



Teorie kognitivních systémů

6 Bayesovské učení

- Základy počtu pravděpodobnosti
- Podmíněná pravděpodobnost
- Bayesova věta
- Optimální a naivní bayesovský klasifikátor
- strategie výběru hypotéz (MAP, ML, MDL)
- Aplikace NBK





Dvě elementární role bayesovských metod

Na bayesovských metodách jsou postaveny dva užitečné a **praktické učící se algoritmy**:

- **Naivní bayesovský klasifikátor (*Naïve Bayes Classifier*)**
- **Bayesovské sítě (*Bayesian Belief Networks*)**
 - odhad parametrů a učení (tj. automatické poznání závislostní struktury)

Bayesovské metody poskytují užitečný výchozí **koncepční rámec**:

- poskytují referenční standard (*Gold Standard*) k hodnocení výkonu jiných učících se algoritmů – bayesovský optimální klasifikátor (*Bayes Optimal Classifier, BOC*)
- jedna z možností implementace Occamovy břitvy (*Minimum Description Length, MDL*)





Bayesovské učení

Pravděpodobnostní rámec usuzování

Bayesovské odvozování poskytuje pravděpodobnostní aparát k usuzování: **Nejlepší hypotéza je ta, která je nejvíce pravděpodobná s ohledem na pozorovaná data a apriorní znalost** (ta je vyjádřena apriorní p-stí konkrétní hypotézy)

- pozorovaná data → zvýšení/snížení odhadu p-sti hypotézy (hypotéza je pak odolnější vůči zašuměným datům)
- lze provádět pravděpodobnostní předpovědi (jaká je šance?)
- kombinuje předpovědi více hypotéz podle jejich p-stí
- kombinuje apriorní znalost (apriorní pravděpodobnosti hypotéz) s pozorovanými daty





Základní vzorce pravděpodobnosti Kolmogorovovy axiomy

Ω značí **universum** (Sample Space), tj. množinu všech hodnot x , jichž může nabývat náhodná proměnná X , tj. $X: \Omega$.

Kolmogorovovy axiomy pak říkají, že

$$\forall x \in \Omega : 0 \leq P(X = x) \leq 1 \quad (1)$$

$$P(\Omega) \equiv \sum_{x \in \Omega} P(X = x) = 1 \quad (2)$$

$$\forall X_1, X_2, \dots, X_n : i \neq j \Rightarrow X_i \wedge X_j = 0 \quad (3)$$

$$P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i) \quad (4)$$



Základní vzorce pravděpodobnosti pro potřeby bayesovské klasifikace

Pravděpodobnost $P(A \wedge B)$ konjunkce jevů A a B:

$$P(A \wedge B) = P(A|B) P(B) = P(B|A) P(A)$$

Pravděpodobnost $P(A \vee B)$ disjunkce jevů A a B:

$$P(A \wedge B) = P(A) + P(B) - P(A \wedge B)$$

Věta o úplné pravděpodobnosti: Jsou-li jevy A_1, \dots, A_n disjunktní a tvoří úplný systém, tj. $P(A_1) + P(A_2) + \dots + P(A_n) = 1$, pak:

$$P(B) = \sum_{i=1}^n P(B|A_i) P(A_i)$$



Podmíněná pravděpodobnost

Definice

Podmíněná p-st (Conditional Probability) –

při podmínění se redukuje množina všech možných výsledků náhodného pokusu A jen na ty výsledky, které vyhovují podmínce B.

P-st jevu A podmíněná jevem B (stručněji **podmíněná p-st jevu A za podmínky B**), kde $P(B) > 0$, je hodnota:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Při neměnném podmiňujícím jevu má podmíněná p-st stejné vlastnosti jako p-st bez podmiňování.



Statistická nezávislost

Intuitivní odvození

Mějme dva jevy:

- A: Je středa.
- B: Je slunečno.

}

zřejmě →

$$P(B|A) = P(B)$$

protože je nemyslitelné, že by počasí záviselo na dni v týdnu...

