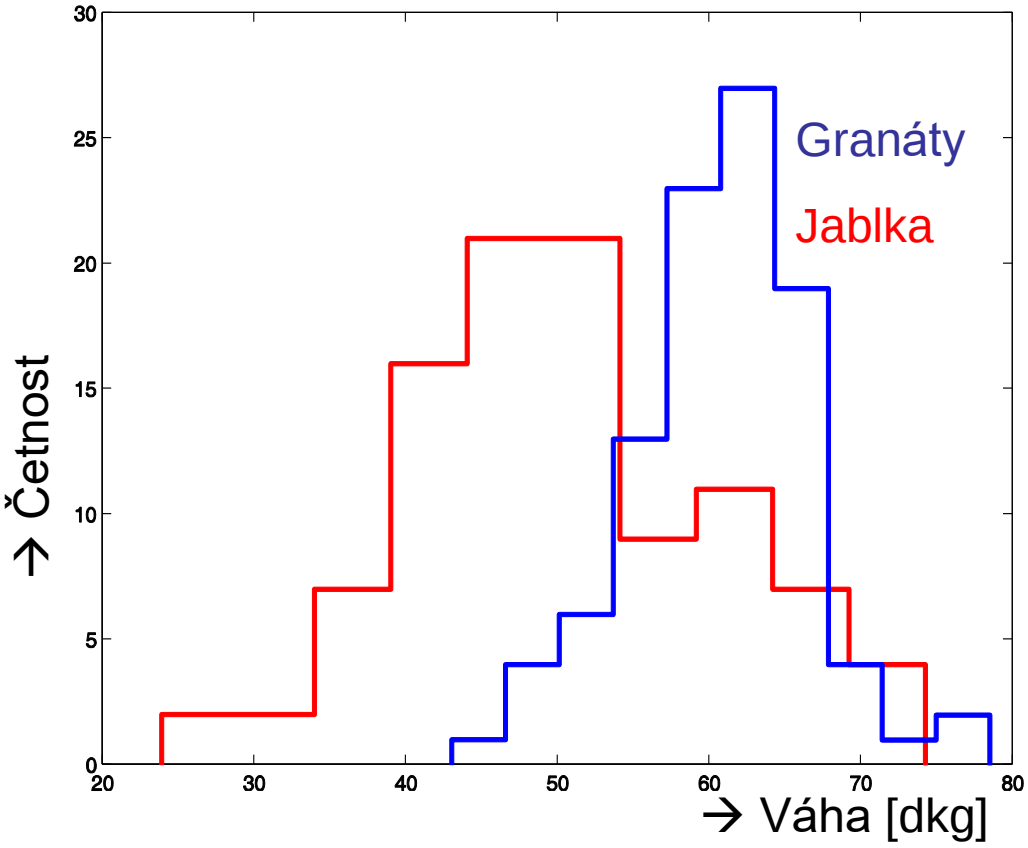
Strojové učení a rozpoznávání

Bayesovská rozhodovací teorie

Lukáš Burget

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Extrakce příznaků



Pravděpodobnosti - diskrétní příznaky

- Uvažujme diskrétní příznaky – „váhové kategorie“
- Nechť tabulka **reflektuje skutečné pravděpodobnosti** jednotlivých kategorií

1	6	12	15	12	2	2	50
4	23	50	14	6	3	1	100
nejlehčí	lehčí	lehký	střední	těžký	těžší	nejtěžší	
0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Apriorní pravděpodobnost – Stav věci

- Hádej co mám za zády, jablko nebo granát?
- Klasifikační pravidlo:
  - Vyber čeho je nejvíc
  - Třída s největší apriorní pravděpodobností (a-priori probability)

$$P(granát) = \frac{50}{150} \qquad P(jablko) = \frac{100}{150}$$

$$\sum_{\omega} P(\omega) = 1$$

1	6	12	15	12	2	2	50
4	23	50	14	6	3	1	100
nejlehčí	lehčí	lehký	střední	těžký	těžší	nejtěžší	
0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Společná pravděpodobnost

- Je to těžké. Hádej co to je?
- Klasifikační pravidlo:
  - Ve sloupci váhové kategorie vyber nejčastější třídu
  - Třída s největší společnou pravděpodobností (joint probability) – pravděpodobnost chlívečku.
  - … ale také největší podmíněnou pravděpodobností (viz další slajd)

$$P(granát, těžký) = \frac{12}{150} \qquad P(jablko, težký) = \frac{6}{150}$$

$$\sum_{\omega,x} P(\omega, x) = 1$$

1	6	12	15	12	2	2	50
4	23	50	14	6	3	1	100
nejlehčí	lehčí	lehký	střední	těžký	těžší	nejtěžší	
0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Podmíněná pravděpodobnost

- Je to těžké. S jakou pravděpodobností je to granát?
- Podmíněnou pravděpodobnost (conditional probability) - pravděpodobnost chlívečku dáno sloupec

$$P(granát|těžký) = \frac{12}{12+6}$$

1	6	12	15	12	2	2	50
4	23	50	14	6	3	1	100
nejlehčí	lehčí	lehký	střední	těžký	těžší	nejtěžší	
0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Ještě nějaké další pravděpodobnosti

$$P(granát) = \frac{50}{150} \qquad P(granát|těžký) = \frac{12}{12+6}$$

$$P(těžký) = \frac{12+6}{150} \qquad P(těžký|granát) = \frac{12}{50}$$

$$P(granát, těžký) = P(granát|těžký)P(těžký) = \frac{12}{150}$$

$$P(granát, těžký) = P(těžký|granát)P(granát) = \frac{12}{150}$$

1	6	12	15	12	2	2	50
4	23	50	14	6	3	1	100
nejlehčí	lehčí	lehký	střední	těžký	těžší	nejtěžší	
0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Bayesův teorém

Posteriorní pravděpodobnost (posterior probability)

Věrohodnost (likelihood)

Apriorní pravděpodobnost (prior probability)

$$P(\omega|x) = \frac{p(x|\omega)P(\omega)}{p(x)}$$

Evidence

- Věrohodnost nás zatím moc nezajímala, ale za chvíli to bude to hlavní co se budeme snažit odhadovat z trénovacích dat.
- Již dříve jsme viděli že (product rule):

$$P(\omega, x) = P(x|\omega)P(\omega)$$

- Pro evidenci platí (sum rule):

$$P(x) = \sum_{\omega} P(\omega, x)$$

např.: $P(\text{těžký}) = P(\text{granát}, \text{těžký}) + P(\text{jablko}, \text{těžký}) = \frac{12}{150} + \frac{6}{150}$

Maximum a-posteriori (MAP) klasifikátor

- Mějme 2 třídy ω_1 a ω_2
  - Pro daný příznak x vyber třídu ω s větší posteriorní pravděpodobností $P(\omega|x)$
  - Vyber ω_1 pouze pokud:

$$P(\omega_1|x) > P(\omega_{2,}|x)$$

$$\frac{P(x|\omega_1)P(\omega_1)}{\cancel{P(x)}} > \frac{P(x|\omega_2)P(\omega_2)}{\cancel{P(x)}}$$

$$P(\omega_1, x) > P(\omega_{2,}, x)$$

Maximum a-posteriori (MAP) klasifikátor

- Pro každé x minimalizuje pravděpodobnost chyby:
  P(chyby|x) = P(ω_1|x) pokud vybereme ω_2
  P(chyby|x) = P(ω_2|x) pokud vybereme ω_1
  Pro dané x vybíráme třídu ω s větším P(ω|x) ➔ minimalizace chyby

- Musíme ovšem znát
  - P(ω|x)
  - nebo P(x,ω)
  - nebo P(x|ω) a P(ω),

  které reflektují skutečná rozložení pro rozpoznávaná data

- Obecně pro N tříd
  - Vyber třídu s největší posteiorní pravděpodobností:

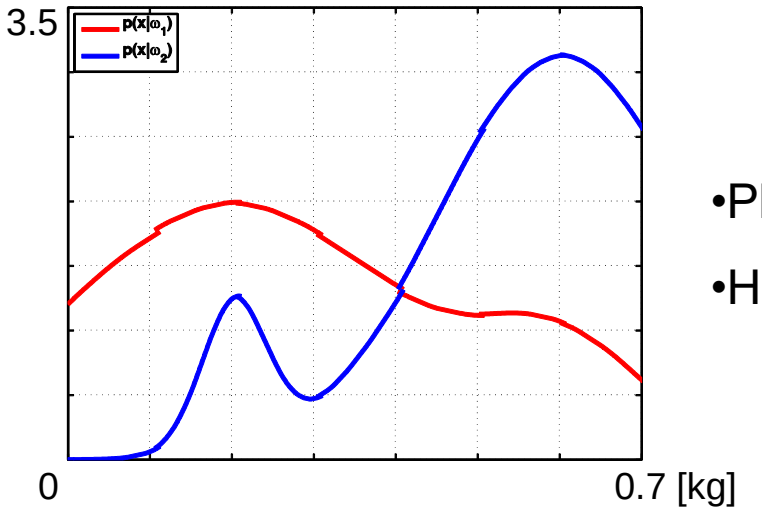
$$\arg\max_\omega P(\omega|x) = \arg\max_\omega P(x|\omega)P(\omega)$$

Spojité příznaky

- $P(.)$ – bude pravděpodobnost
- $p(.)$ – bude hodnota funkce rozložení pravděpodobnosti

$$P(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

- Bude nás zajímat funkce rozložení pravděpodobnosti příznaků podmíněné třídou $p(x|\omega)$

