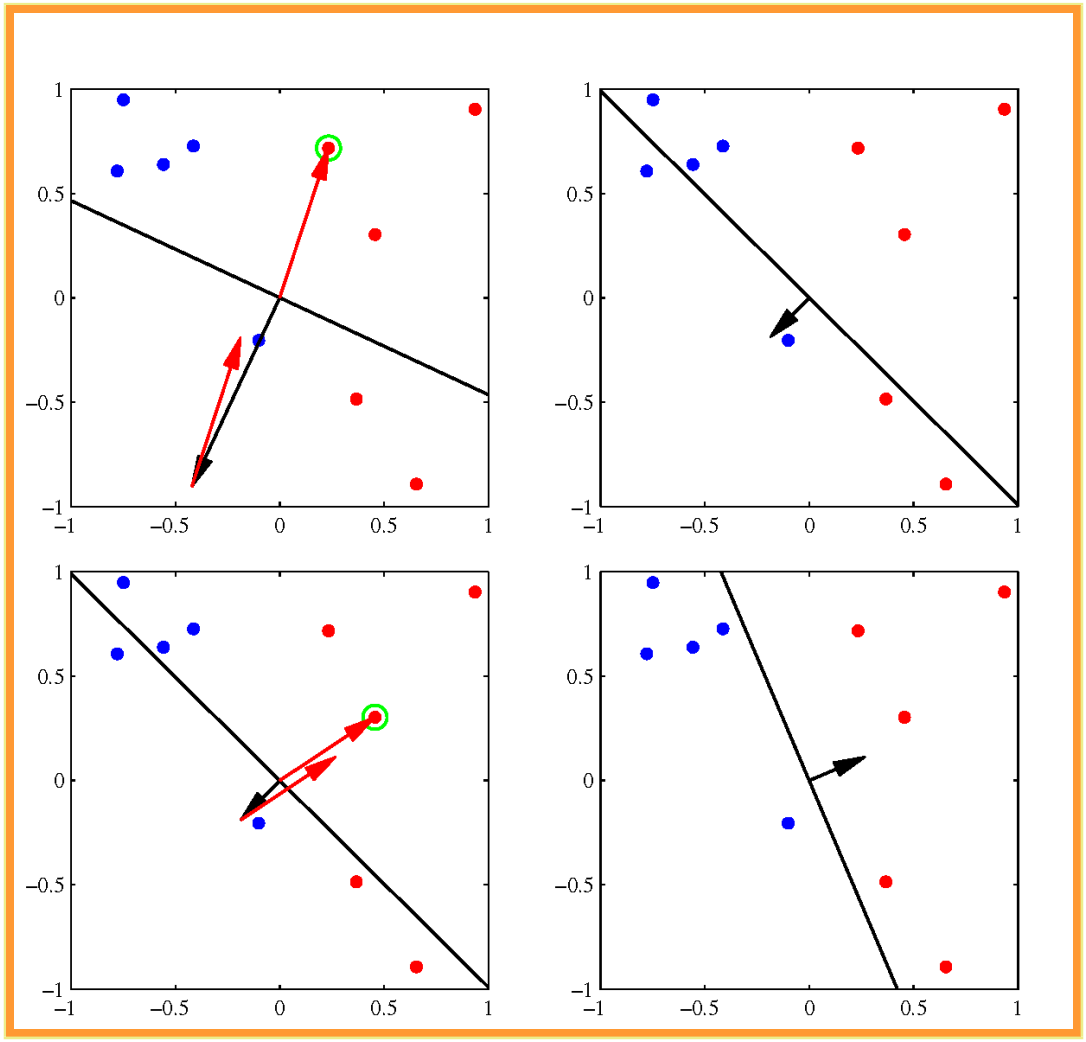
Percepton II

- Sjednocení terminologie se cvičením - rozšíření na K tříd
  - Nechť η =1, $g_s(x) = \vec{w_s}^t \vec{x} + b_s$, $s = 1 \ldots K$
  - $\vec{w_2} = \vec{w_2} + \vec{x}$, $b_2 = b_2 + 1$, pokud $\vec{x}$ má patřít do třídy $s = 2$, váhy pro správnou třídu posílím
  - $\vec{w_4} = \vec{w_4} - \vec{x}$, $b_4 = b_4 - 1$, $\vec{x}$ bylo vyhodnoceno jako třída $s = 4$, váhy pro špatnou třídu oslabím
- grafické znázornění rozhodovací hranice (změna symbolů)
  - $y(\vec{x}) = \vec{w}^t \vec{x} + w_0$, $y(\vec{x_a}) = y(\vec{x_b})$
  - $\vec{x_a}$ a $\vec{x_b}$ leží na rozhodovací hranici, tedy $\vec{w}^t(\vec{x_a} - \vec{x_b}) = 0)$
  - w je ortonormální na rozhodovací hranici, $\omega_0(b)$ je posunutí [Bishop]