Aplikujme poznatky z teorie pravděpodobnosti:

$$P(\neg B|A) = P(\neg B)$$

$$P(A|B) = P(A)$$

$$P(A \cap B) = P(A) P(B)$$

$$P(\neg A \cap B) = P(\neg A) P(B)$$

$$P(A \cap \neg B) = P(A) P(\neg B)$$

$$\begin{aligned} P(\neg A \cap \neg B) &= \\ &= P(\neg A) P(\neg B) \end{aligned}$$

Jevy A a B jsou **nezávislé** tehdy, je-li $P(A|B) = P(A|\neg B) = P(A)$, tj. $P(A)$ je stejná, ať B nastane nebo nenastane nebo o B vůbec nic nevíme...





Bayesova věta

Základ bayesovské klasifikace

Bayesova věta (Bayes' Theorem) – kvanitifikuje souvislost podmíněné p-stí nějakého jevu s opačnou podmíněnou p-stí.

Mějme dva náhodné jevy A a B s p-stmi $P(A)$ a $P(B)$, $P(B) > 0$.
Potom platí:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

$P(A|B)$ je podmíněná p-st jevu A za předpokladu, že nastal jev B, a naopak $P(B|A)$ je podmíněná p-st jevu B za předpokladu, že nastal jev A.

Této souvislosti si nezávisle na Thomasi Bayesovi (1702–1761) povšiml roku 1774 i Pierre-Simon Laplace...



Bayesova věta

Poněkud morbidní příklad

Pacient podstoupí vyšetření k prokázání choroby, jehož **výsledek je pozitivní**. Zároveň známe následující statistické údaje:

- **pozitivní spolehlivost testu je 98%**, tj. je-li výsledek testu pozitivní, pacient v 98% případů skutečně chorobu má
- **negativní spolehlivost testu je 97%**, tj. je-li výsledek testu negativní, pacient v 97% skutečně chorobu nemá
- dlouhodobé výzkumy ukazují, že touto chorobou trpí 0.8% populace



$$\begin{array}{ll} P(\text{choroba}) = 0.008 & P(\neg \text{choroba}) = 0.992 \\ P(+ | \text{choroba}) = 0.98 & P(- | \text{choroba}) = 0.02 \\ P(+ | \neg \text{choroba}) = 0.03 & P(- | \neg \text{choroba}) = 0.97 \end{array}$$



Bayesova věta

Poněkud morbidní příklad

Trpí tedy pacient touto chorobou?

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

$$P(\text{choroba}) = 0.008 \quad P(\neg\text{choroba}) = 0.992$$

$$P(+|\text{choroba}) = 0.98 \quad P(-|\text{choroba}) = 0.02$$

$$P(+|\neg\text{choroba}) = 0.03 \quad P(-|\neg\text{choroba}) = 0.97$$

$$P(\text{choroba}|+) = \frac{P(+|\text{choroba}) P(\text{choroba})}{P(+)} = \\ = \frac{0.98 * 0.008}{P(+)} = \mathbf{0.00784} / P(+)$$

$$P(\neg\text{choroba}|+) = \frac{P(+|\neg\text{choroba}) P(\neg\text{choroba})}{P(+)} = \\ = \frac{0.03 * 0.992}{P(+)} = \mathbf{0.02976} / P(+)$$





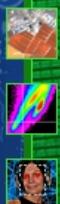
Bayesova věta – rozšířená podoba

Základ bayesovské klasifikace

Nechť A_1, \dots, A_n je úplný systém jevů a B je libovolný jev. Pak pro $k = 1, \dots, n$ platí, že:

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^n P(A_i)P(B|A_i)}$$

Bayesova věta se také nazývá **věta o pravděpodobnosti příčin**, protože se často používá v případech, kdy chceme z p-sti výskytu jevu B učinit závěry o p-stech jeho možných příčin A_k .





