Přednáška 11: Bayesovská klasifikace; Úvod do grafických modelů a hierarchického učení

Ondřej Tichý

Bayesovské metody v machine learningu (BML)

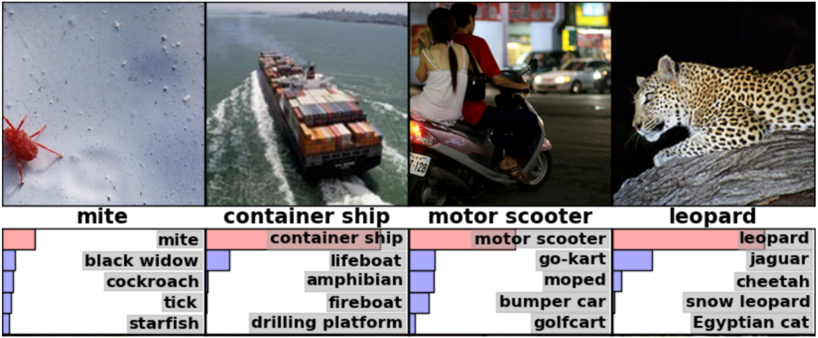
May 13, 2018

Obsah přednášky

- bayesovská klasifikace
  - principy
  - bayesian spam filtering
- grafické modely
  - hierarchické učení
  - demonstrace na maticové dekompozici

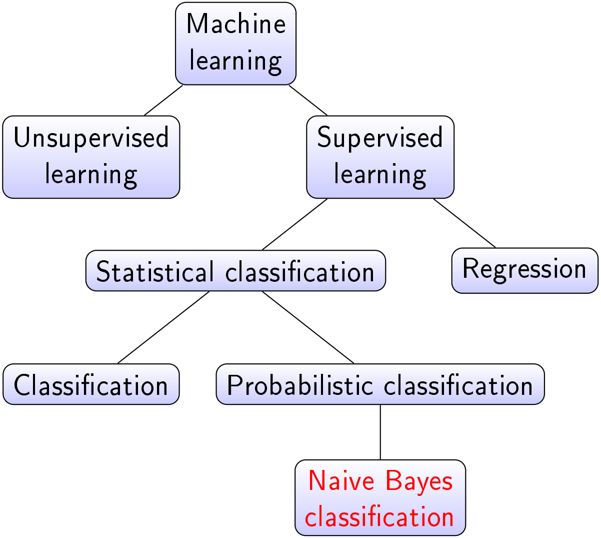
Bayesovská klasifikace

Úvod



Bayesovská klasifikace

Zařazení



Naive Bayes Klasifikátor

Princip

- mějme vektor pozorování $\mathbf{x} = (x_1, \ldots, x_n)$, který chceme přiřadit ke třídě C_k ze setu $\{C_1, \ldots, C_K\}$.
- teroreticky: máme tabulku pro příslušnost x_i k jednotlivým třídám a vyjádříme

$$p\left(C_k|\mathbf{x}\right), \tag{1}$$

  což není vždy vhodné a možné.

Naive Bayes Klasifikátor

Princip

- mějme vektor pozorování $\mathbf{x} = (x_1, \ldots, x_n)$, který chceme přiřadit ke třídě C_k ze setu $\{C_1, \ldots, C_K\}$.
- teroreticky: máme tabulku pro příslušnost x_i k jednotlivým třídám a vyjádříme
$$p\left(C_k|\mathbf{x}\right), \tag{1}$$
což není vždy vhodné a možné.
- s využitím Bayesova teorému
$$p\left(C_k|\mathbf{x}\right) = \frac{p\left(C_k\right) p\left(\mathbf{x}|C_k\right)}{p\left(\mathbf{x}\right)} \tag{2}$$
  - $p\left(C_k|\mathbf{x}\right)$ - aposteriorní rozdělení
  - $p\left(C_k\right)$ - apriorní rozdělení
  - $p\left(\mathbf{x}|C_k\right)$ - model (věrohodnost)
  - $p\left(\mathbf{x}\right)$ - marginála přes $\mathbf{x}$ (nezávisí na C_k)

