FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BAYESOVSKÉ UČENÍ

Autor textu:
Ing. Petr Honzík, Ph.D.

Květen 2014

Komplexní inovace studijních programů a zvyšování kvality výuky na FEKT VUT v Brně
OP VK CZ.1.07/2.2.00/28.0193

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Obsah přednášky

1. **Bayesův teorém**
2. Brutální Bayesovský klasifikátor (BBK)
3. Maximální aposteriorní pravděpodobnost (MAP)
4. Optimální Bayesovský klasifikátor (OBK)
5. Gibbsův algoritmus (GiA)
6. **Naivní Bayesovský klasifikátor (NBK)**

Thomas Bayes

- duchovní, 2. polovina 18. století
- zavrhován statistiky (nepodložené empirické metody)
- zabýval se otázkou, jak pozdější zkušenosti uvést do souladu s původními předpoklady (dynamické „ověřování hypotéz“ s jejich dodatečnou korekcí)
- **(a)priorní pravděpodobnost** (dána na počátku) vs. **(a)posteriorní pravděpodobnost** (vyplývající z následné analýzy, ověřené zkušeností)
- Principiálně blízké přirozenému lidskému uvažování

? uveď rozdíl mezi apriorní a aposteriorní pravděpodobností

Bayesův vzorec

$$p(H_k \mid D) = \frac{p(D \mid H_k) \cdot p(H_k)}{\sum_{i=1}^{K} p(D \mid H_i) \cdot p(H_i)} = p(H_k) \frac{p(D \mid H_k)}{p(D)}$$

H_k je **hypotéza**

D je realita, data, konkrétní měření

p(H_k) zkušenost; **apriorní pravděpodobnost**

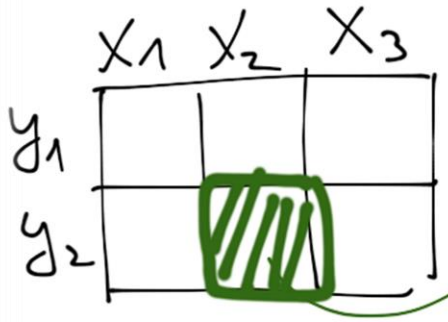
p($D|H_k$) je **věrohodnost**

p(D) pravděpodobnost nastolení měření (dat, údajů) D, **evidence**

p($H_k|D$) je **aposteriorní pravděpodobnost** platnosti konkrétní hypotézy

? uveď a popiš Bayesův vzorec

Bayesův vzorec ze sdružené pravděpodobnosti



$$p(x_2, y_2) = p(y_2 | x_2)$$

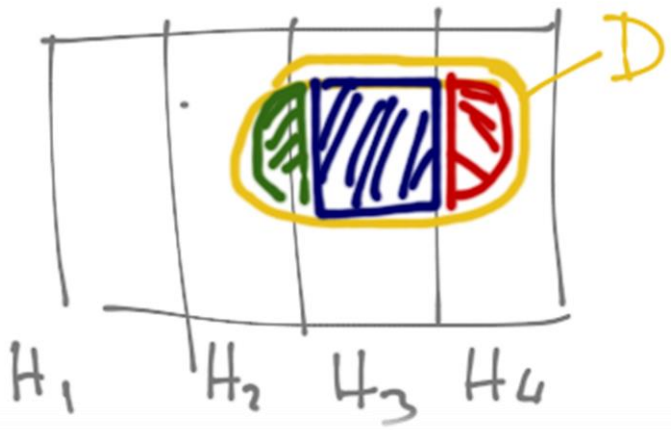
$$p(x_2, y_2) = p(x_2 | y_2) \cdot p(y_2)$$