Bayesova věta

Bayesovská (epistemologická) interpretace

Pravděpodobnost je v B. I. chápána jako míra důvěry (*Degree of Belief*): Bayesova věta tedy dává do souvislosti míru důvěry ve výrok **před** a **po** uvedení důkazu.

Mějme výrok A a důkaz B:

- $P(A)$ – **prior** – je počáteční důvěra ve výrok A
- $P(A|B)$ – **posterior** – je důvěra ve výrok A po provedení B
- $P(B|A) / P(B)$ – je míra „podpory“ důkazu B výroku A

Na základě bayesovské interpretace Bayesovy věty lze provádět tzv. **bayesovské odvozování** (*Bayesian Inference*) či bayesovskou statistiku.

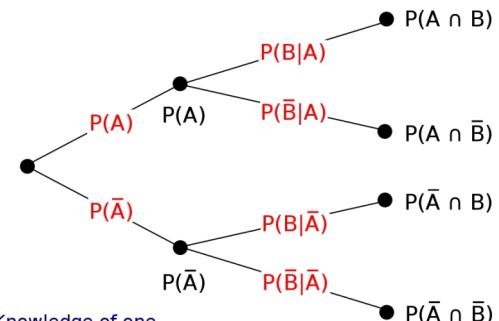




Bayesova věta

Frekventistická (standardní) interpretace

Pravděpodobnost je ve F. I. chápána jako vyjádření podílu výsledků náhodného jevu (tj. jejich relativní četnosti při počtu pokusů → ∞)

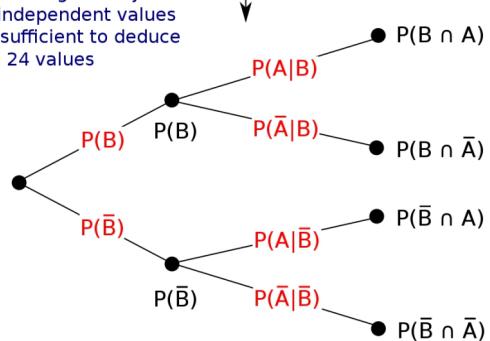


Knowledge of one diagram is sufficient to deduce the other

Use Bayes' Theorem to convert between diagrams

$$P(\alpha|\beta) P(\beta) = P(\alpha \cap \beta) = P(\beta|\alpha) P(\alpha)$$

Knowledge of any 3 independent values is sufficient to deduce all 24 values





Bayesova věta

Základ bayesovské klasifikace

Bayesovu větu lze zformulovat také v běžných pojmech takto:

$$\text{aposteriorní p-st} = \frac{\text{apriorní p-st} \times \text{p-st jevu}}{\text{důkaz (data)}}$$

Při aplikaci do oblasti strojového učení, tj. výběru hypotézy při znalosti nějakých trénovacích dat, dostáváme tuto podobu:

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)} = \frac{P(h \wedge D)}{P(D)}$$

$P(h)$ – apriorní p-st hypotézy h

$P(D)$ – apriorní p-st trénovacích dat D

$P(h|D)$ – p-st hypotézy h , podmíněná pozorováním dat D

$P(D|h)$ – p-st dat D , podmíněná hypotézou h , tj. p-st, že „uvidíme“ data D , platí-li hypotéza h (generativní model)

aposteriorní p-st
hypotézy h



Naivní bayesovský klasifikátor

Typické příklady užití NBK

Klasifikace textů pomocí NBK –

zařazení textu do třídy podle tématu (např. na články o politice, počasí, sportu, atd.). NBK patří mezi **nejlepší známé učící se algoritmy** ke klasifikaci textů (\rightarrow *Multinomial Naive Bayes*).

Filtrování spamu –

nejlepší známé užití NBK; klasifikace textu do dvou disjunktivních tříd SPAM a HAM, které tvoří úplný systém (*tertium non datur*).

Doporučovací systémy (*Recommender Systems*) –

kombinace data miningu a strojového učení ke zpracování dat uživatele (např. zpráv na sociálních sítích) s cílem predikovat, zda takový uživatel má zájem o určitý zdroj (služby, zboží).