Naive Bayes Klasifikátor

Princip

$$p\,(C_k|\mathbf{x}) \propto p\,(C_k)\,p\,(\mathbf{x}|\,C_k) = p\,(x_1, x_2, \ldots, x_n, C_k) \tag{3}$$

- aplikujeme řetězové pravidlo (n-krát)

$$\begin{aligned} p\,(x_1, x_2, \ldots, x_n, C_k) =& p\,(x_1|x_2, \ldots, x_n, C_k) \times \\ & \times\, p\,(x_2|x_3, \ldots, x_n, C_k) \ldots p\,(x_n|\,C_k)\,p\,(C_k) \end{aligned}$$

Naive Bayes Klasifikátor

Princip

$$p\left(C_k|\mathbf{x}\right) \propto p\left(C_k\right) p\left(\mathbf{x}|C_k\right) = p\left(x_1, x_2, \ldots, x_n, C_k\right) \tag{3}$$

- aplikujeme řetězové pravidlo (n-krát)

$$\begin{aligned} p\left(x_1, x_2, \ldots, x_n, C_k\right) =& p\left(x_1|x_2, \ldots, x_n, C_k\right) \times \\ & \times p\left(x_2|x_3, \ldots, x_n, C_k\right) \ldots p\left(x_n|C_k\right) p\left(C_k\right) \end{aligned}$$

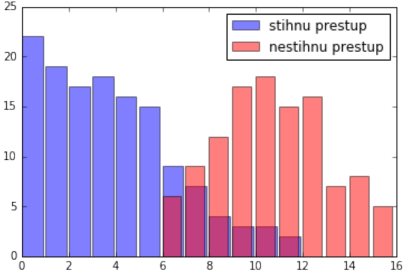
- proč "naivní" Bayes: předpokládáme podmíněnou nezávislost mezi prvky vektoru **x**:

$$p\left(x_1, x_2, \ldots, x_n, C_k\right) = p\left(x_1|C_k\right) p\left(x_2|C_k\right) \ldots p\left(x_n|C_k\right) p\left(C_k\right)$$

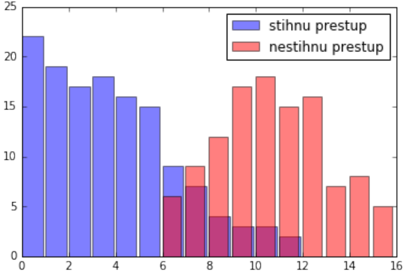
- pravděpodobnost příslušnosti k dané třídě:

$$p\left(C_k|\mathbf{x}\right) \propto p\left(C_k\right) \prod_{i=1}^{n} p\left(x_i|C_k\right) \tag{4}$$

Naive Bayes Klasifikátor

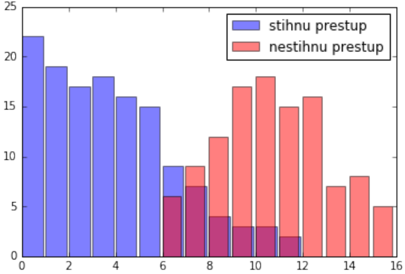


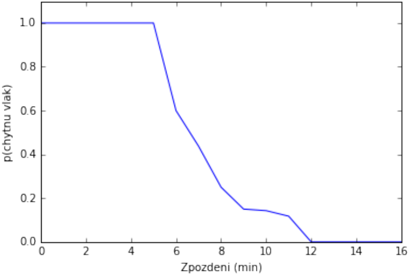
Naive Bayes Klasifikátor



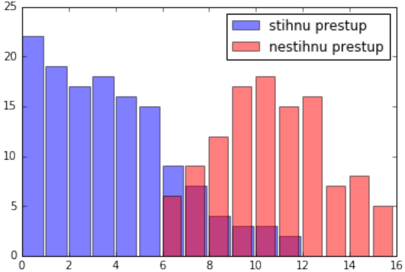
$$p(\text{stihnu}|\text{zpozdeni} = 6) = \frac{9}{9+6} = 0.6 \tag{5}$$