$$p(y_2, x_2) = p(y_2 | x_2) \cdot p(x_2)$$

$$\Rightarrow \boxed{p(x_2 | y_2) = \frac{p(y_2 | x_2) \cdot p(x_2)}{p(y_2)}}$$

? uveď a popiš Bayesův vzorec

Bayesův vzorec graficky – 1/3

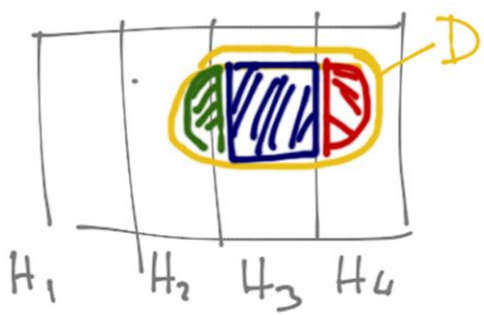


$$\sum P(H_i) = 1$$

$$\forall_{i \neq j} : H_i \cap H_j = \{\emptyset\}$$

? uveď a popiš Bayesův vzorec

Bayesův vzorec graficky – 2/3



$$\sum_i p(H_i) = 1$$

$$\forall_{i \neq j} : H_i \cap H_j := \{\emptyset\}$$

$$= p(D)$$

$$= p(D|H_2) \cdot p(H_2)$$

$$\Rightarrow \boxed{p(H_2|D) = \frac{\text{}}{\text{  }} = \frac{p(D|H_2)\,p(H_2)}{p(D)}}$$

? uveď a popiš Bayesův vzorec

Bayesův vzorec graficky – 3/3

$$p(D) = p(D|H_2)\cdot p(H_2) + p(D|H_3)\,p(H_3) + p(D|H_4)\cdot p(H_4)$$

$$\Rightarrow \boxed{p(H_2|D) = \frac{p(D|H_2)\cdot p(H_2)}{\sum p(D|H_i)\,p(H_i)}}$$

? uveď a popiš Bayesův vzorec

Bayesovské učení - poznámky

- Klademe si dvě základní otázky:
  - Jaká **hypotéza** (model) o modelovaném systému je s největší pravděpodobností platná?
  - Jaká je **predikce** nové instance na základě známých hypotéz (modelů)?
- Výhody
  - lze kombinovat **předchozí znalost** s **pozorovanými měřeními** (apriorní pravděpodobnost, evidence)
  - intuitivní řešení **blízké lidskému uvažování**
- Nevýhody
  - narůstající **složitost** hledání řešení s rostoucím počtem hypotéz – v obecném případě

Příklad 1 – Bayes

- V pytlíku je 7 hracích kostek. 4 jsou normální, na 1 padají víc šestky, na 2 víc jedničky.
- Jaká je pravděpodobnost, že vytáhnu normální kostku? (apriorní pravděpodobnost)
- **Jaká je pravděpodobnost, že** když po 20 hodech (experiment E) padla šestka 6-krát, **mám kostku *Normální*, *Šestkovou*, *Jedničkovou*?** (aposteriorní pravděpodobnost)
- (když $p(6|N)=1/6$, $p(6|Š)=1/3$, $p(6|J)=2/15$)

$p(N|E)=0{,}52$; $p(Š|E)=0{,}36$; $p(J|E)=0{,}12$

Ale co teď s tím?

- Víme, že s
  - p=0,52 držíme v ruce kostku normální
  - p=0,36 kostku šestkovou
  - p=0,12 kostku jedničkovou.

- Jak z toho ale predikovat?
- Jak odhadnout pravděpodobnost, s jakou padne číslo „6“ v dalším hodu?