Percepton - grafické znázornění učení

- grafické znázornění [Bishop]

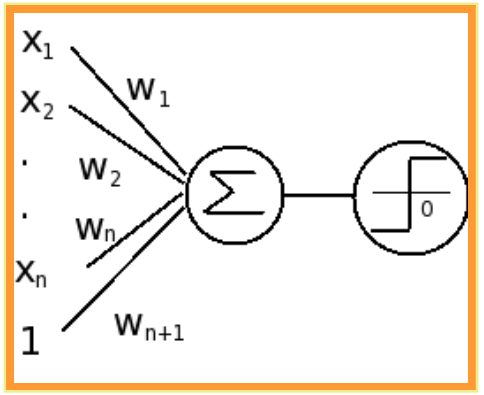


Perceptron: lineární diskriminace

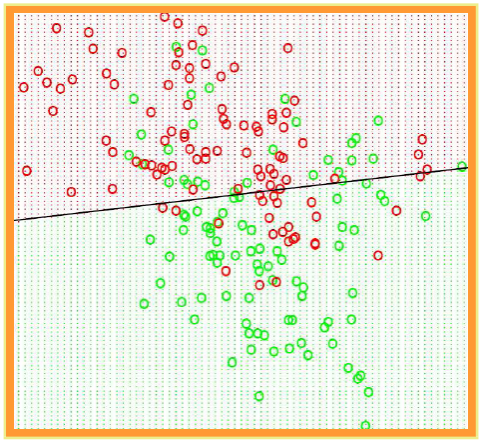
- Perceptron je využíván obecně pro úlohy lineární klasifikace, nejen za předpokladu normálně rozdělených příznaků.
- Je-li zadaný vzorek dvou tříd lineárně separabilní, perceptronový algoritmus skončí v konečném čase s nulovou chybou (pro $\theta = 0$ a dostatečně malé $\eta(k)$).
- Vzorek lineárně neseparabilní v $\Re^n$ může být separabilní po *transformaci* do $\Re^{n'}$ $n' > n$. Např. pro $\vec{x} \in \Re^n$, $T(\vec{x}) =$

$$[\underbrace{x(1), \ldots, x(n)}_{=\vec{x}}, x^2(1), x(1)x(2), x(1)x(3), \ldots x^2(n)]$$

- Tj. přidány kvadratické členy. Lineární diskriminace v transformovaném prostoru odpovídá nelineární (zde kvadratické) diskriminaci v původním prostoru $\Re^n$.
- Lineární diskriminační metody, jako je perceptron, mohou být tedy pomocí tranformace využity i pro hledání nelineární diskriminace.
- DEMO - iterace rozhodovací hranice ve 2D případu