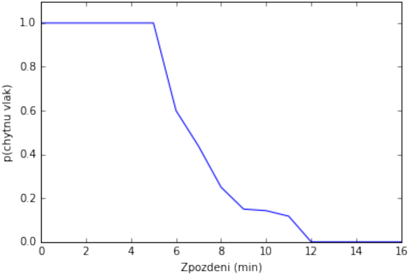
Naive Bayes Klasifikátor





Naive Bayes Klasifikátor





- MAP klasifikace: $\widehat{k} = \arg\max_{k\in\{1,\dots,K\}} p\left(C_k\right) \prod_{i=1}^{n} p\left(x_i | C_k\right)$

Naive Bayes Klasifikátor

Jednoduchý příklad 2

- předpokládejme nákupní košík podle pohlaví

	masový výrobek	mléčný výrobek	pečivo
muž (M)	40%	10%	50%
žena (Z)	10%	70%	20%

- chceme určit pohlaví nakupujícího podle nákupního košíku, v němž je
  - 1 položka masový výrobek
  - 1 položka mléčný výrobek

Naive Bayes Klasifikátor

Jednoduchý příklad 2

- předpokládejme nákupní košík podle pohlaví

	masový výrobek	mléčný výrobek	pečivo
muž (M)	40%	10%	50%
žena (Z)	10%	70%	20%

- chceme určit pohlaví nakupujícího podle nákupního košíku, v němž je
  - 1 položka masový výrobek
  - 1 položka mléčný výrobek
- přímá aplikace:

$$\begin{aligned} p(M|\text{maso, mleko}) &\propto p(\text{maso, mleko}|M)p(M) = \\ &= 0.4 \times 0.1 \times 0.5 = 0.02 \\ p(Z|\text{maso, mleko}) &\propto p(\text{maso, mleko}|Z)p(Z) = \\ &= 0.1 \times 0.7 \times 0.5 = 0.035 \end{aligned}$$

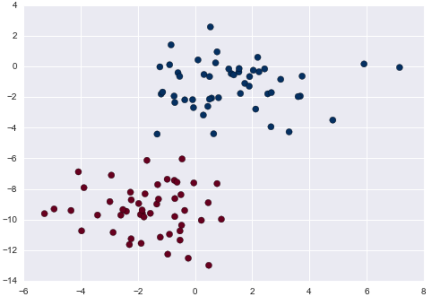
tzn. na 36% muž a na 64% žena.

Naive Bayes a další apriorní předpoklady

Apriorno

- pro třídy $p(C_k)$
- pro model $p(\mathbf{x}|C_k)$:
  - apriorní předpoklad normálního rodělení: pro C_k mějme příslušné μ_k a σ_k^2, pak

$$p(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \tag{6}$$

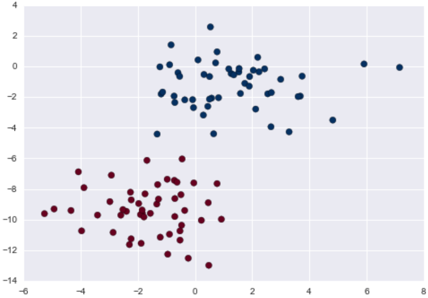


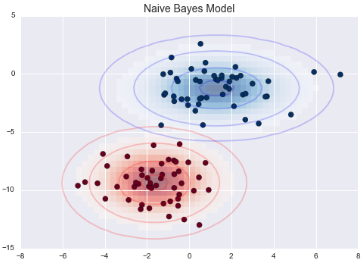
Naive Bayes a další apriorní předpoklady

Apriorno

- pro třídy $p(C_k)$
- pro model $p(\mathbf{x}|C_k)$:
  - apriorní předpoklad normálního rodělení: pro C_k mějme příslušné μ_k a σ_k^2, pak

$$p(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right) \tag{6}$$





Naive Bayes a další apriorní předpoklady

Apriorno

- multinomiální rozdělení: parametrizováno vektorem (pro $k$tou třídu)

$$\theta_{C_k} = \left(\theta_{C_k 1}, \ldots, \theta_{C_k n}\right), \tag{7}$$

  - n je počet příznaků (např. délka slovníku v klasifikaci dokumentů)
  - $\theta_{C_k i}$ je pravděpodobnost $p\left(x_i | C_k\right)$, tzn. že se např. slovo objeví v dané třídě (pomocí počtu výskytů)

- Bernoulli rozdělení: jako multinomiální, ale výskyt ano-ne

$$p\left(x_i | C_k\right) = p\left(i | C_k\right) x_i + \left(1 - p\left(i | C_k\right)\right)\left(1 - x_i\right). \tag{8}$$

  - x_i je binární
  - $p\left(i | C_k\right)$ je výskyt $i$tého slova

Naive Bayes

Spam filtering

- Jak poznat spam od legitimní zprávy?