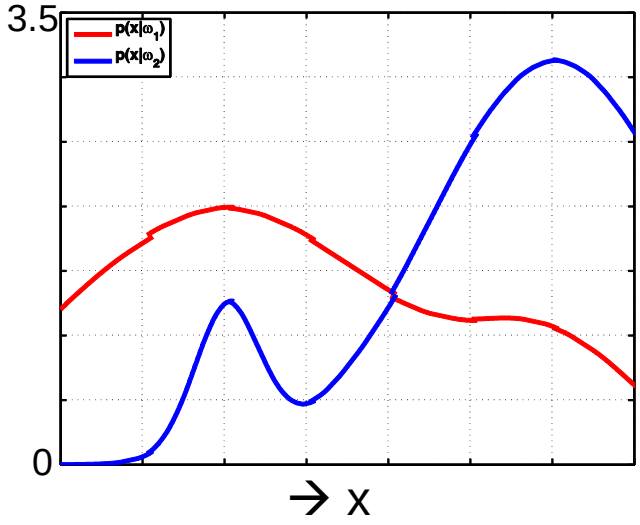


- Plocha pod funkci musí být 1
- Hodnoty mohou být ale libovolné kladné

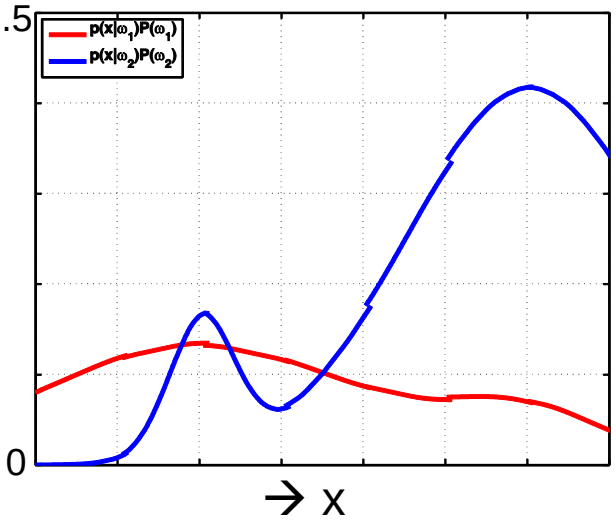
Bayesův teorém – spojité příznaky

$$P(\omega|x) = \frac{p(x|\omega)P(\omega)}{p(x)}$$

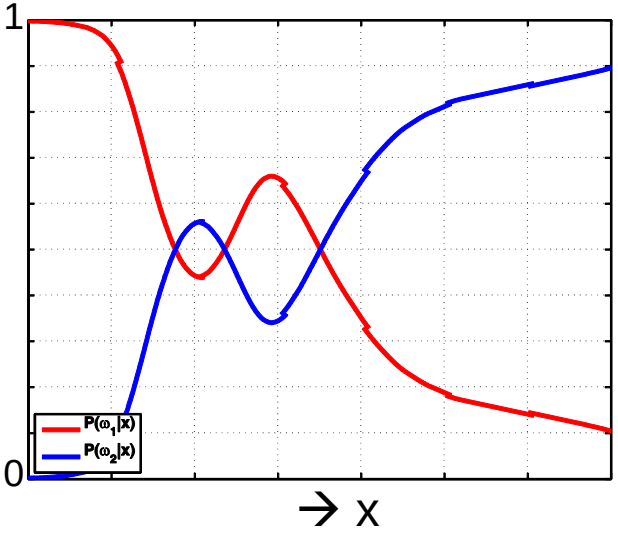
$p(x|\omega)$



$p(\omega, x) = p(x|\omega)P(\omega)$



$P(\omega|x)$



MAP klasifikátor – spojité příznaky

- Opět se budeme rozhodovat podle:

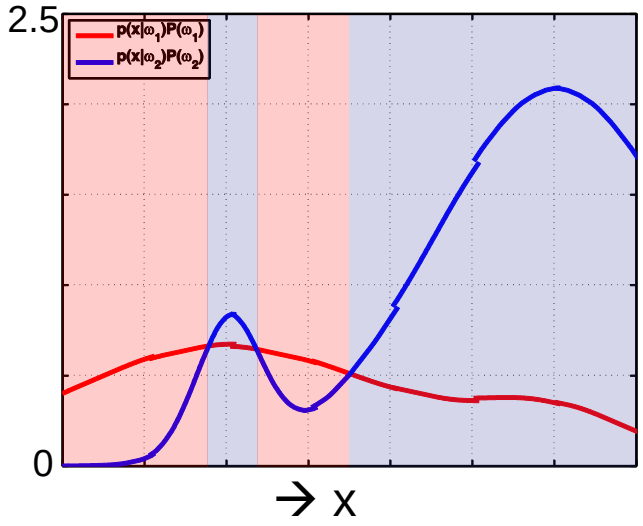
$$p(\omega_1, x) > p(\omega_{2,}, x)$$

nebo

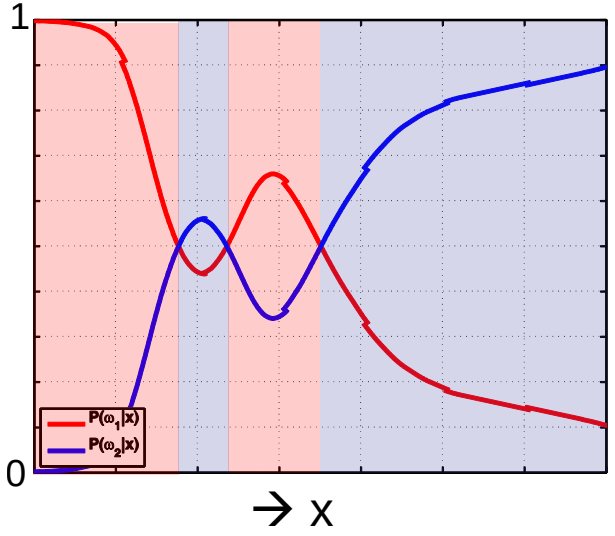
$$P(\omega_1 |x) > P(\omega_{2,} |x)$$

Na obrazcích vidíme, že obě pravidla vedou ke stejným rozhodnutím

$P(\omega, x)$



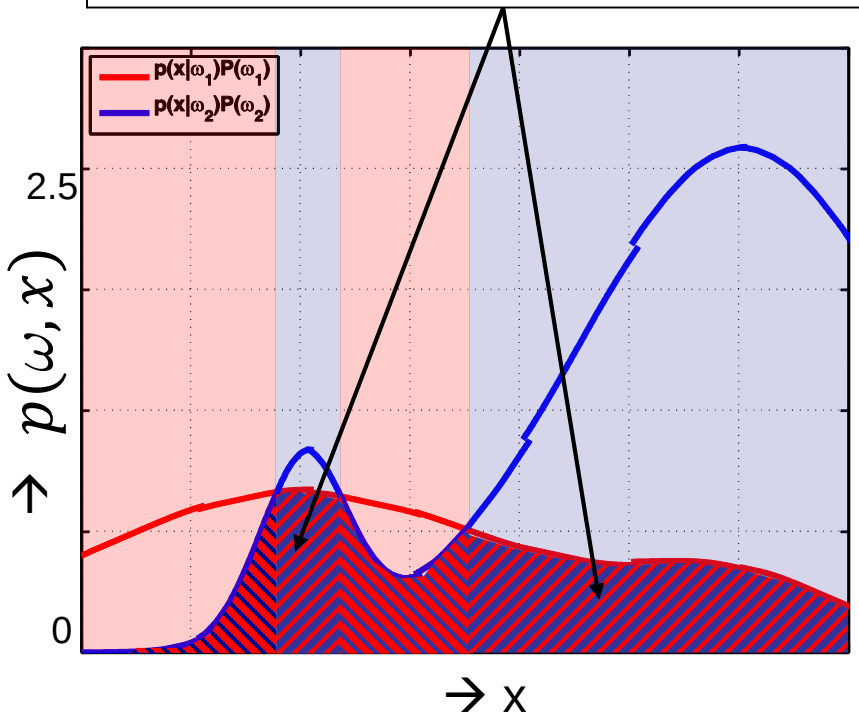
$P(\omega|x)$



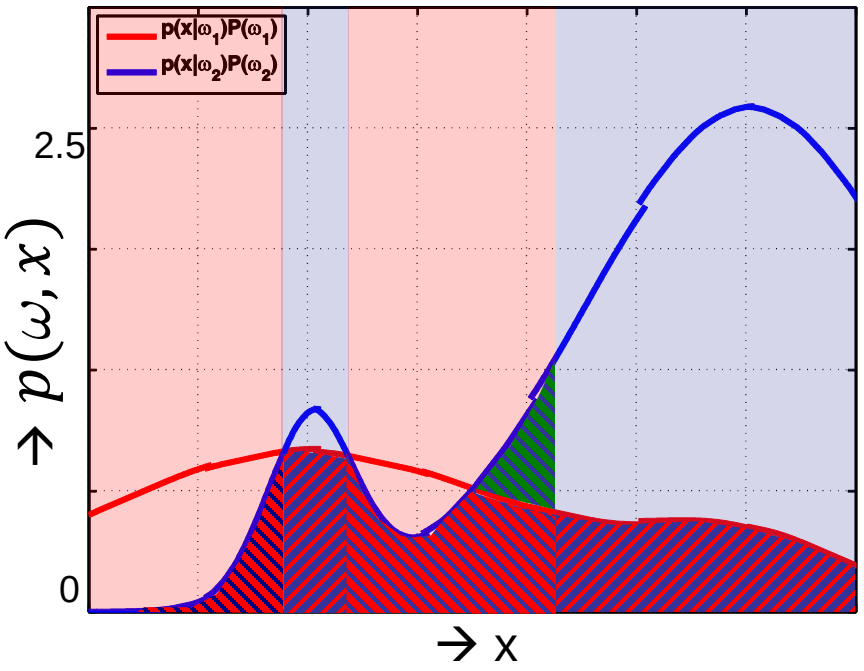
MAP klasifikátor – pravděpodobnost chyby

- Říkali jsme, že MAP klasifikátor minimalizuje pravděpodobnost chyby
- Plocha pod funkci společného rozložení pravděpodobnosti p(ω,x) v určitém intervalu x je pravděpodobnost výskytu vzoru třídy ω s příznakem v daném intervalu
- Jaká je tedy celková pravděpodobnost, že klasifikátor udělá chybu?