Schema perceptronu


Lineárně neseparabilní problém

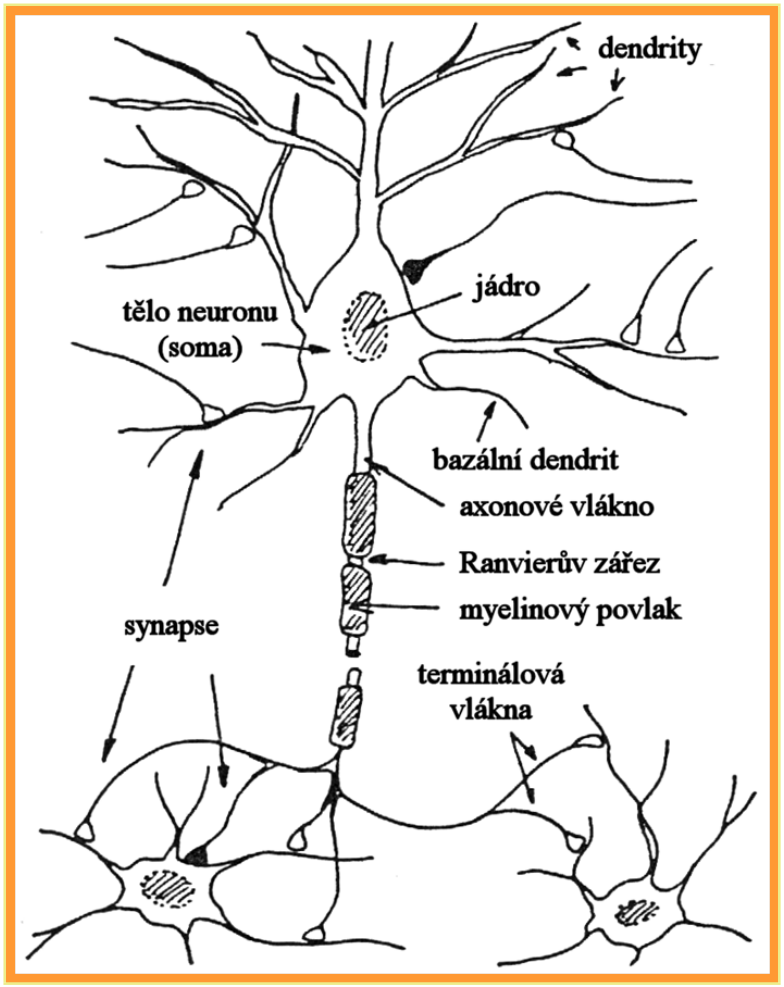
Percepton - historie

- Frank Rosenblatt - HW realizace perceptonu v roce 1958

- Učení jednoduchých písmen a tvarů - inspirace neuronovými sítěmi v mozku
- Světla osvětlují objekt, foceno na pole 20 x 20 cadmium-sulfádových buněk, dostáváme tedy obrázek o 400 pixelech
- Přepojovací deska - možnost různé konfigurace vstupů
- Adaptivní váhy - potenciometr ovládány elektrickým motorem - automatické nastavování vah učícím algoritmem
- Implementace na počítači Mark 1 (Hardvard-IBM spolupráce) - 765000 součastek, 16 m dlouhý 2.4 m vysoký, 2 m široký, 3 operace za sekundu, násobení trvalo 6 vteřin

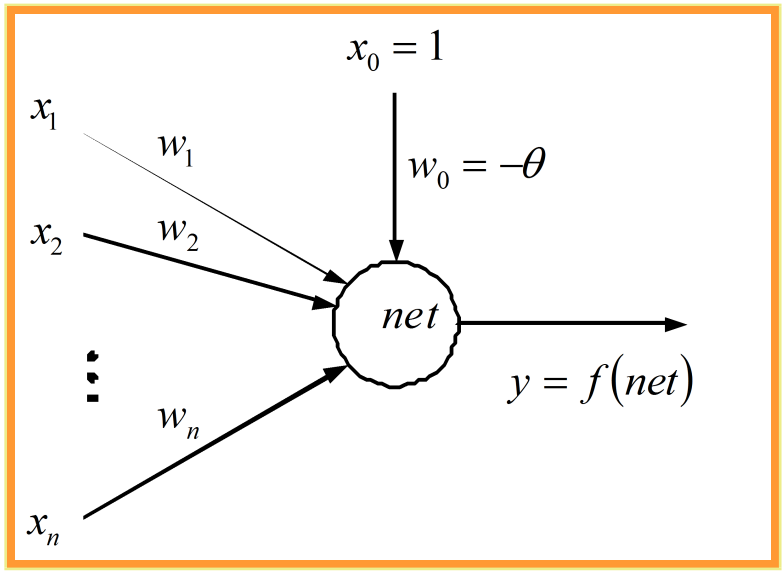
Neuronové sítě

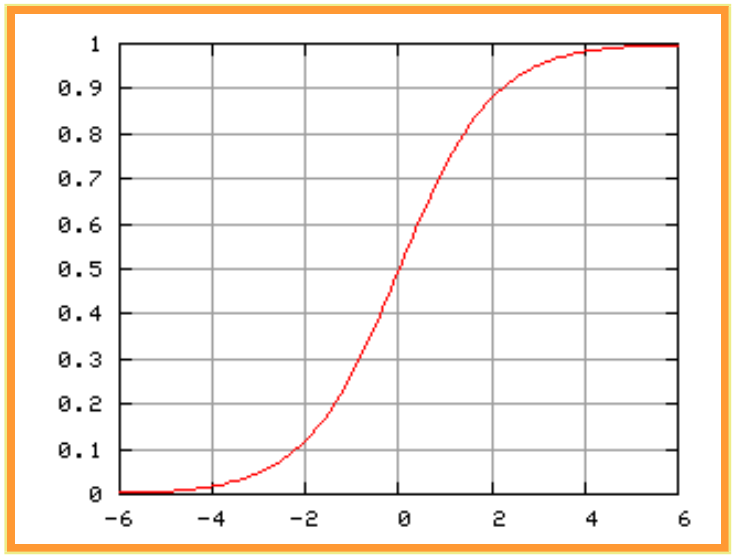
- Umělá neuronová sít je distribuovaný výpočetní systém sestávající z dílčích podsystémů (neuronů), který je inspirován neurofyziologickými poznatky o struktuře a činnosti neuronu a nervových systému živých organizmů, a který je ve vetší či menší míře realizuje.
- Neurony mezi sebou komunikují na základě přenášení elektrického potenciálu pomocí synapsí
- Schopnost extrahovat a reprezentovat závislosti v datech, které nejsou zřejmé
- Schopnost učit se, schopnost zevšeobecňovat
- Schopnost rešit silně nelineární úlohy