Naive Bayes

Spam filtering

- Jak poznat spam od legitimní zprávy?
- náš příklad: podle slov, které obsahuje.
- označení: S - “spam”, H - “ham”.
- pravděpodobnost, že email obsahující slovo “viagra” je spam:

Naive Bayes

Spam filtering

- Jak poznat spam od legitimní zprávy?
- náš příklad: podle slov, které obsahuje.
- označení: S - “spam”, H - “ham”.
- pravděpodobnost, že email obsahující slovo “viagra” je spam:

$$p\left(S|\text{"viagra"}\right) = \frac{p\left(\text{"viagra"}|S\right) p\left(S\right)}{p\left(\text{"viagra"}|S\right) p\left(S\right) + p\left(\text{"viagra"}|H\right) p\left(H\right)}$$

- samozřejmě, jedno slovo nestačí...

Naive Bayes

Spam filtering

- pokud předpokládáme nezávislý výskyt N slov v textu: $S_1, \ldots, S_N$
- potom [Graham (2002): A Plan for Spam]:

$$p\left(S|S_1, \ldots, S_N\right) = \frac{p\left(S|S_1\right) \ldots p\left(S|S_N\right)}{p\left(S|S_1\right) \ldots p\left(S|S_N\right) + \left(1 - p\left(S|S_1\right)\right) \ldots \left(1 - p\left(S|S_N\right)\right)}$$

- problém s neznámými nebo s velmi málo vyskytujícími se slovy

Naive Bayes

Spam filtering

- pokud předpokládáme nezávislý výskyt N slov v textu: $S_1, \ldots, S_N$
- potom [Graham (2002): A Plan for Spam]:

$$p(S|S_1, \ldots, S_N) = \frac{p(S|S_1) \ldots p(S|S_N)}{p(S|S_1) \ldots p(S|S_N) + (1 - p(S|S_1)) \ldots (1 - p(S|S_N))}$$

- problém s neznámými nebo s velmi málo vyskytujícími se slovy...můžeme:
  - ignorovat
  - korigovat pravděpodobnost

$$\widetilde{p}(S|\text{"lihovina"}) = \frac{\check{c} \cdot p(S) + n \cdot p(S|\text{"lihovina"})}{\check{c} + n}, \qquad (9)$$

kde $\check{c}$ je zvolené číslo a n je počet výskytů během učení klasifikátoru.

Naive Bayes

Spam filtering - další možnosti

- zpracování jen daného počtu slov s největší informační hodnotou
- zpracování delších řetězců
- výhody:
  - relativně jednoduchý princip
  - pro každého uživatele jiné váhy a tedy specifické chování
- nevýhody:

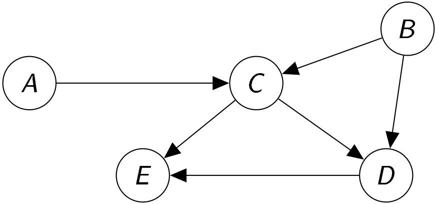
Naive Bayes

Spam filtering - další možnosti

- zpracování jen daného počtu slov s největší informační hodnotou
- zpracování delších řetězců
- výhody:
  - relativně jednoduchý princip
  - pro každého uživatele jiné váhy a tedy specifické chování
- nevýhody:
  - Bayesian poisoning (např. přidání spousty legitivních slov k textu spamu)
  - necitlivé na záměnu písmene ve slovu ("gooqle" přečteme stále jako "google")
  - obrázky

Grafické modely

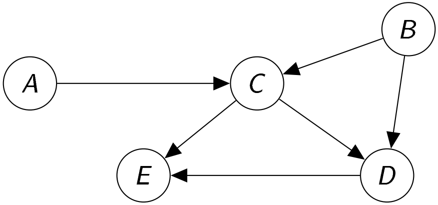
Proč? Jak?