Pravděpodobnost, že červená třída je chybně klasifikována jako modrá



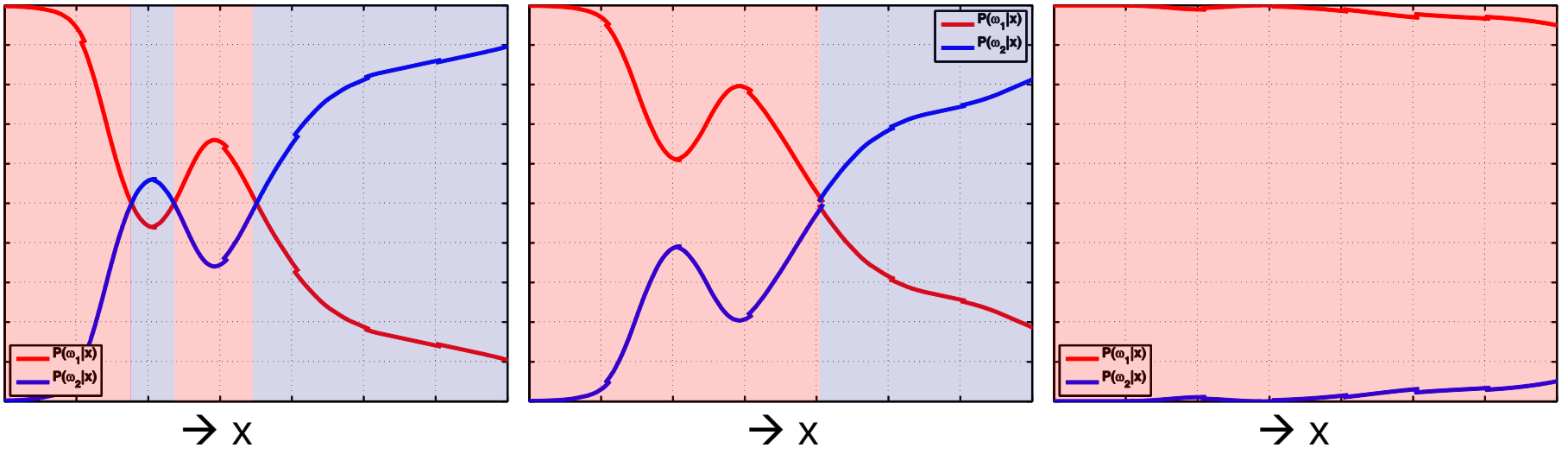
Jakákoli snaha posunout hranice povede jen k větší chybě



Posteriorní pravděpodobnosti pro různé apriorní pravděpodobnosti

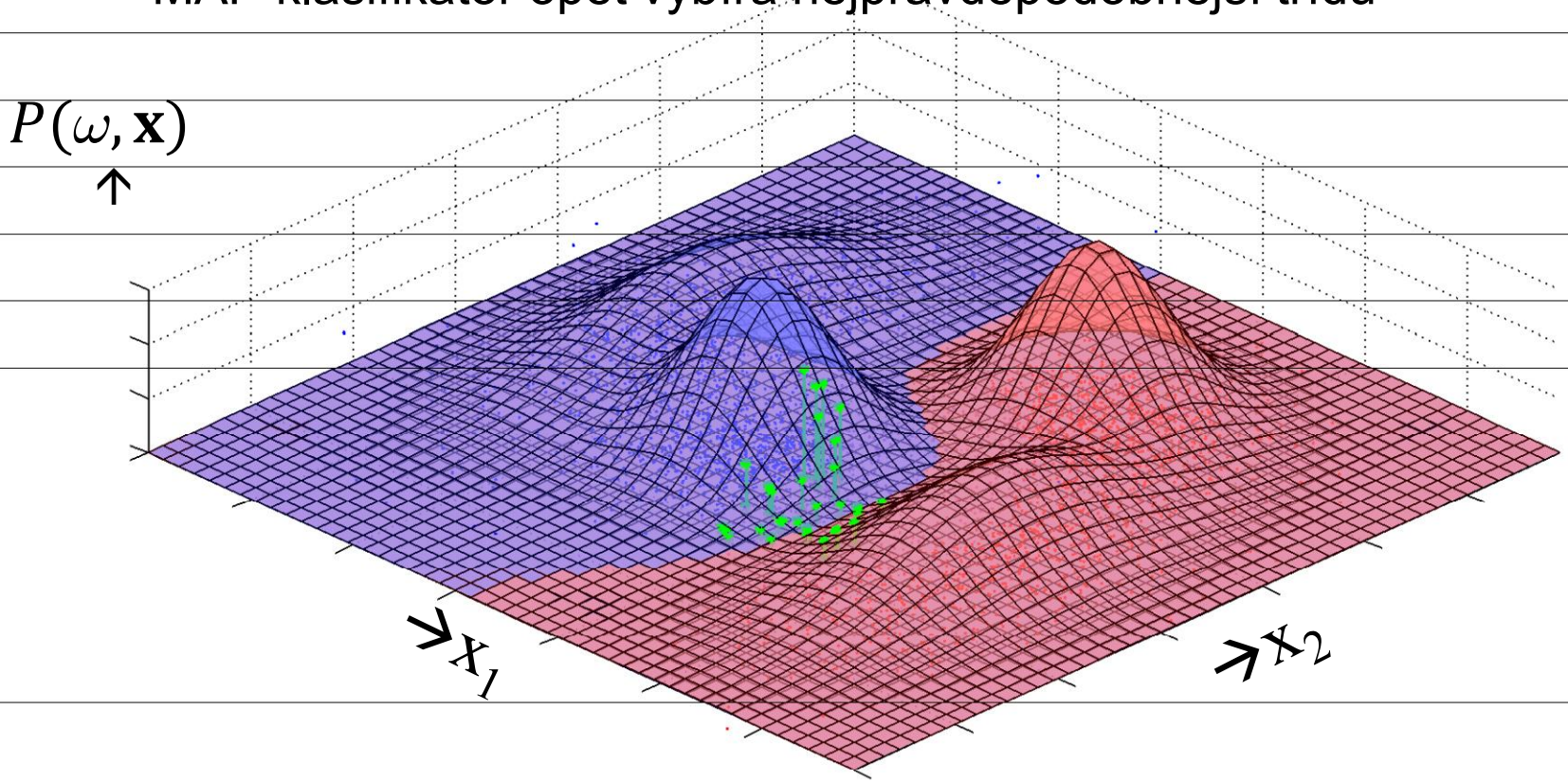
- Změna apriorních pravděpodobností tříd může vézt k různým rozhodnutím

$P(\omega_1) = \frac{1}{3}, P(\omega_2) = \frac{2}{3}$ $P(\omega_1) = \frac{1}{2}, P(\omega_2) = \frac{1}{2}$ $P(\omega_1) = \frac{99}{100}, P(\omega_2) = \frac{1}{100}$



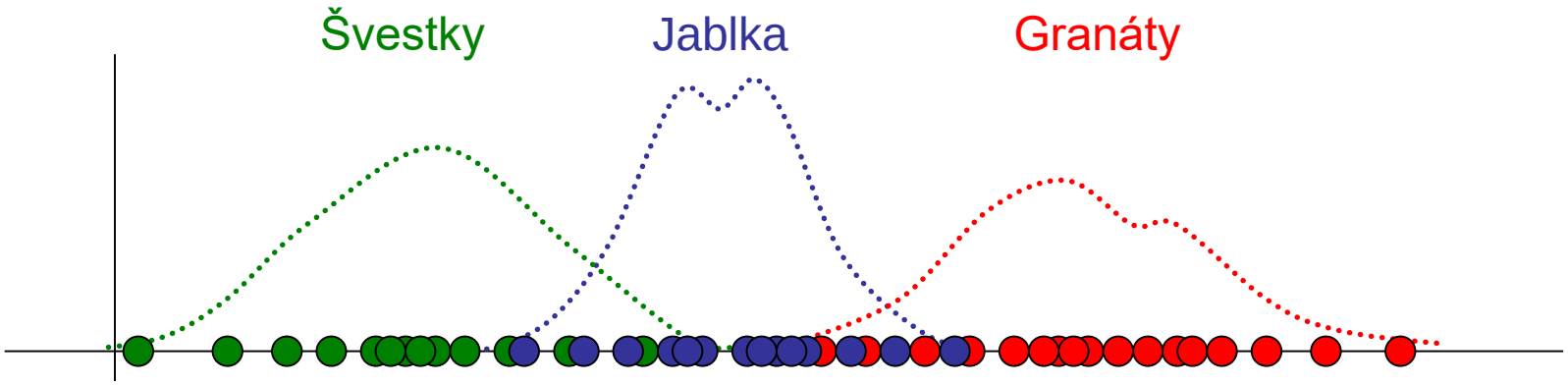
Vícerozměrné příznaky

- Místo jednorozměrného příznaku máme N rozměrný příznakový vektor
  - $\mathbf{x} = [x_1, x_2, \ldots, x_N]$
  - např. [váha, červenost]
  - MAP klasifikátor opět vybírá nejpravděpodobnější třídu



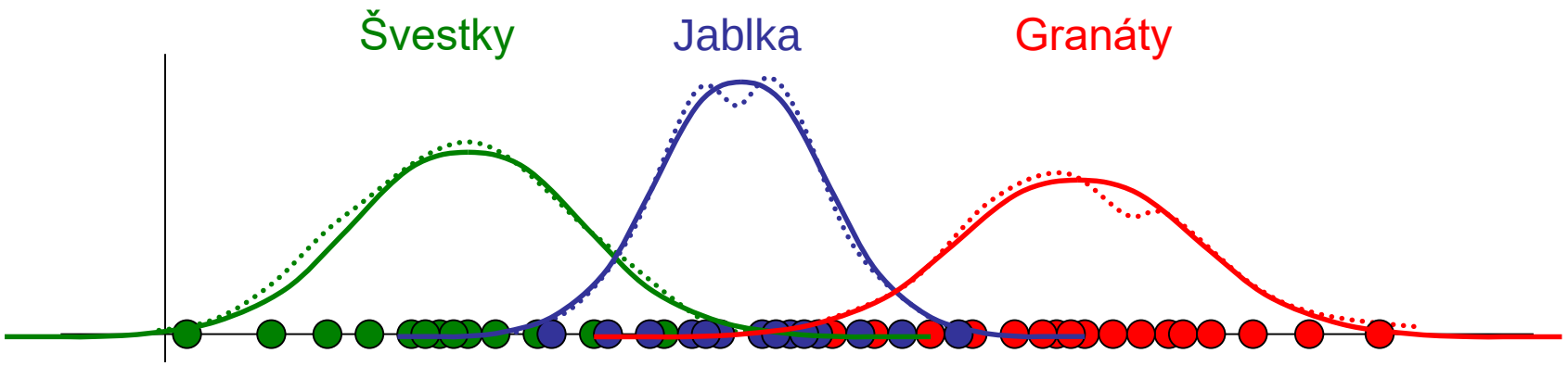
Parametrické modely

- Pro rozpoznávání s MAP klasifikátorem jsme doposud předpokládali, že známe skutečná rozloženi
  - $P(\omega|\mathbf{x})$
  - nebo $p(\mathbf{x},\omega)$
  - nebo $p(\mathbf{x}|\omega)$ a $P(\omega)$
- Ve skutečnosti ale většinou známe jen trénovací vzory
- Pokusíme se tato rozložení odhadnout z dat – budeme trénovat statistické modely