Strategie výběru hypotézy

Volba hypotézy bayesovskými prostředky

- **Bayesova věta** –
pro výběr hypotézy (tj. klasifikaci neznámého vzorku) lze použít přímo Bayesovy věty, ale obvykle nejsme schopni vyjádřit $P(D)$ přímo (a i kdybychom byli, bude to pořád jen **odhad**)...
- **MAP** – Maximum A Posteriori → nejpravděpodobnější hypotéza podmíněná pozorovanými vstupními daty.
- **ML** – Maximum Likelihood → nejvěrohodnější hypotéza za předpokladu rovnoměrné distribuce hypotéz, což umožňuje výpočet dále zjednodušit.
- **MDL** – Minimum Description Length → Occamova břitva

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)}$$



Maximum a posteriori (MAP)

Výběr hypotézy s maximální aposteriorní p-stí

Obecně hledáme **nejpravděpodobnější hypotézu**, jsou-li dány trénovací data:

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)}$$

$$h_{MAP} = \arg \max_{h \in H} P(h|D)$$

$$= \arg \max_{h \in H} \frac{P(D|h) P(h)}{P(D)}$$

$$= \arg \max_{h \in H} P(D|h) P(h)$$

}

tj. vybereme hypotézu, která maximalizuje tento výraz

p-st pozorovaných dat za podmínky dané hypotézy (např. spolehlivost testu na průkaz onemocnění)

apriorní p-st hypotézy **obvykle známe** (např. výzkumem zjištěnou frekvenci výskytu onemocnění)



Maximum Likelihood (ML)

Výběr hypotézy s maximální věrohodností

Uvážíme-li při výběru hypotézy s maximální aposteriorní p-stí

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)}$$

$$h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)$$

situaci, kdy jsou všechny hypotézy stejně pravděpodobné, tj. $P(h_i) = P(h_j), \forall i, j$, redukuje se výběr na tu hypotézu, při které – pokud by nastala – bychom s největší pravděpodobností pozorovali data D :

$$h_{ML} = \arg \max_{h_i \in H} P(D|h_i)$$

→ **hypotéza s maximální věrohodností (Maximum Likelihood (ML) Hypothesis)**





Minimum Description Length (MDL)

Výběr hypotézy s minimální délkou kódu

Occamova břitva: Dáváme přednost „nejkratší“ hypotéze

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)}$$

MDL: Vybíráme takovou hypotézu h , která minimalizuje výraz

$$h_{MDL} = \operatorname{argmin}_{h \in H} L_{C_1}(h) + L_{C_2}(D|h)$$

kde $L_C(x)$ je délka zakódování x pomocí kódu C .

Příklad: H = rozhodovací stromy, D = odpovědi učitele (tags)

- $L_{C_1}(h)$ je počet bitů nutných k popsání stromu h
- $L_{C_2}(D|h)$ je počet bitů nutných k popsání D za podmínky h , což je rovno 0, jsou-li vzorky popsány hypotézou dokonale.

Tento člen nabude nenulové hodnoty pouze v případě výjimek → protiváhou velikosti stromu je chyba učení...





Minimum Description Length (MDL)

Výběr hypotézy s minimální délkou kódu

Zajímavý fakt z TI: Optimální kód (s nejmenší očekávanou délkou) pro jev x s pravděpodobností $p(x)$ má $-\log_2(p(x))$ bitů.

zlogaritmujeme

$$\left\{ \begin{array}{lcl} h_{MAP} & = & \arg \max_{h \in H} P(D|h)P(h) \\ & = & \arg \max_{h \in H} \log_2 P(D|h) + \log_2 P(h) \\ & = & \arg \min_{h \in H} -\log_2 P(D|h) - \log_2 P(h) \end{array} \right.$$

Z poznatků o MDL můžeme členy interpretovat:

- $-\log_2(P(h))$ je délka hypotézy h v optimálním kódu
- $-\log_2(P(D|h))$ je délka D za podmínky h v optimálním kódu

→ vybereme hypotézu, která minimalizuje výraz
 $L(h) + L(<\text{chyba klasifikace}>)$



Využití hypotéz MAP a ML

Paradigma rozpoznávacího systému s HMM

Rozpoznávací systém – např. ASR, NLP, OCR, diagnostika

Forward problém (jeden krok při výpočtu maximálně věrohodnému odhadu):

Dáno: hypotéza h , pozorování (data) D

Odhadujeme: $P(D|h)$, tj. „p-st, že hypotéza h generuje data D “

Backward problém (rozpoznávání / jeden krok predikce):

Dáno: hypotéza h , pozorování D

Maximalizujeme: $P(h(X=x)|h, D)$, tj. hledáme nejlepší třídu x

Problém (učení) **Forward-Backward** –

Dáno: prostor hypotéz H , data D

Hledáme: $h \in H$ takovou, že $P(h|D)$ je maximální (MAP)



Bayesovský optimální klasifikátor (Bayes Optimal Classifier)

- v praxi většinou nehledáme nejpravděpodobnější hypotézu h za podmínky trénovacích dat D, ale nejpravděpodobnější klasifikaci \hat{v} vzorku x za podmínky pozorovaných dat → **klasifikační úloha**.
- $h_{MAP}(x)$ je klasifikace vypočítaná z nejpravděpodobnější hypotézy za podmínky dat (a apriorních znalostí), ale **nikoliv nejpravděpodobnější klasifikace!**
- nejpravděpodobnější klasifikace

$$\hat{v} = \arg \max_{v_j} P(v_j | D)$$

$$P(v_j | D) = \sum_{h_i \in H} P(v_j | h_i) P(h_i | D)$$





Bayesovský optimální klasifikátor

Výpočet nejpravděpodobnější klasifikace

$$\hat{v} = \arg \max_{v_j} P(v_j|D)$$

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

→ Bayesovský optimální klasifikátor má podobu

$$\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Žádný klasifikátor pracující se stejným prostorem hypotéz H a stejnými apriorními znalostmi nemůže v průměru ($\# \rightarrow \infty$) překonat Bayesovský optimální klasifikátor!



Bayesovský optimální klasifikátor

Příklad

Jaká je nejpravděpodobnější klasifikace neznámého vzorku x ?

- **Učení** – tři možné hypotézy (z trénovacích dat D):

$$P(h_1|D) = .4, P(h_2|D) = .3, P(h_3|D) = .3$$

- **Klasifikace** neznámého vzorku x :

$$h_{MAP} = h_1(x) = \boxed{+}, \quad h_2(x) = -, \quad h_3(x) = - \quad \leftarrow \text{MAP}$$

samotný MAP
by „vybral“ h_1

$$P(h_1|D) = .4, \quad P(-|h_1) = 0, \quad P(+|h_1) = 1$$

$$P(h_2|D) = .3, \quad P(-|h_2) = 1, \quad P(+|h_2) = 0$$

$$P(h_3|D) = .3, \quad P(-|h_3) = 1, \quad P(+|h_3) = 0$$

$$\text{BOC} \left\{ \begin{array}{l} \sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4 \\ \sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6 \end{array} \right. \text{ vybírá maximum}$$

$$h_{optimal} = \arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \boxed{-}$$



Naivní bayesovský klasifikátor

Charakteristika, vlastnosti

Jeden z nejstarších a současně nejpraktičtějších učících se algoritmů (spolu s rozhodovacími stromy, neuronovými sítěmi a k-NN):

- velmi jednoduchý pravděpodobnostní klasifikátor založený na aplikaci Bayesovy věty a silných (naivních) předpokladů o **podmíněné nezávislosti** příznaků klasifikovaných vzorů

Podmínky užití:

- středně až velmi velká trénovací množina
- **příznaky jsou nezávislé** za podmínky dané klasifikace

Úspěšné aplikace NBK:

- diagnostika
- klasifikace textových dokumentů (spam/ham, kategorie, ...)