- každý bod grafu reprezentuje náhodnou proměnnou
- hrany reprezentují statistickou závislost mezi proměnnými

Grafické modely

Proč? Jak?



- každý bod grafu reprezentuje náhodnou proměnnou
- hrany reprezentují statistickou závislost mezi proměnnými
- ekvivalentní zápis:

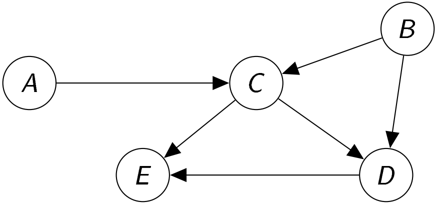
$$p(A, B, C, D, E) = p(A)\, p(B)\, p(C|A, B)\, p(D|B, C)\, p(E|C, D)$$

- formálně:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p\left(X_i | X_{\text{parents}(i)}\right) \tag{10}$$

Grafické modely

Proč? Jak?

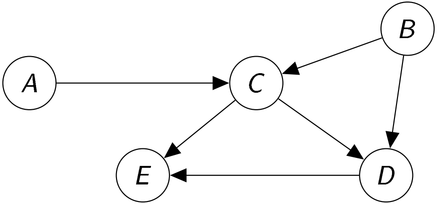


Jaké to má výhody?

- intuitivní reprezentace a vizualizace vztahů a vazeb mezi proměnnými
- vazby jsou přímo viditelné
  - např. otázka "Je A závislé na B (s předpokladem znalostí hodnot uzlu C)?"
- možnost tzv. message-passing přístupu
  - určení $p\left(A|C=c\right)$?

Grafické modely

Proč? Jak?



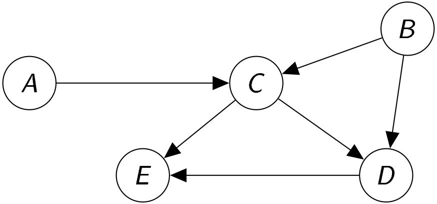
$$p\left(A|C=c\right)=\frac{p\left(A,C=c\right)}{p\left(C=c\right)} \tag{11}$$

- naivně:

$$p\left(A,C=c\right)=\sum_{B,D,E}p\left(A,B,C=c,D,E\right) \tag{12}$$

Grafické modely

Proč? Jak?



$$p\,(A|C=c) = \frac{p\,(A, C=c)}{p\,(C=c)} \tag{13}$$

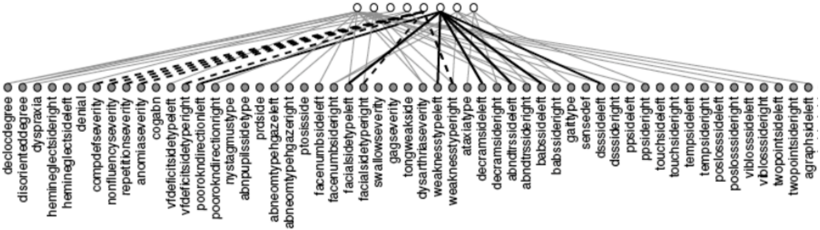
- efektivně:

$$\begin{aligned}
p\,(A, C=c) &= \sum_{B,D,E} p\,(A)\,p\,(B)\,p\,(C=c|A,B)\,p\,(D|B,C=c)\,p\,(E|C=c,D) \\
&= \sum_{B} p\,(A)\,p\,(B)\,p\,(C=c|A,B) \sum_{D} p\,(D|B,C=c) \sum_{E} p\,(E|C=c,D) \\
&= \sum_{B} p\,(A)\,p\,(B)\,p\,(C=c|A,B)
\end{aligned}$$