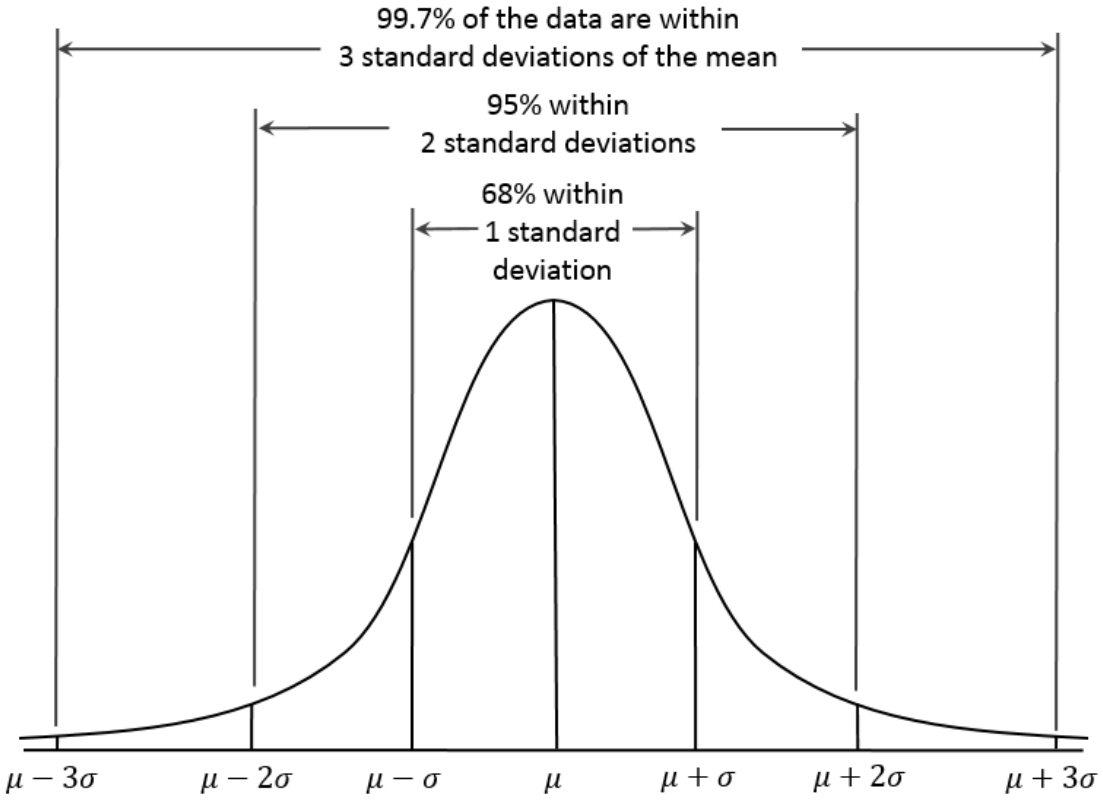
Parametrické modely

- Můžeme se pokusit modelovat přímo posteriorní pravděpodobnost, a tu použít přímo k rozpoznávání $P(\omega|\mathbf{x})$
  - tzv. **diskriminativní trénování**
  - Ale o tomto bude řeč až později
- Běžnější je odhadovat rozložení $p(\mathbf{x}|\omega)$ a $\mathrm{P}(\omega)$
- Tato rozložení popisují předpokládaný proces generování dat – **generativní modely**
- Nejprve se musíme rozhodnout pro formu modelu, který použijeme (např. gaussovské rozložení)



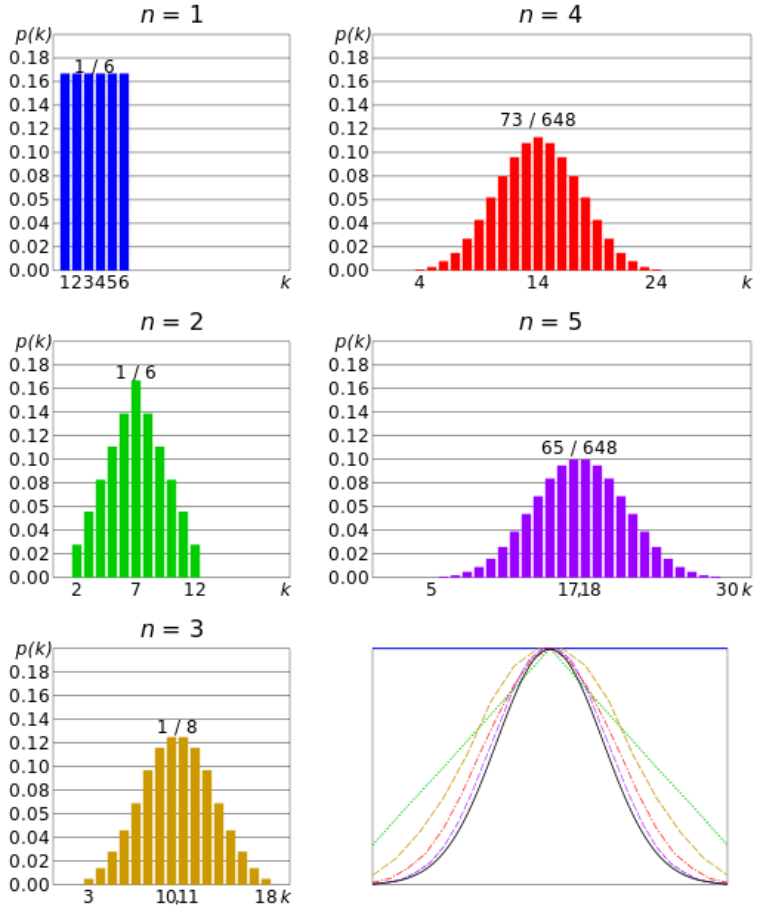
Gaussovské rozložení (jednorozměrné)

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Proč gaussovské rozložení?

- Přirozeně se vyskytuje
- Centralní limitní teorém: Sečtení hodnot mnoha bezávysle vygenerovaných nahodných čísel nám da vzorek z Gaussova rozložení
- Příklady:
  - Sečtení hodnot z N hracích kostek
  - Galton's board
    https://www.youtube.com/watch?v=03tx4v0i7MA



Proč gaussovské rozložení?

- Jednoduché a dobře se s ním pracuje
  - V logaritmické doméně je to jen kvadratická funkce

$$\log\mathcal{N}(x;\mu,\sigma^2) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}(x-\mu)^2 = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} + K$$

  - Věrohodnost množiny pozorování $\mathbf{x} = [x_1, x_2, x_3, \ldots x_N]$ je

$$p(\mathbf{x}|\mu,\sigma^2) = p(x_1, x_2, x_3, \ldots x_N|\mu,\sigma^2) = \prod_i \mathcal{N}(x_i;\mu,\sigma^2)$$

$$\log p(\mathbf{x}|\mu,\sigma^2) = \sum_i \log\mathcal{N}(x_i;\mu,\sigma^2) = \sum_i \log\mathcal{N}(x_i;\mu,\sigma^2)$$

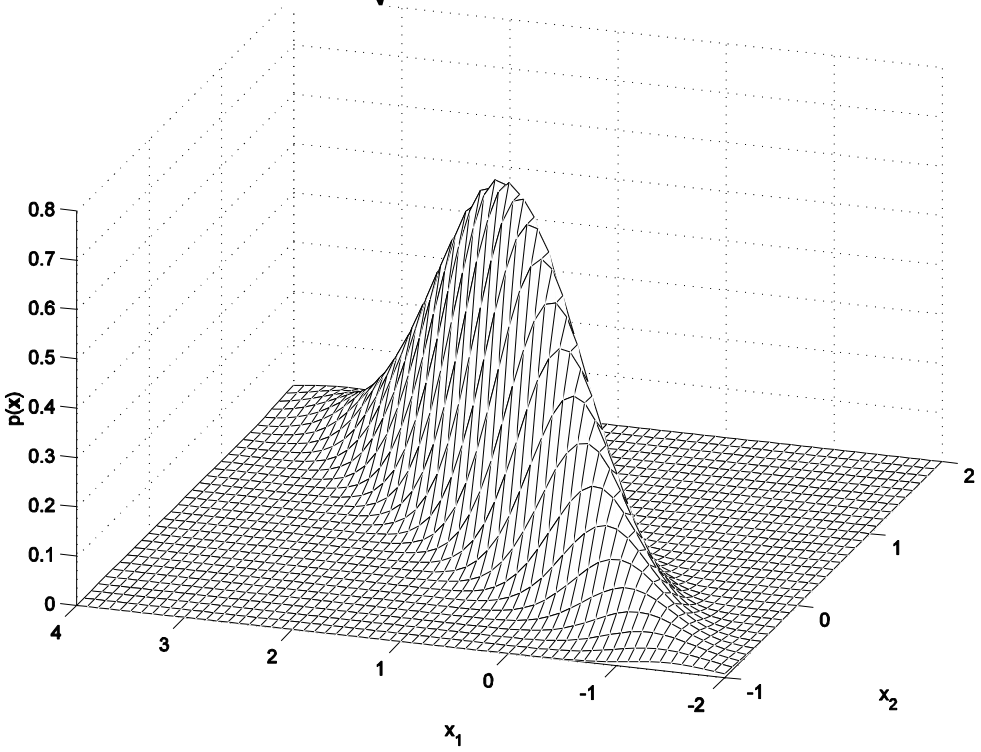
$$= -\frac{1}{2\sigma^2}\sum_i x_i^2 + \frac{\mu}{\sigma^2}\sum_i x_i - N\frac{\mu^2}{2\sigma^2} + NK$$

Postačující statistiky
(Sufficient statistics)

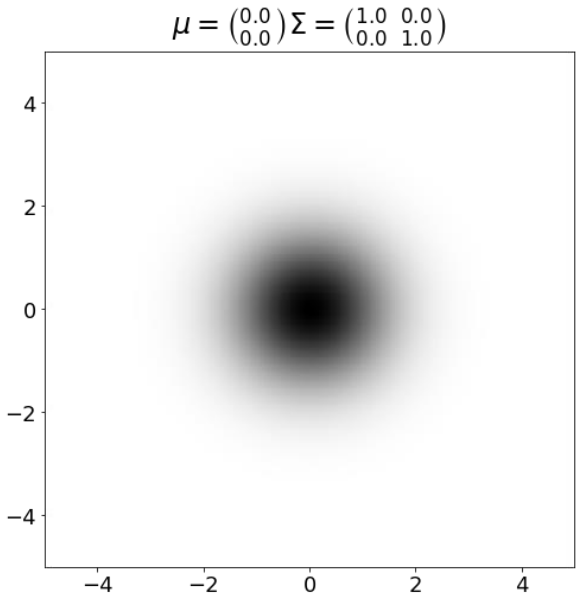
Gaussovské rozložení (vícerozměrné)