Naivní bayesovský klasifikátor

Odvození

Předpokládejme rozhodovací pravidlo $f: X \rightarrow V$, přičemž každá instance $x \in X$ je popsána vektorem příznaků $[a_1, a_2, \dots, a_n]$. Nejpravděpodobnější hodnota $f(x)$ je dána

$$\begin{aligned} v_{MAP} &= \operatorname{argmax}_{v_j \in V} P(v_j | a_1, a_2 \dots a_n) \\ v_{MAP} &= \operatorname{argmax}_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)} \\ &= \operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j) \end{aligned}$$

← nelze vyjádřit Bayes

Naivní bayesovský předpoklad:

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$$

Předpokládáme-li, že jsou příznaky **nezávislé**, p-st P se redukuje na sdruženou p-st výskytu příznaků a_i za podmínky klasifikace v_j ...



Naivní bayesovský klasifikátor

Podmíněná nezávislost – teoretický základ

Definice **podmíněné nezávislosti** –

X je podmíněně nezávislé na Y za podmínky Z , jestliže hustota pravděpodobnosti X je nezávislá na hodnotě Y , známe-li hodnotu Z , tj.:

$$P(X|Y, Z) = P(X|Z)$$

Jinými slovy: X a Z jsou nezávislé na Y tehdy a jen tehdy, když znalost toho, že jev Y nastal, neposkytuje žádné informace o pravděpodobnosti toho, zda v závislosti na výskytu X nastane Z nebo naopak.

Příklad: Hrom H je podmíněně nezávislý na dešti D , za podmínky výskytu blesku B , tj. $P(H|D, B) = P(H|B)$



Naivní bayesovský klasifikátor

Matematický model

Nejpravděpodobnější klasifikace $f(x)$ vzoru x je dáná

$$v_{MAP} = \operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j)$$

Naivní
bayesovský
předpoklad

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$$



Výsledná podoba naivního bayesovského klasifikátoru:

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$



Naivní bayesovský klasifikátor Algoritmus

```
NB_Learn( trénovací data  $[a_1, \dots, a_n; v_j]_{j=1}^m$  ) {  
    for každou klasifikaci učitele  $v_j$   $j = 1 \dots m$  {  
         $\hat{P}(v_j) \leftarrow$  odhad  $P(v_j)$ ;  
        for každou hodnotu příznaku  $a_i$   $i = 1 \dots n$  {  
             $\hat{P}(a_i|v_j) \leftarrow$  odhad  $P(a_i|v_j)$ ;  
        }  
    }  
}
```

```
NB_Classify( neznámý vzorek  $x$  ) {  
    return  $v_{NB} = \operatorname{argmax}_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$   
}
```



Naivní bayesovský klasifikátor

Ukázka – Hrajeme dnes tenis?

Trénovací množina vypadá po 14-ti dnech pozorování takto:

Den	Počasí	Teplota	Vlhkost	Vítr	Tenis
1	jasno	vysoká	vysoká	mírný	ne
2	jasno	vysoká	vysoká	silný	ne
3	zataženo	vysoká	vysoká	mírný	ano
4	déšť	normální	vysoká	mírný	ano
5	déšť	nízká	normální	mírný	ano
6	déšť	nízká	normální	silný	ne
7	zataženo	nízká	normální	silný	ano
8	jasno	normální	vysoká	mírný	ne
9	jasno	nízká	normální	mírný	ano
10	déšť	normální	normální	mírný	ano
11	jasno	normální	normální	silný	ano
12	zataženo	normální	vysoká	silný	ano
13	zataženo	vysoká	normální	mírný	ano
14	déšť	normální	vysoká	silný	ne

učitel



Naivní bayesovský klasifikátor

Ukázka – Hrajeme dnes tenis?