Definice neuronu

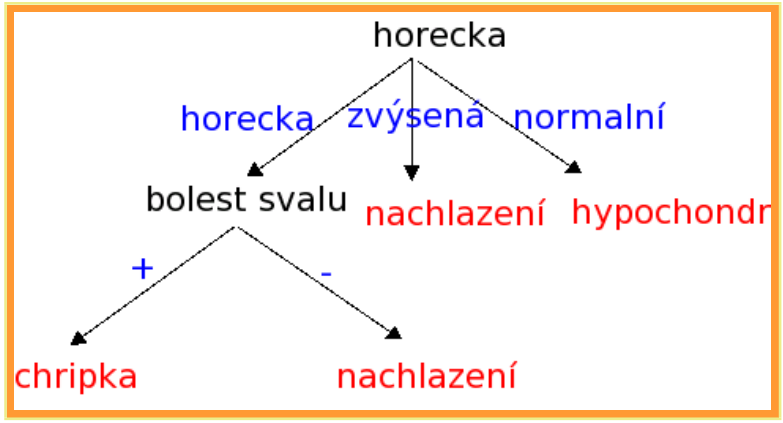
- Neuron je základní výpocetní jednotkou neuronových sítí
- Vstupy x_i jsou váhovány parametry ω_i
- $net = \sum_{i=1}^{n} x_i \omega_i + w_0 = \sum_{i=0}^{n}$
- Zavedení nelinearity do neuronové sítě $y = f(net)$, nejčastěji sigmoid třida, např, $\tanh$ či logistická funkce $y = f(net) = \frac{1}{1+\exp^{-\lambda * net}}$
- Přednášky pokračují v angličtině na základě kapitoly 6 knihy Duda, Pattern Classification





Rozhodovací strom

- V mnoha aplikacích strojového učení vyžadujeme klasifikační model přímo srozumitelný člověku.
- Příkladem srozumitelného modelu je **rozhodovací strom**.
- Souvislý graf bez smyček s ohodnocenými uzly a hranami.



- Označme i-tou složku příznaku příkladu jako $x(i)$. Uzly stromu (mimo listy) odpovídají nějaké složce příznaku $x(i)$.
  - Má-li $x(i)$ konečný obor hodnot: každá hrana vycházející z toho uzlu odpovídá jedné hodnotě z tohoto oboru hodnot. Uzel označený $x(i)$ a z něho vycházející hrana s označená hodnotou v definují *podmínku*