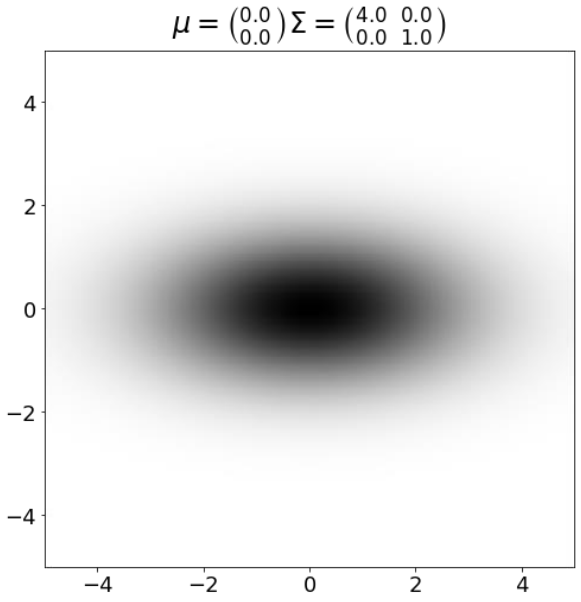
$$p(x_1, \ldots, x_D) =$$

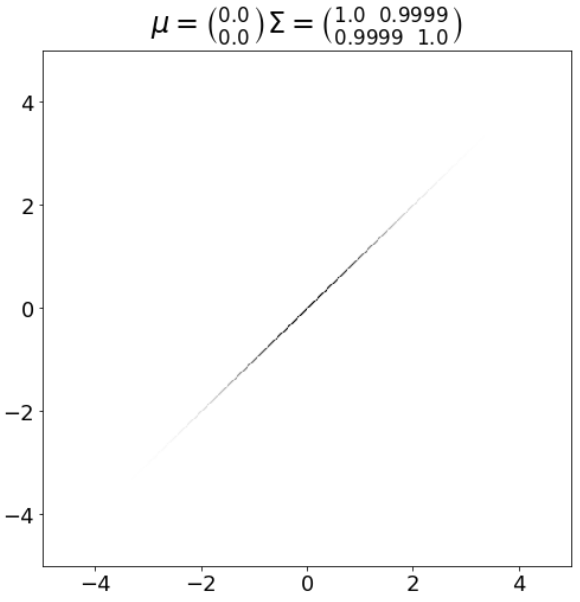
$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

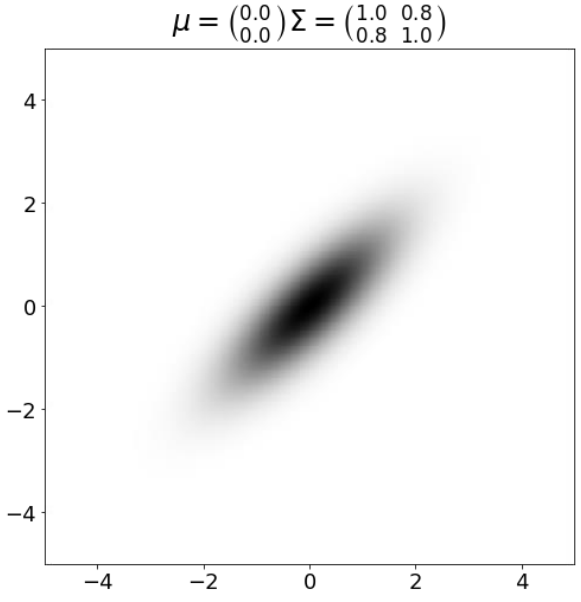


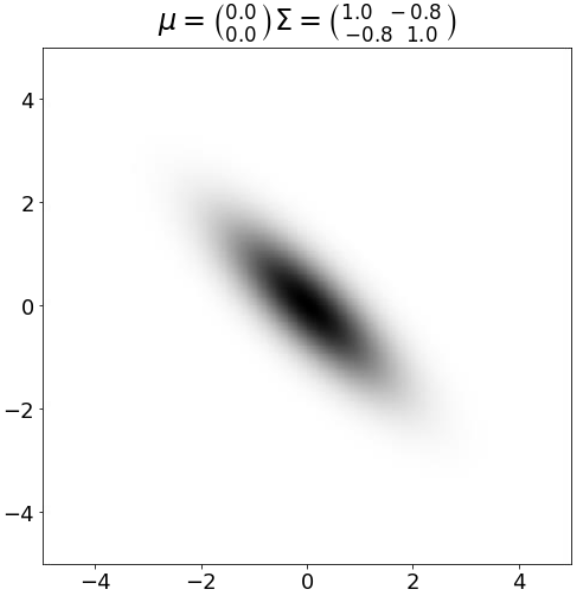
Příklady dvourozměrných gaussovek











Odhad parametrů rozložení s maximální věrohodností

- Mějme trénovací vzory (pozorování) $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, u kterých předpokládáme že jsou generovaná *nezávislé ze stejného rozložení (idependent, identically distributed - i.i.d.)* $p(\mathbf{x}|\boldsymbol{\eta})$ popsaného parametry $\boldsymbol{\eta}$
  - Např gaussovské rozložení s parametry $\boldsymbol{\eta} = \{\mu, \sigma^2\}$
- Maximálně věrohodným odhadem (*Maximum Likelihood Estimate*) jsou ty parametry $\widehat{\boldsymbol{\eta}}^{ML}$, které maximalizují funkci věrohodnosti:

$$\widehat{\boldsymbol{\eta}}^{ML} = \arg\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\eta}|\mathbf{X}) = \arg\max_{\boldsymbol{\eta}} p(\mathbf{X}|\boldsymbol{\eta}) = \arg\max_{\boldsymbol{\eta}} \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\eta})$$

- V následujících příkladech předpokládáme, že odhadujeme parametry nezávisle pro jednotlivé třídy. Pro zjednodušení notace tedy u rozložení neuvádíme závislost na třídě ω, pouze na jejích parametrech $\boldsymbol{\eta}$.
- Modely rozložení kterými se budeme zabývat jsou:
  - Gaussovské rozloženi
  - Směs gaussovských rozložení (Gaussian Mixture Model, GMM)
  - V následujících přednáškách přibudou další (např. HMM)

Proč odhad s maximální věrohodností

- Použijme opět Bayesův vzorec, ale tentokrát na vyjádření toho, jak jsou které parametry rozložení η pravděpodobné po tom co jsme vyděli trénovací data $\mathbf{X}$

$$p(\boldsymbol{\eta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\eta})}{p(\mathbf{X})}$$

- $p(\mathbf{X})$ nezavysí na $\boldsymbol{\eta}$ a je tedy pro dané $\mathbf{X}$ konstanta
- $p(\boldsymbol{\eta})$ je apriorní rozložení parametrů: naše prvotní představa o tom jak jsou které parametry pravděpodobné než vidíme trénovací data.
- Uvážíme-li *ignorantní prior* $p(\boldsymbol{\eta}) = konstanta$ (jakékoli parametry jsou pro nás stejně pravděpodobné) potom

$$\arg\max_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\mathbf{X}) = \arg\max_{\boldsymbol{\eta}} p(\mathbf{X}|\boldsymbol{\eta})$$

⇒ maximálně věrohodný odhad jsou ty nejpravděpodobnější parametry!

Ještě jednou: Proč gaussovské rozložení?

- Jednoduché a dobře se s ním pracuje
  - V logaritmické doméně je to jen kvadratická funkce

$$\log\mathcal{N}(x;\mu,\sigma^2) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}(x-\mu)^2 = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} + K$$

  - Věrohodnost množiny pozorování $\mathbf{x} = [x_1, x_2, x_3, \ldots x_N]$ je

$$p(\mathbf{x}|\mu,\sigma^2) = p(x_1, x_2, x_3, \ldots x_N|\mu,\sigma^2) = \prod_i \mathcal{N}(x_i;\mu,\sigma^2)$$

$$\log p(\mathbf{x}|\mu,\sigma^2) = \sum_i \log\mathcal{N}(x_i;\mu,\sigma^2) = \sum_i \log\mathcal{N}(x_i;\mu,\sigma^2)$$

$$= -\frac{1}{2\sigma^2}\sum_i x_i^2 + \frac{\mu}{\sigma^2}\sum_i x_i - N\frac{\mu^2}{2\sigma^2} + NK$$

Postačující statistiky
(Sufficient statistics)

ML odhad pro Gaussovku

$$\arg\max_{\mu,\sigma^2} p(\mathbf{x}|\mu,\sigma^2) = \arg\max_{\mu,\sigma^2} \log p(\mathbf{x}|\mu,\sigma^2) = \arg\max_{\mu,\sigma^2} \sum_n \log \mathcal{N}(x_n;\mu,\sigma^2)$$

$$= \arg\max_{\mu,\sigma^2} \left(-\frac{1}{2\sigma^2}\sum_n x_n^2 + \frac{\mu}{\sigma^2}\sum_n x_n - N\frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi)}{2} \right)$$

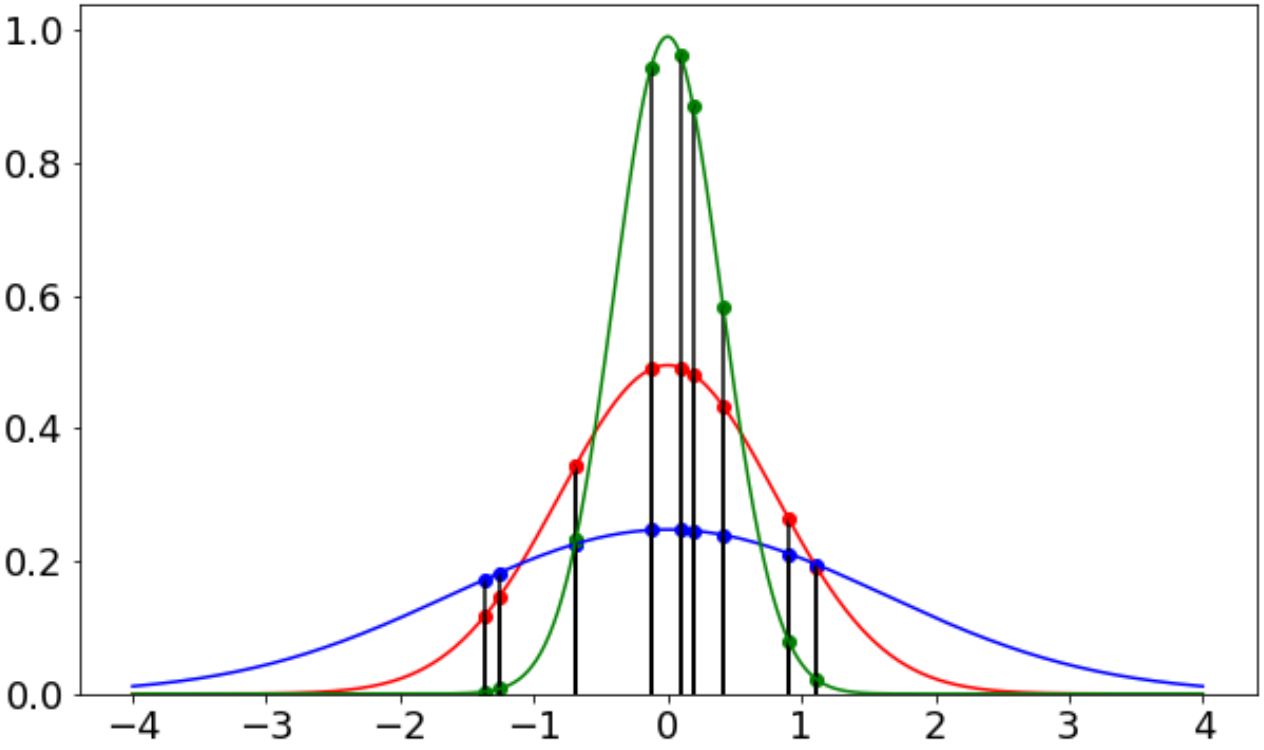
$$\frac{\partial}{\partial\mu}\log p(\mathbf{x}|\mu,\sigma^2) = \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_n x_n^2 + \frac{\mu}{\sigma^2}\sum_n x_n - N\frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi)}{2} \right)$$

$$= \frac{1}{\sigma^2}\left(\sum_n x_n - N\mu\right) = 0 \quad \Rightarrow \quad \hat{\mu}_{ML} = \frac{1}{N}\sum_n x_n$$