Trénovací množinou natrénujeme NBK – všechny dílčí podmíněné p-sti z ní lze zjistit (resp. jejich odhad), také p-st odpovědi učitele, tj. hypotézy **ano** či **ne**...

Klasifikace neznámého vzorku x =

[počasí = jasno, teplota = nízká, vlhkost = vysoká, vítr = silný]

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$



$$P(\text{ano}) \times P(\text{počasí} = \text{jasno} | \text{ano}) \times P(\text{teplota} = \text{nízká} | \text{ano}) \times \\ \times P(\text{vlhkost} = \text{vysoká} | \text{ano}) \times P(\text{vítr} = \text{silný} | \text{ano}) = \textcolor{red}{0.005}$$



$$P(\text{ne}) \times P(\text{počasí} = \text{jasno} | \text{ne}) \times P(\text{teplota} = \text{nízká} | \text{ne}) \times \\ \times P(\text{vlhkost} = \text{vysoká} | \text{ne}) \times P(\text{vítr} = \text{silný} | \text{ne}) = \textcolor{green}{0.021}$$



Naivní bayesovský klasifikátor

Poznámky

- předpoklad podmíněné nezávislosti příznaků za podmínky klasifikace $P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$ je často nesprávný (a proto nesplněný)
- **ale NBK funguje překvapivě dobře i tak**
- není nutné mít správné odhady aposteriorních p-stí $\hat{P}(v_j | x)$ je ovšem nutné, aby platilo, že
$$\operatorname{argmax}_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i | v_j) = \operatorname{argmax}_{v_j \in V} P(v_j) P(a_1 \dots, a_n | v_j)$$
- matematické detaily a podrobný rozbor je v článku:
Domingos, P. & Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Proc. of the 13th Int'l Conference on Machine Learning*, pp. 105–112, Bari, Italy.
<http://homes.cs.washington.edu/~pedrod/papers/mlc96.pdf>
- aposteriorní p-sti NBK jsou v praxi obvykle nereálně blízko hodnot 0 nebo 1



Naivní bayesovský klasifikátor

Poznámky

Celkem běžný problém: Co když žádný z trénovacích vzorků klasifikovaných učitelem jako v_j nemá příznak s hodnotou a_i ?

Potom $\hat{P}(a_i|v_j) = 0$, a tedy $\hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = 0$

Typickým řešením je výpočet tzv. **m -odhadu** $\hat{P}(a_i|v_j)$

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}$$

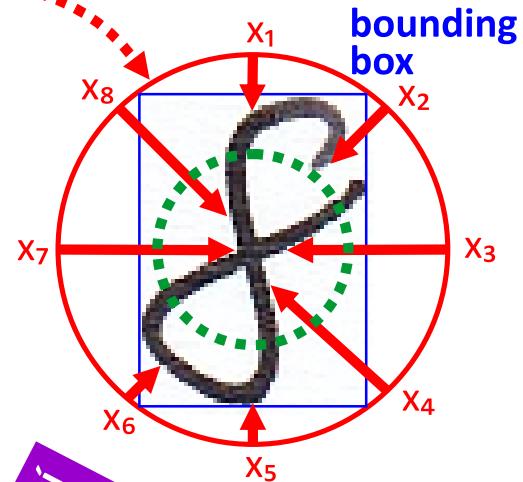
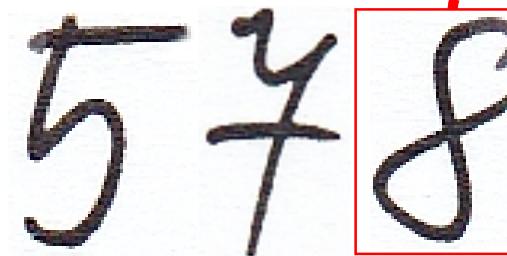
- n je počet trénovacích vzorků zařazených do j -té třídy, $v = v_j$,
- n_c je počet vzorků, kde $v = v_i$ a současně $a = a_i$,
- p je apriorní odhad $\hat{P}(a_i|v_j)$
- m je váha apriorního odhadu, tj. vlastně počet „virtuálních“ trénovacích vzorků