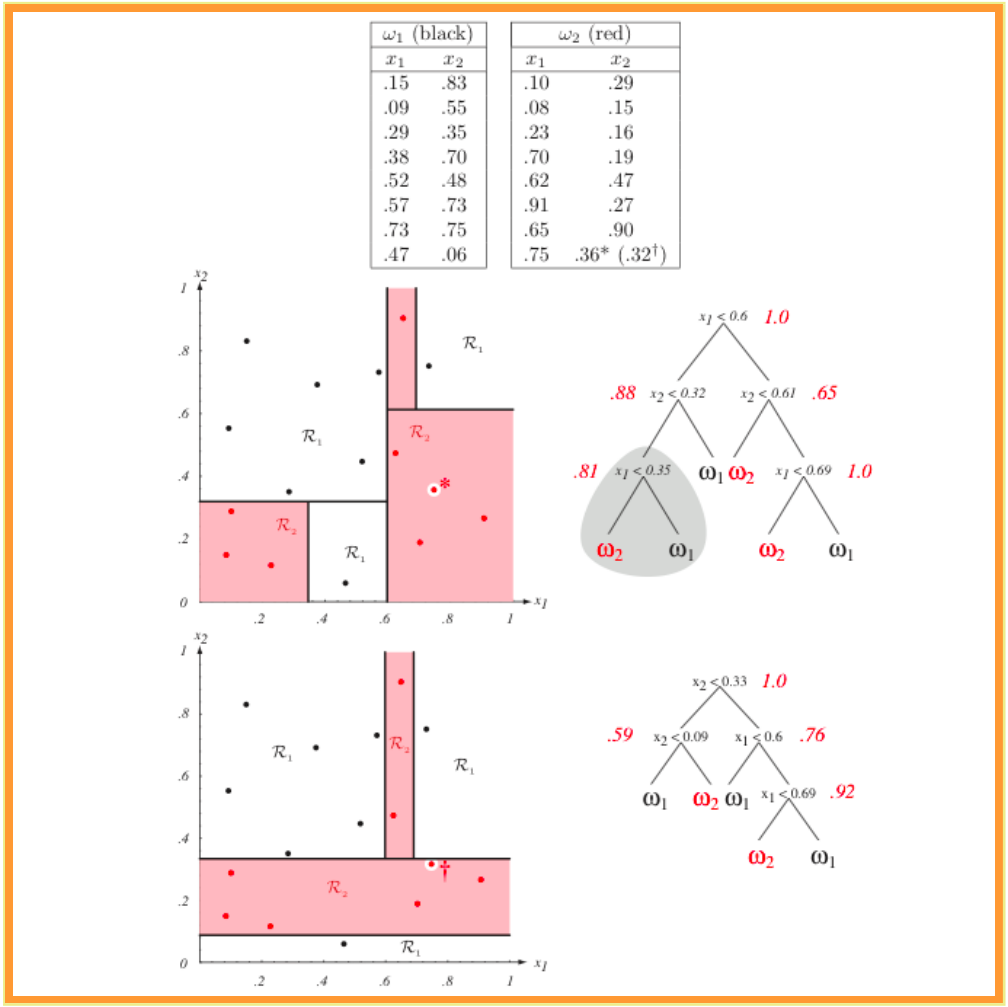
$$x(i) = v$$

  - Je-li $x(i)$ reálné: každá hrana vycházející z toho uzlu odpovídá nějakému reálnému intervalu. Např pro dvě hrany: $(-\infty, h)$ a (h, ∞), kde h je nějaká prahová hodnota. Uzel označený $x(i)$ a z něho vycházející hrana s označená intervalem I definují *podmínku*

$$x(i) \in I$$

Diskriminační hranice rozhodovacího stromu

- Diskriminační hranice rozhodovacího stromu pro reálné příznaky jsou nadroviny rovnoběžné s osami.
- Příklad pro binární klasifikaci v $\Re^2$:



ω_1 (black)		ω_2 (red)	
x_1	x_2	x_1	x_2
.15	.83	.10	.29
.09	.55	.08	.15
.29	.35	.23	.16
.38	.70	.70	.19
.52	.48	.62	.47
.57	.73	.91	.27
.73	.75	.65	.90
.47	.06	.75	.36* (.32†)

Priklad

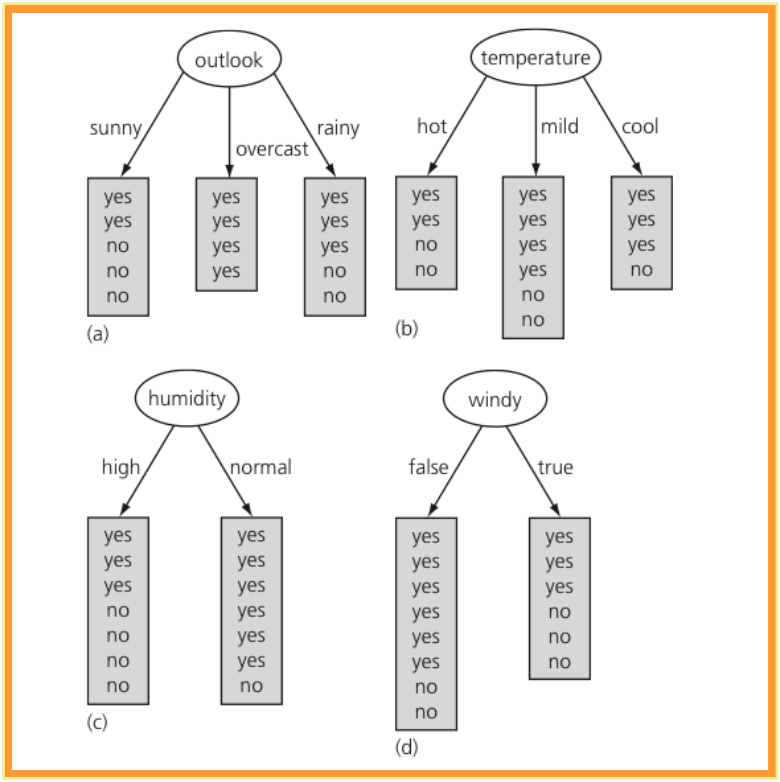
- Mam 9 pozitivnich a 5 negativnich prikladu