Postačující statistiky
(Sufficient statistics)

and similarly: $\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_n (x_n - \hat{\mu}_{ML})^2 = \frac{1}{N}\sum_n x_n^2 - \hat{\mu}^2_{ML}$

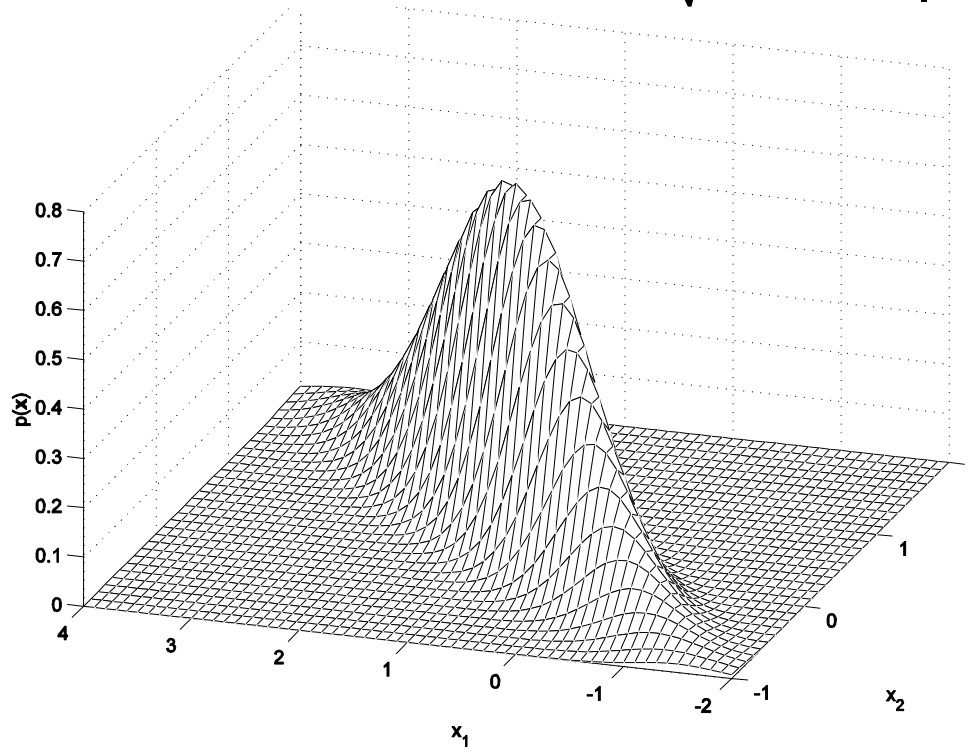
ML odhad pro Gaussovku



- Černé svislé čary představují trénovací vzory (pozorování)
- Červena gaussovka odpovídá maximálně věrohodnému odhadu maximum likelihood estimate
  - Násobek výšek červených teček bude největší možné číslo

Gaussovské rozložení
(vícerozměrné)

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



ML odhad of parametrů:

$$\boldsymbol{\mu} = \frac{1}{T}\sum_i \mathbf{x}_i$$

$$\boldsymbol{\Sigma} = \frac{1}{T}\sum_i (\mathbf{x}_\mathrm{i} - \boldsymbol{\mu})(\mathbf{x}_\mathrm{i} - \boldsymbol{\mu})^T$$

Diskrétní rozložení

$$p(\omega|\boldsymbol{\pi}) = \mathrm{Cat}(\omega|\boldsymbol{\pi}) = \pi_\omega$$

- Např. k popisu apriorní pravděpodobnosti tříd

$p(\omega\|\boldsymbol{\pi}) =$			
	0.5	0.3	0.2

- Speciální binární případ je Bernoulliho rozložení
- $\omega \in \{švestka, jablko, granát\}$
  nebo ω může být jednoduše index třídy $\omega \in \{1,2,\dots,C\}$
- $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_C]$ - parametry rozložení jsou pravděpodobnosti tříd
- Věrohodnost trénovacích vzorů $\boldsymbol{\omega} = [\omega_1, \omega_2, \dots, \omega_N]$

$$P(\boldsymbol{\omega}|\boldsymbol{\pi}) = \prod_n \mathrm{Cat}(\omega_n|\boldsymbol{\pi}) = \prod_n \pi_{\omega_n} = \prod_c \pi_c^{m_c}$$

kde m_c je počet pozorování třídy c (tedy čísla z naší tabulky)

ML odhad pro diskrétní rozložení

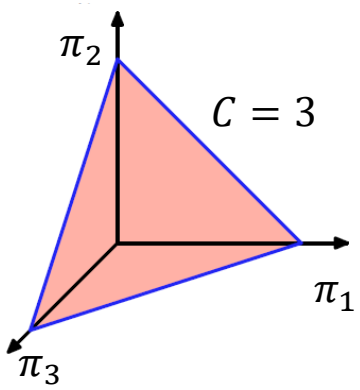
$$\arg\max_{\boldsymbol{\pi}} p(\mathbf{x}|\boldsymbol{\pi}) = \arg\max_{\boldsymbol{\pi}} \log p(\mathbf{x}|\boldsymbol{\pi}) = \arg\max_{\boldsymbol{\pi}} \log \prod_{n=1}^{N} \mathrm{Cat}(x_n|\boldsymbol{\pi})$$

$$= \arg\max_{\boldsymbol{\pi}} \log \prod_{c} \pi_c^{m_c} = \arg\max_{\boldsymbol{\pi}} \sum_{c} m_c \log \pi_c$$

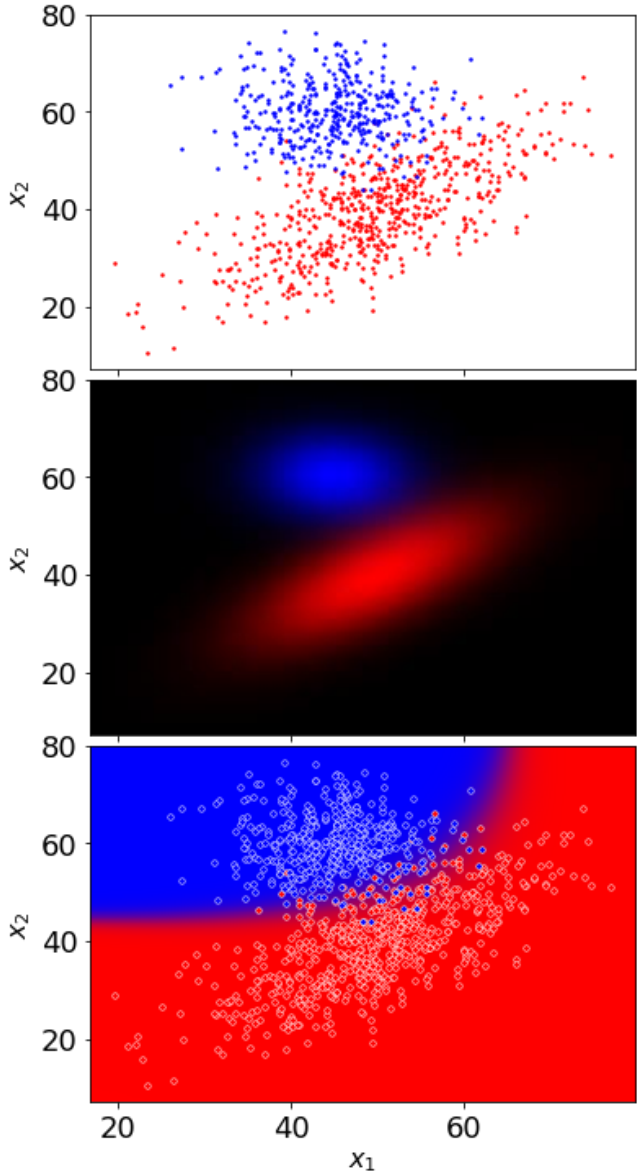
Potřebujeme lagrangeovy multiplikátory λ pro zajičtění podminký $\sum_k \pi_k = 1$

$$\frac{\partial}{\partial \pi_c} \log p(\mathbf{x}|\boldsymbol{\pi}) = \frac{\partial}{\partial \pi_c} \left(\sum_k m_k \log \pi_k - \lambda \left(\sum_k \pi_k - 1 \right) \right) = \frac{m_c}{\pi_c} - \lambda = 0$$

$$\Rightarrow \pi_c = \frac{m_c}{\lambda} = \frac{m_c}{N}$$



Gaussovský klasifikátor – 2D data



- Apriorní pravděpodobnosti tříd mohou byt ML odhadnuty jako poměry počtu příkladů