Grafické modely

Použití

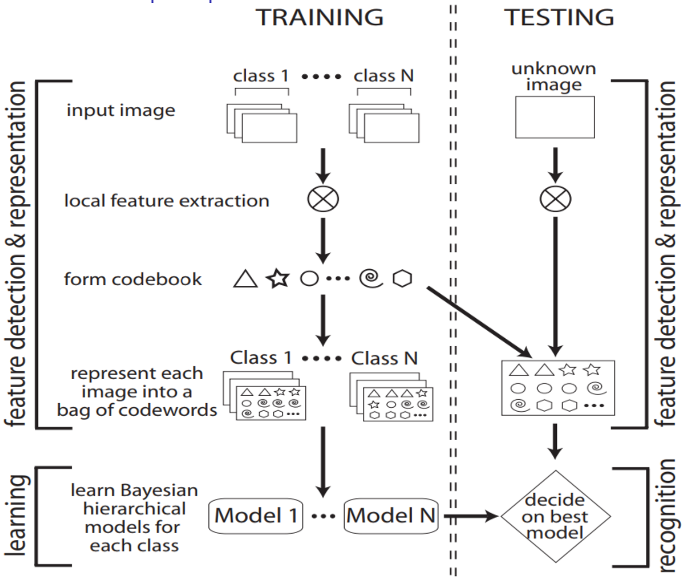
- rozpoznání a analýza řeči
- rozpoznání a analýza obrazu
- aplikace v medicíně (závislosti atd.)



[Ghahramani et all]

Naive Bayes

Struktura reálného přístupu



[Fei-Fei et all]

- Mějme pozorování (data) D. Jak se "naučit" parametry $\boldsymbol{\theta}$ z D?
- Příklad na maticové dekompozici:

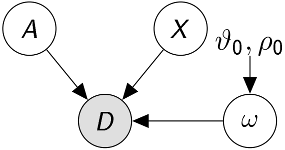
$$\begin{aligned} f\left(D|A,X,\omega\right) =& \mathcal{N}\left(AX^{T},\omega^{-1}I_{p}\otimes I_{n}\right), \\ f(\omega) =& \mathcal{G}\left(\vartheta_{0},\rho_{0}\right), \\ f\left(A\right) =& \mathcal{N}\left(\mathbf{0},I_{p}\otimes I_{r}\right), \\ f(X) =& \mathcal{N}\left(\mathbf{0},I_{n}\otimes I_{r}\right). \end{aligned}$$

Grafické modely a hierarchické učení

Příklad na maticové dekompozici

- Mějme pozorování (data) D. Jak se "naučit" parametry $\boldsymbol{\theta}$ z D?
- Příklad na maticové dekompozici:

$$\begin{aligned} f(D|A,X,\omega) =& \mathcal{N}\left(AX^T, \omega^{-1} I_p \otimes I_n\right), \\ f(\omega) =& \mathcal{G}(\vartheta_0, \rho_0), \\ f(A) =& \mathcal{N}(\mathbf{0}, I_p \otimes I_r), \\ f(X) =& \mathcal{N}(\mathbf{0}, I_n \otimes I_r). \end{aligned}$$

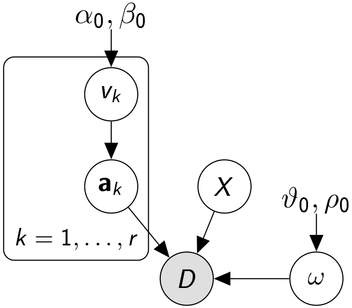


- Zkoumali jsme odhad počtu komponent - váhy $v_1, \ldots, v_r$ na obrázky

$$
\begin{aligned}
f\left(D|A,X,\omega\right) = \mathcal{N}\left(AX^T, \omega^{-1} I_p \otimes I_n\right),\\
f(\omega) =& \mathcal{G}\left(\vartheta_0, \rho_0\right),\\
f\left(A|V\right) =& \mathcal{N}\left(\mathbf{0}, I_p \otimes V^{-1}\right)\\
f(v_k) =& \mathcal{G}(\alpha_0, \beta_0),\\
f(X) =& \mathcal{N}\left(\mathbf{0}, I_n \otimes I_r\right).
\end{aligned}
$$

- Zkoumali jsme odhad počtu komponent - váhy $v_1, \dots, v_r$ na obrázky

$$
\begin{aligned}
f\left(D|A,X,\omega\right) = \mathcal{N}\left(AX^T, \omega^{-1} I_p \otimes I_n\right), \\
f(\omega) = &\mathcal{G}\left(\vartheta_0, \rho_0\right), \\
f\left(A|V\right) = &\mathcal{N}\left(\mathbf{0}, I_p \otimes V^{-1}\right) \\
f(v_k) = &\mathcal{G}(\alpha_0, \beta_0), \\
f(X) = &\mathcal{N}\left(\mathbf{0}, I_n \otimes I_r\right).
\end{aligned}
$$