Table 1.2 The weather data.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

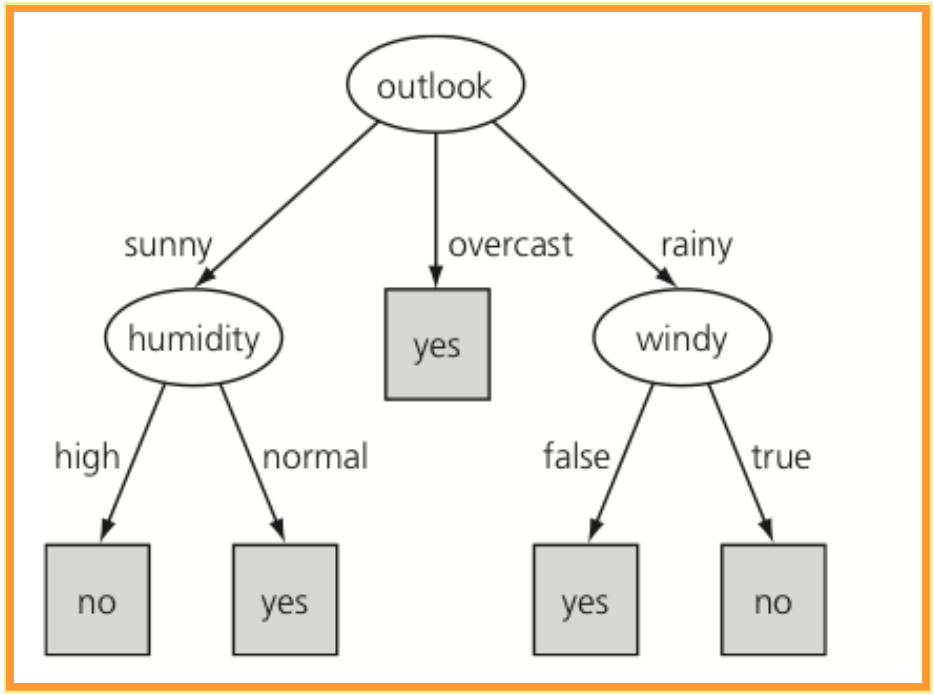
Konstrukce rozhodovacího stromu

- Pro tvorbu rozhodovacího stromu z trénovacích dat se užívá strategie ‘rozděl a panuj’.
- Zavedeme míru informace
- $info([2,3]) = 0.971$,$info([4,0]) = 0.0$, $info([3,2]) = 0.971$
- $info([2,3],[4,0],[3,2]) = (\frac{5}{14})0.971 + (\frac{4}{14})0 + (\frac{5}{14})0.971 = 0.693.$