$$P(c) = \frac{N_c}{\sum_k N_k} \quad P(blue) = \frac{400}{400 + 600}$$

- Probability density function for each class is assumed to be 2D Gaussian

$$p(\mathbf{x}|c) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

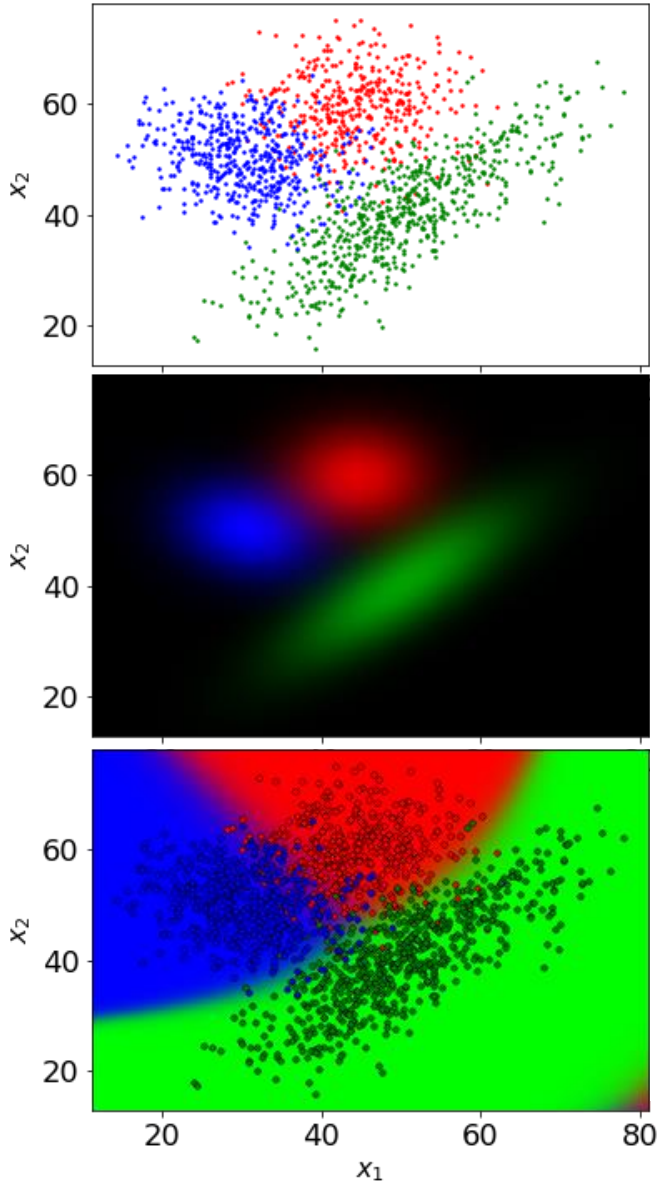
and its parameters ML estimated as

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{n=1}^{N_c} \mathbf{x}_{cn} \quad \boldsymbol{\Sigma}_c = \frac{1}{N_c}\sum_{n=1}^{N_c} (\mathbf{x}_{cn} - \boldsymbol{\mu}_c)(\mathbf{x}_{cn} - \boldsymbol{\mu}_c)^T$$

- Class posterior probability for new observations is obtained from the prior and class pdf-s using Bayes rule:

$$P(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)P(c)}{\sum_k p(\mathbf{x}|k)P(k)}$$

Gaussian classifier – more classes



- Class priors can be ML estimated as the proportions of the example counts

$$P(c) = \frac{N_c}{\sum_k N_k} \quad P(blue)\frac{400}{400+600}$$

- Probability density function for each class is assumed to be 2D Gaussian

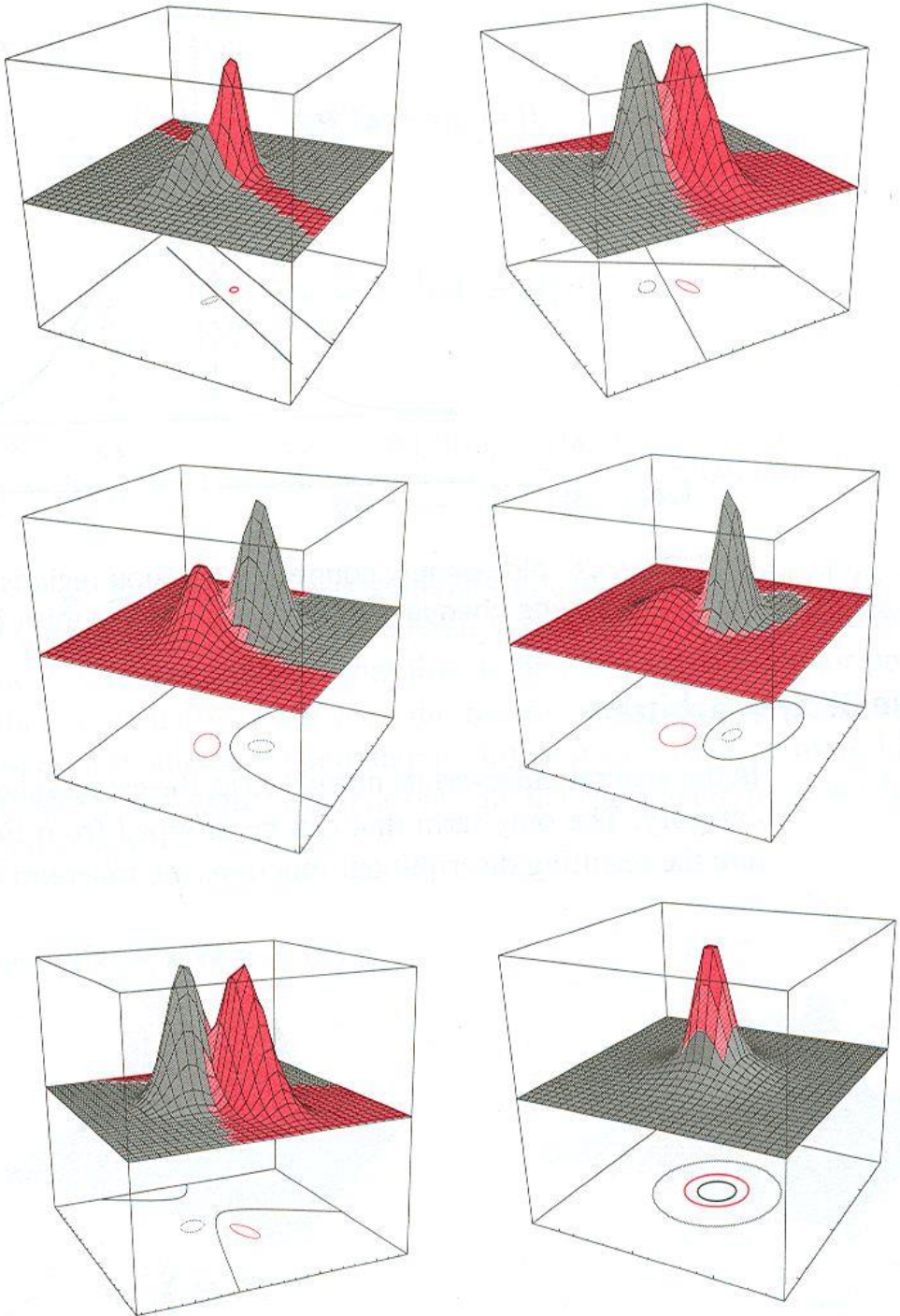
$$p(\mathbf{x}|c) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

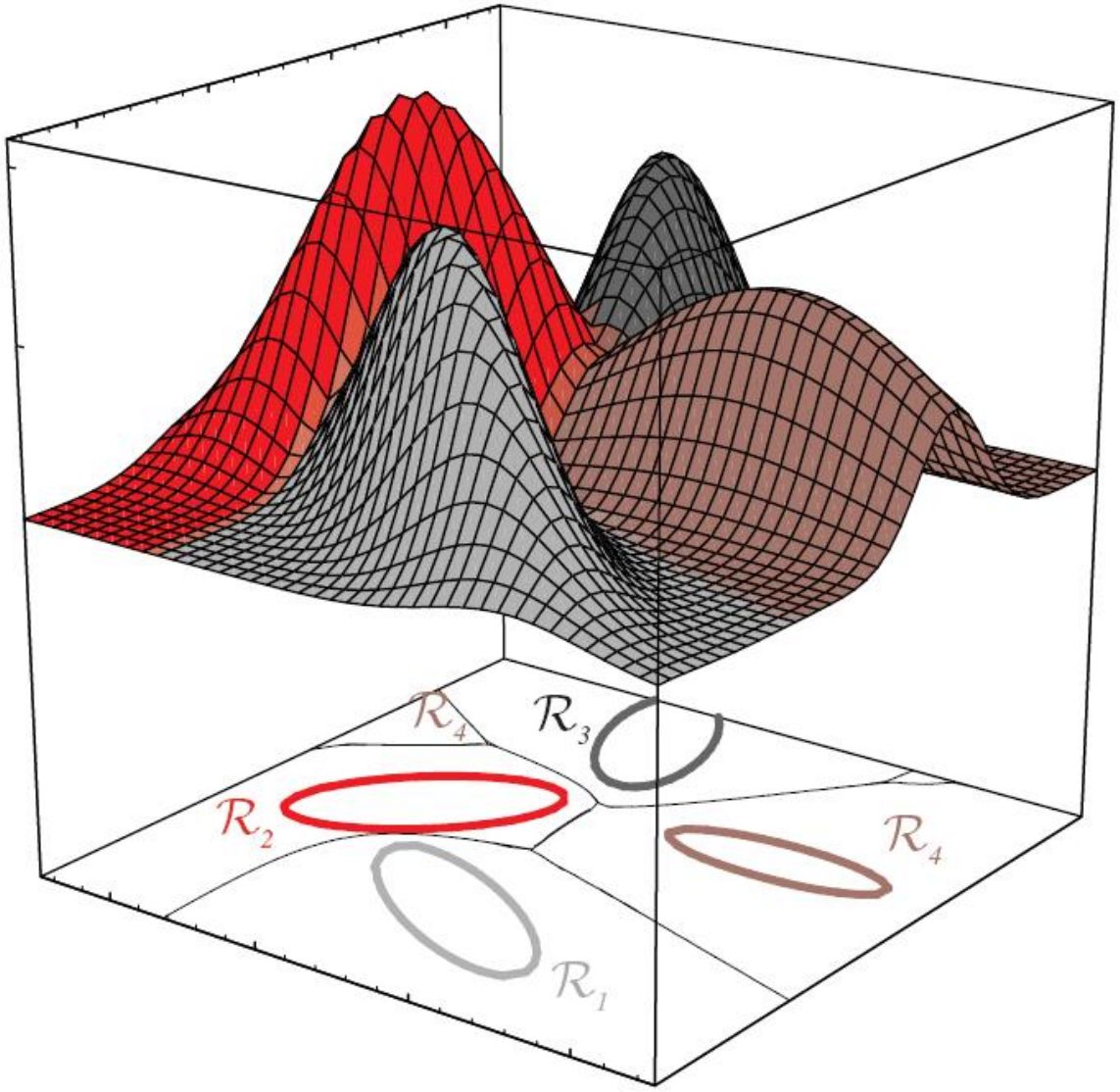
and its parameters ML estimated as

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{n=1}^{N_c}\mathbf{x}_{cn} \quad \boldsymbol{\Sigma}_c = \frac{1}{N_c}\sum_{n=1}^{N_c}(\mathbf{x}_{cn}-\boldsymbol{\mu}_c)(\mathbf{x}_{cn}-\boldsymbol{\mu}_c)^T$$

- Class posterior probability for new observations is obtained from the prior and class pdf-s using Bayes rule:

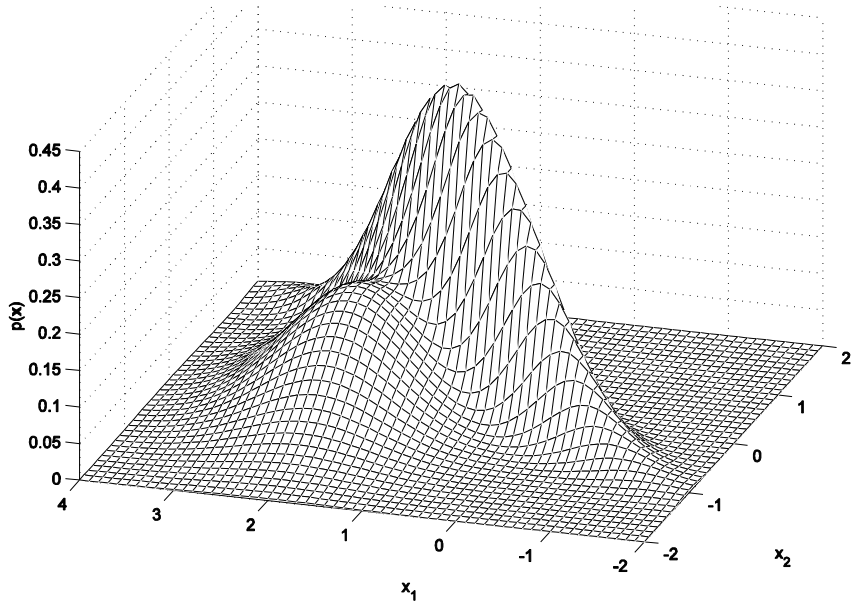
$$P(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)P(c)}{\sum_k p(\mathbf{x}|k)P(k)}$$




$\mathcal{R}_4$
$\mathcal{R}_3$
$\mathcal{R}_2$
$\mathcal{R}_4$
$\mathcal{R}_1$

Směs gaussovských rozložení
(Gaussian Mixture Model – GMM)

$$P(\mathbf{x}|\Theta) = \sum_c \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) P_c$$



kde

$$\Theta = \{P_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}$$

$$\sum_c P_c = 1$$

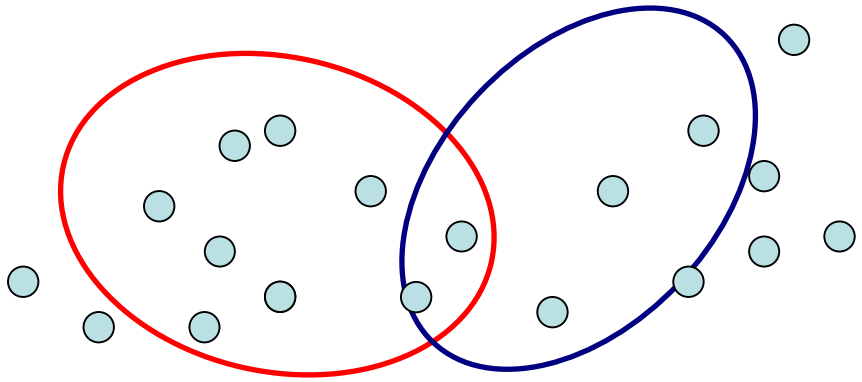
Směs gaussovských rozložení

$$P(\mathbf{x}) = \sum_c p(\mathbf{x}|c)P(c) = \sum_c \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)P_c$$

- Vzoreček můžeme chápat jen jako něco co definuje tvar funkce hustoty pravděpodobnosti…
- nebo jej můžeme vidět jako složitější generativní model,který generuje příznaky následujícím způsobem:
  - Napřed je jedna z gaussovských komponent vybrána tak aby respektovala apriorní pravděpodobnosti P_c
  - Příznakový vektor se generuje z vybraného gaussovského rozložení.
- Pro vyhodnoceni modelu ale nevíme, která komponenta příznakový vektor generovala a proto musíme marginalizovat (suma přes gaussovské komponenty násobené jejich “apriorními” pravděpodobnostmi)

Trénování GMM – Viterbi training

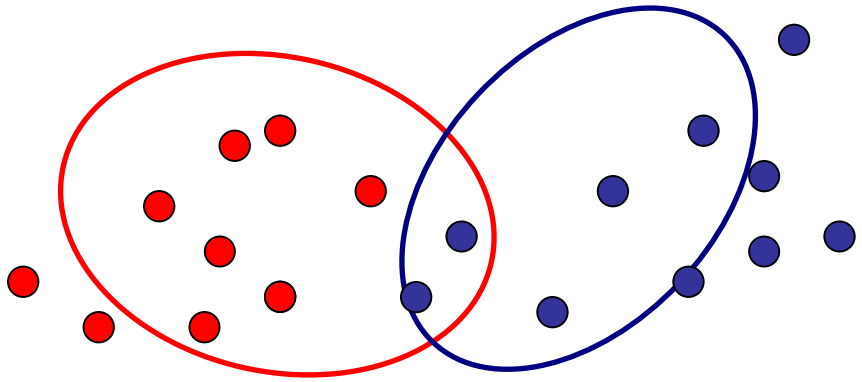
- Intuitivní ale nepřesný iterativní algoritmus pro ML trénování GMM parametrů



Trénování GMM – Viterbi training

- Intuitivní ale nepřesný iterativní algoritmus pro ML trénování GMM parametrů

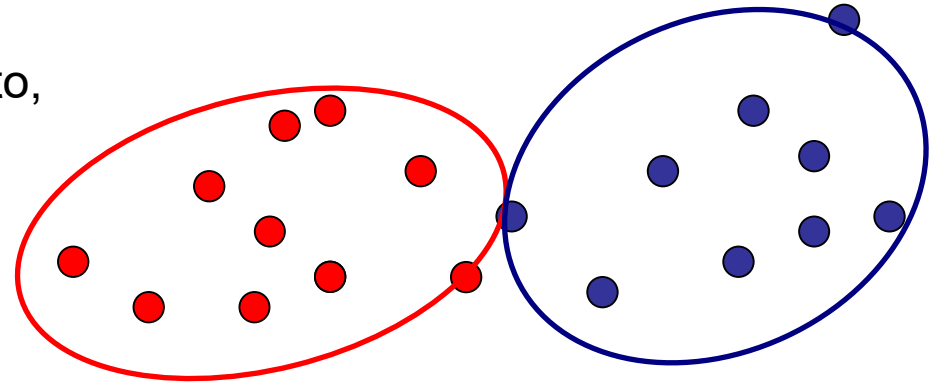
- Současným modelem klasifikujeme data jako kdyby by jednotlivé Gaussovky modelovaly různé třídy a váhy byly apriorní pravděpodobnosti tříd (přesto, že všechnadata patří do jedné třídy, kterou se snažíme modelovat).



Trénování GMM – Viterbi training

- Intuitivní ale nepřesný iterativní algoritmus pro ML trénování GMM parametrů

- Současným modelem klasifikujeme data jako kdyby by jednotlivé Gaussovky modelovaly různé třídy a váhy byly apriorní pravděpodobnosti tříd (přesto, že všechnadata patří do jedné třídy, kterou se snažíme modelovat).

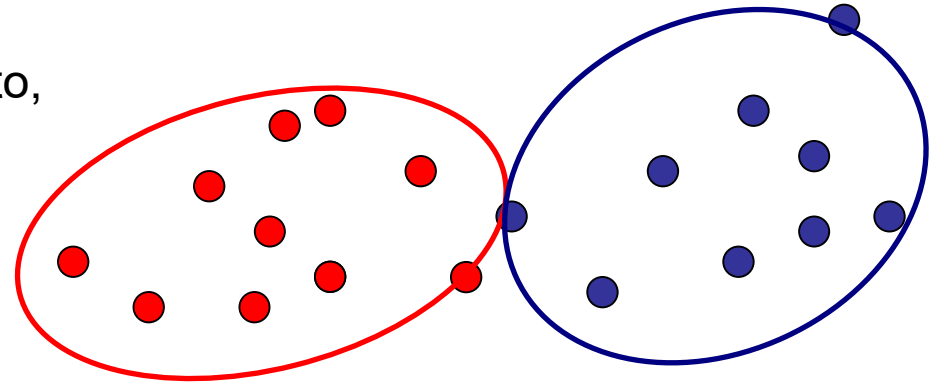


- Nové parametry každé Gaussovky odhadneme na datech k ní přiřazených v předchozím kroku. Nové váhy jsou dány poměry možství dat přiřazených Gausovkám.

Trénování GMM – Viterbi training

- Intuitivní ale nepřesný iterativní algoritmus pro ML trénování GMM parametrů

- Současným modelem klasifikujeme data jako kdyby by jednotlivé Gaussovky modelovaly různé třídy a váhy byly apriorní pravděpodobnosti tříd (přesto, že všechnadata patří do jedné třídy, kterou se snažíme modelovat).



- Nové parametry každé Gaussovky odhadneme na datech k ní přiřazených v předchozím kroku. Nové váhy jsou dány poměry možství dat přiřazených Gausovkám.

- Předchozí dva kroky opakujeme až do konvergence.

Trénování GMM – EM algorithm

- **Expectation Maximization** je iterativní algoritmus pro trénování různých generativních modelů se skrytými proměnnými (latent or hidden variables), jehož kazdá iterace vede ke zvýšení věrohodnosti (likelihood) trenovacích dat. Nezaručuje ale nalezení globalního optima.
- Zde ukazujeme pouze výsledek aplikace EM na trénování GMM.
- Algoritmus je podobny předchozímu Viterbi trénování, s tím rozdílem, že (místo tvrdých přiřazení) jsou data Gaussovkám přiřazena "měkce" pomocí vah – posteriorních pravděpodobností $P(c|\mathbf{x}_i)$ spočítaných současným modelem. Parametry $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$ jsou potom pocítány pomocí váhovaných (namísto prostých) průměrů.

$$\gamma_{ci} = \frac{\mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_c^{(old)}, \boldsymbol{\Sigma}_c^{(old)}\right) P_c^{(old)}}{\sum_k \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k^{(old)}, \boldsymbol{\Sigma}_k^{(old)}\right) P_k^{(old)}} = \frac{p(\mathbf{x}_i|c)P(c)}{\sum_k p(\mathbf{x}_i|k)P(k)} = P(c|\mathbf{x}_i)$$

$$\boldsymbol{\mu}_c^{(new)} = \frac{1}{\sum_i \gamma_{ci}} \sum_i \gamma_{ci} \mathbf{x}_i \qquad P_c^{(new)} = \frac{\sum_i \gamma_{ci}}{\sum_k \sum_i \gamma_{ki}}$$

$$\boldsymbol{\Sigma}_{\mathrm{c}}^{(new)} = \frac{1}{\sum_i \gamma_{ci}} \sum_i \gamma_{ci} (\mathbf{x}_{\mathrm{i}} - \boldsymbol{\mu}_c)(\mathbf{x}_{\mathrm{i}} - \boldsymbol{\mu}_c)^T$$