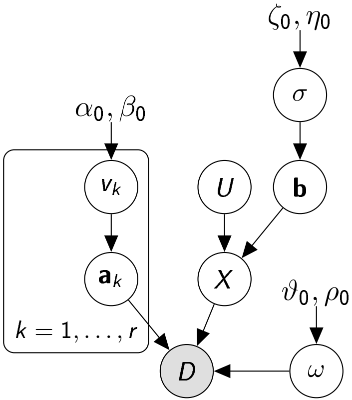


- Říkali jsme si o konvolučním modelu křivek

$$
\begin{aligned}
f\left(D|A,X,\omega\right) &= \mathcal{N}\left(AX^T,\omega^{-1}I_p\otimes I_n\right),\\
f(\omega) &= \mathcal{G}\left(\vartheta_0,\rho_0\right),\\
f\left(A|V\right) &= \mathcal{N}\left(\mathbf{0},I_p\otimes V^{-1}\right)\\
f(v_k) &= \mathcal{G}(\alpha_0,\beta_0),\\
X &= BU,\\
f(U) &= \mathcal{N}\left(\mathbf{0},I_n\otimes I_r\right),\\
f(b|\sigma) &= \mathcal{N}\left(\mathbf{0},\sigma^{-1}I_n\right),\\
f(\sigma) &= \mathcal{G}\left(\zeta_0,\eta_0\right).
\end{aligned}
$$

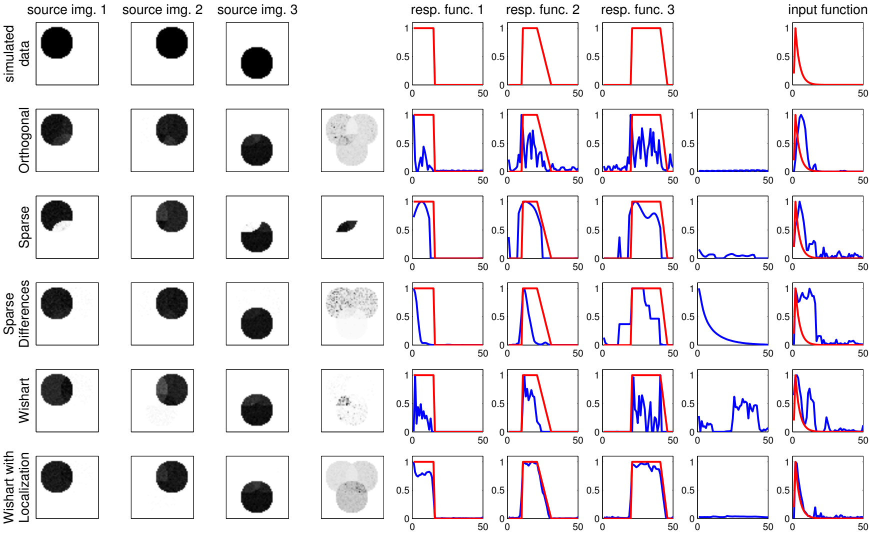
- Říkali jsme si o konvolučním modelu křivek

$$
\begin{aligned}
f\left(D|A,X,\omega\right) &= \mathcal{N}\left(AX^T, \omega^{-1} I_p \otimes I_n\right), \\
f(\omega) &= \mathcal{G}\left(\vartheta_0, \rho_0\right), \\
f\left(A|V\right) &= \mathcal{N}\left(\mathbf{0}, I_p \otimes V^{-1}\right) \\
f(v_k) &= \mathcal{G}(\alpha_0, \beta_0), \\
X &= BU, \\
f(U) &= \mathcal{N}\left(\mathbf{0}, I_n \otimes I_r\right), \\
f(b|\sigma) &= \mathcal{N}\left(\mathbf{0}, \sigma^{-1} I_n\right), \\
f(\sigma) &= \mathcal{G}\left(\zeta_0, \eta_0\right).
\end{aligned}
$$



Grafické modely a hierarchické učení

Aplikace na cvičná data



Grafické modely a hierarchické učení

Aplikace na data z renální scintigrafie