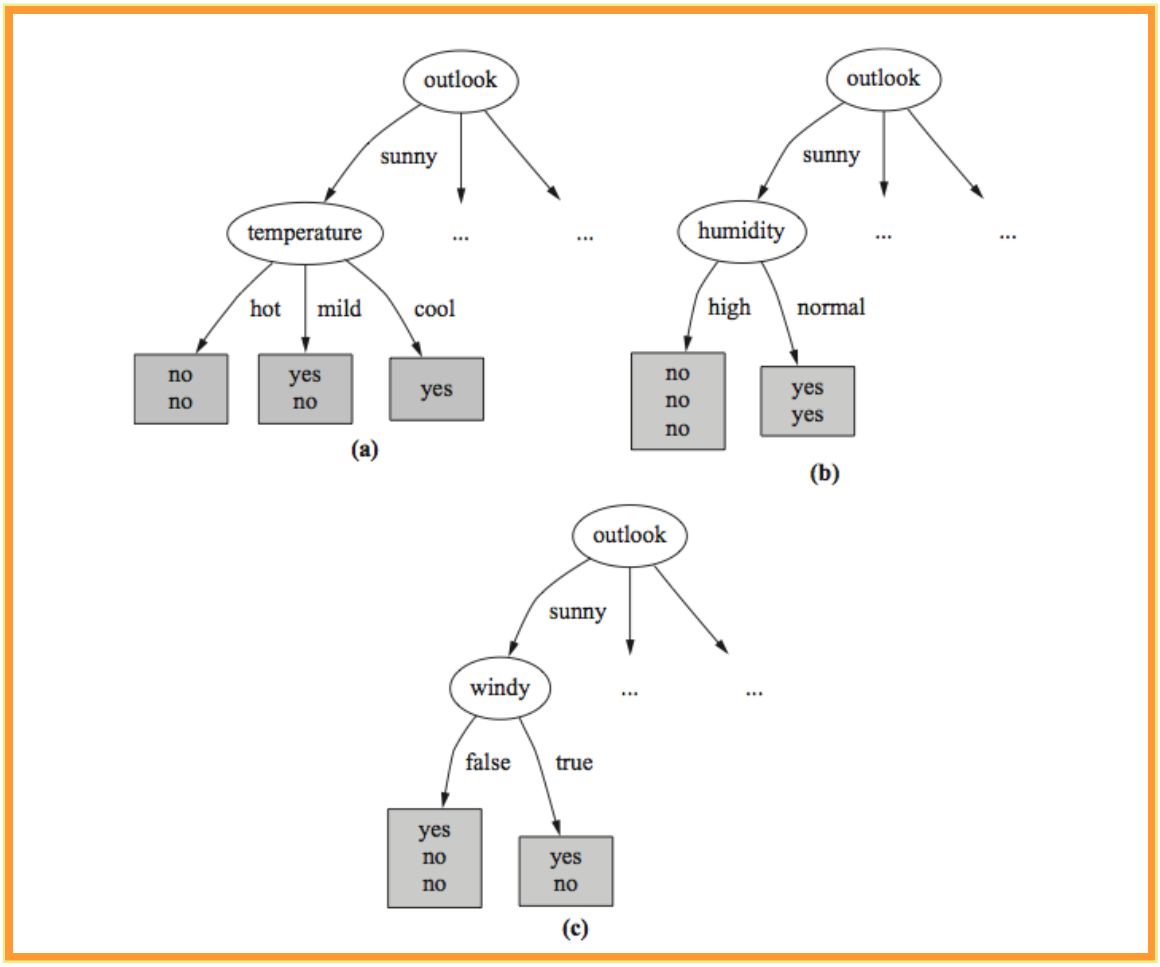
Konstrukce rozhodovacího stromu

- Zavedeme informacni zisk
- V korenu stromu: $info([9,5]) = 0.94$
- Informacni zisk $gain(outlook) = info([9,5]) - info([2,3],[4,0],[3,2]) = 0.940 - 0.693 = 0.247$ bits,
- $gain(temperature) = 0.029$ bits,$gain(humidity) = 0.152$ bits,$gain(windy) = 0.048$ bits
- Vybereme atribut outlook!



Konstrukce rozhodovacího stromu

- Pokracujeme v konstrukci
- $gain(temperature) = 0.571$ bits, $gain(humidity) = 0.971$ bits,$gain(windy) = 0.020$ bits
- vybereme $humidity$



Volba testu v uzlu rozhodovacího stromu

- Jak formalizovat pojem 'nejčistšího' rozdělení množiny příkladů?
- Lze využít pojmu entropie.
- Uvažujme vzorek D, γ tříd, a relativními četnostmi $p_1, p_2, \ldots p_\gamma$ příkladů v těchto třídách. Entropie vzorku je potom

$$H(D) = \sum_{i=1}^{\gamma} -p_i \log_2 p_i$$

- $H(D)$ je minimální ($H(D) = 0$), pokud jsou všechny příklady ve stejné třídě.
- Maximální $H(D) = \log_2 \gamma$, pokud je rozdělení rovnoměrné.
- Heuristika pro výběr $x(i)$ pro nový uzel: snížení entropie po přidání tohoto uzlu. Pro $x(i)$ s konečným oborem:

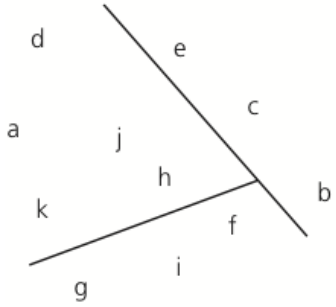
$$G(D, i) = H(D) - \sum_{v_j \in \mathit{Obor}(x(i))} \frac{|D_j|}{|D|} H(D_j)$$

D_j = množina příkladů z D, pro něž platí $x(i) = v_j$

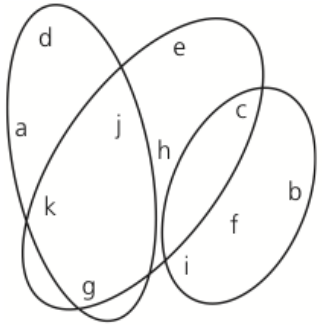
- Součet entropií v následných uzlech vážený relativní velikostí jim přiřazených množin.