Ukázka použití NBK OCR (Optical Character Recognition) PSČ

Naskenované číslice



Apriorní pravděpodobnosti výskytu číslic v PSČ:

$$P("0") = 0.07$$

$$P("1") = 0.16 \text{ (Praha!)} \quad \left. \right\}$$

...

$$P("9") = 0.03$$

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

$$[0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1] = ?$$

$$P("0") \times P(x_1 = 0 | "0") \times \\ \times P(x_2 = 0 | "0") \times \dots$$

$$P("1") \times \dots \times P(x_8 = 1 | "1")$$

$$P("2") \times \dots \times P(x_8 = 1 | "2")$$

...

$$P("8") \times \dots \times P(x_8 = 1 | "8")$$

$$P("9") \times \dots \times P(x_8 = 1 | "9")$$



Klasifikace textu

Úspěšná aplikace NBK

Proč klasifikovat text?

- klasifikace příchozí pošty na spam/ham
- třídění článků na webu podle tématu (zprávy, sport, kultura, politika, věda, ...)
- vyhledávání textu podle zájmů uživatele (recommender)
- prezentace reklamy podle zájmů cílové skupiny

Naivní bayesovský klasifikátor je v této problémové oblasti jedním z nejúspěšnějších učících se algoritmů...

Problém: Jak reprezentovat text jako příznaky, resp. jaké příznaky vybrat pro reprezentaci textového dokumentu?



Klasifikace textu pomocí NBK

Filtrování spamu

Klasifikace **Spam?** : Dokument $\rightarrow \{+, -\}$

- každý textový dokument reprezentujme jako vektor slov **doc**, tj. jeden příznak na pozici slova v dokumentu
- **učení** – z trénovacích vzorků určeme odhady $P(+)$, $P(-)$, $P(doc|+)$, $P(doc|-)$

Předpoklad podmíněné nezávislosti NBK

$$P(doc|v_j) = \prod_{i=1}^{\text{délka}(doc)} P(a_i = w_k | v_j)$$

$P(a_i = w_k | v_j)$ je p-st, že slovo na pozici i je w_k , za podmínky klasifikace v_j

Další předpoklad: $P(a_i = w_k | v_j) = P(a_m = w_k | v_j), \forall i, m$
tzv. „bag-of-words“ model



Klasifikace textu pomocí NBK

Algoritmus učení

NB_Learn_Text(trénovací množina, V)

- vytvořit slovník

slovník \leftarrow všechny jedinečné výrazy z *trénovací množiny*

- výpočet pravděpodobnosti $P(v_j)$ a $P(w_k | v_j)$

for \forall klasifikace $v_j \in V$:

$doc_j \leftarrow$ podmnožina dokumentů z *trénovací množiny*
klasifikovaných učitelem do v_j

$P(v_j) \leftarrow |doc_j| / |\text{trénovací množina}|$

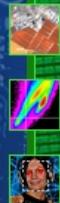
$text_j \leftarrow$ sloučení všech doc_j do jediného dokumentu

$n \leftarrow$ počet slov v $text_j$ (k -duplicity se započítávají $k \times$)

for \forall slova ve slovníku:

$n_k \leftarrow$ počet výskytů slova w_k v $text_j$

$P(w_k | v_j) \leftarrow (n_k + 1) / (n + |\text{slovník}|)$



>

< 757586001

C228



Klasifikace textu pomocí NBK

Klasifikace neznámého vzorku

NB_Classify_Text(doc)

pozice \leftarrow všechny pozice slov v **doc**, na kterých se vyskytují výrazy obsažené ve **slovníku**

return v_{NB} , vypočítaný výrazem

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_{i \in \text{pozice}} P(a_i | v_j)$$



Úspěšnost klasifikace textu pomocí NBK:

Na úloze Twenty NewsGroups (1000 trénovacích dokumentů, zařazení do 20 kategorií) – **89%**