Shlukovaní

- Nemam k dispozici trenovaci mnozinu
- Hledam v datech prirozene shluky
- (a) k-means, (b) fuzzy shlukovani, (c) pravdebodobnosti, napr. Gaussovska smes (EM algortimus), (d) hierarchicke shlukovani (dendogram)



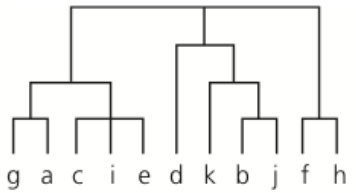
(a)



(b)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

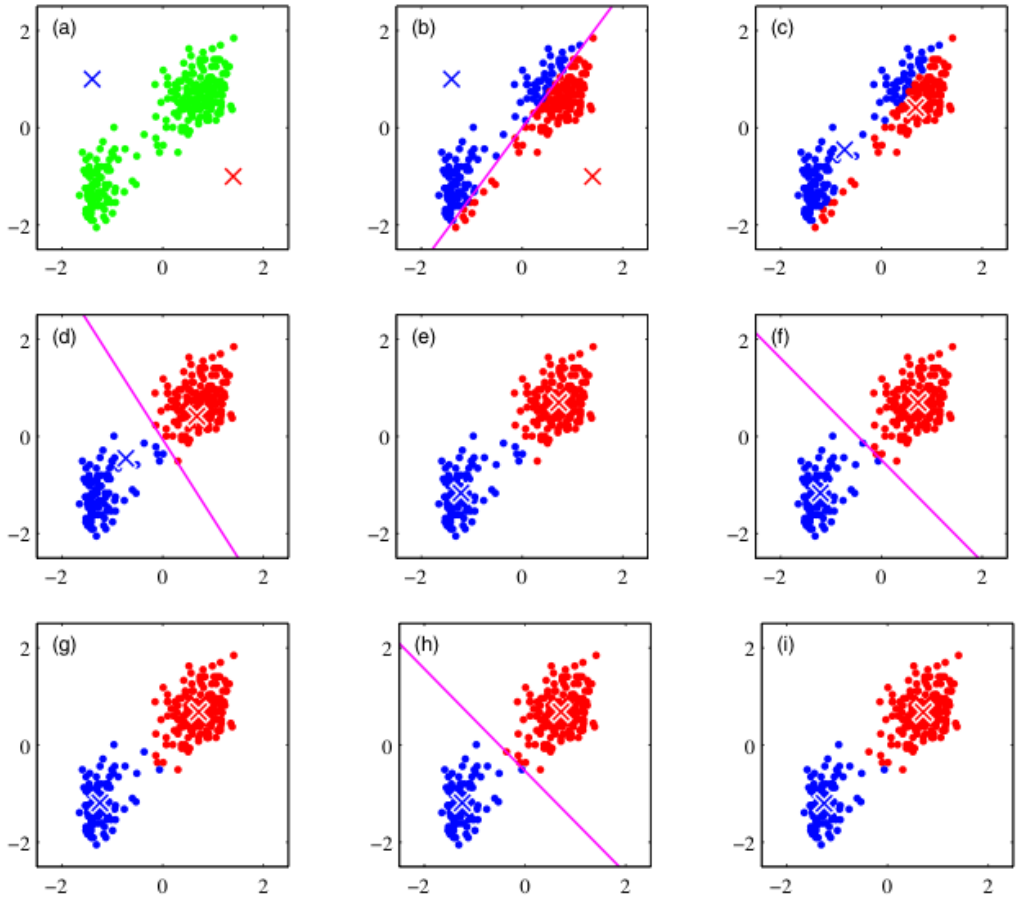
(c)



(d)

K-means

1. begin inicializuj $k, \mu_1, \mu_2, \ldots, \mu_k$
2. do klasifikuj vzorek podle nejblizsiho μ_i
3. prepocitej μ_i
4. until zadna zmena μ_i
5. return $\mu_1, \mu_2, \ldots, \mu_k$
6. end



Hierarchicke shlukovani

- agglomerative: bottom-up $\rightarrow$ merging
- divisive: top-down $\rightarrow$ splitting

1. begin inicializuj $k, \hat{k} \leftarrow n, \mathcal{D}_i \leftarrow \{X_i\}, i = 1, \ldots, n$
2. do $\hat{k} = \hat{k} - 1$
3. Najdi nejblizsi shluky, napr. $\mathcal{D}_i$ a $\mathcal{D}_j$
4. until $k = \hat{k}$
5. return k shluku
6. end

- $d_{min}(x, x^{'}) = \min \|x - x^{'}\|$, $x \in \mathcal{D}_i$, $x^{'} \in \mathcal{D}_i$

Hierarchicke shlukovani - priklad