Brutální Bayesovský klasifikátor

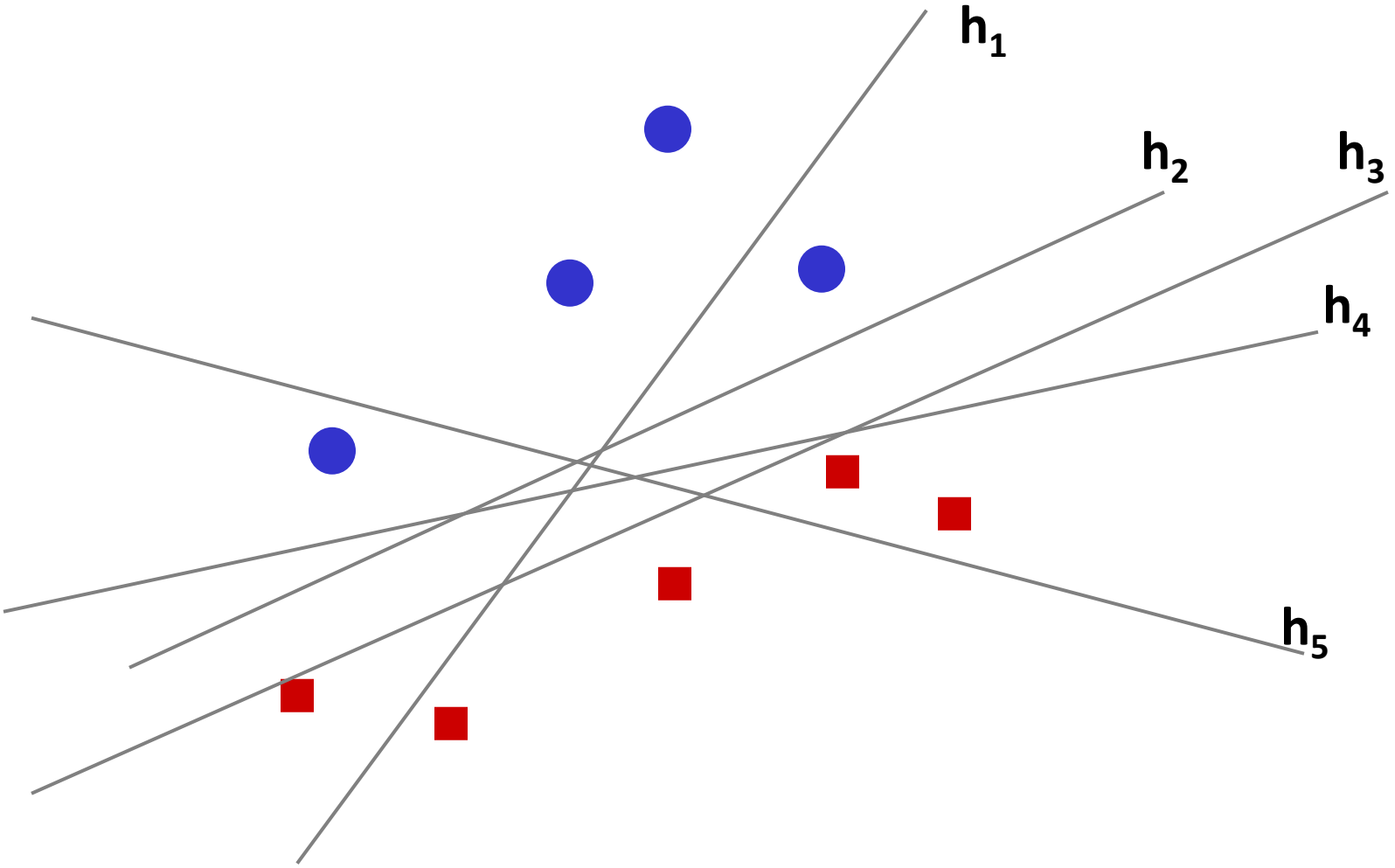
- Princip
  - **Všechny hypotézy konzistentní s daty mají stejnou pravděpodobnost** P(h_j), hypotézy nekonzistentní jsou vyloučeny
  - Pokud jsou trénovací data zašuměná, budou správné hypotézy zamítnuty, možné zamítnutí všech hypotéz
- Predikce
  - pokud *K* modelů predikuje na trénovacích datech správně, jsou si podle BBK tyto hypotézy rovnocené; predikce je průměr nebo nejčastější třída

$$P(D|h_j) = \begin{cases} 1; \forall y_i : y_i = h_j(x_i) \\ 0; \exists y_i : y_i \neq h_j(x_i) \end{cases}$$

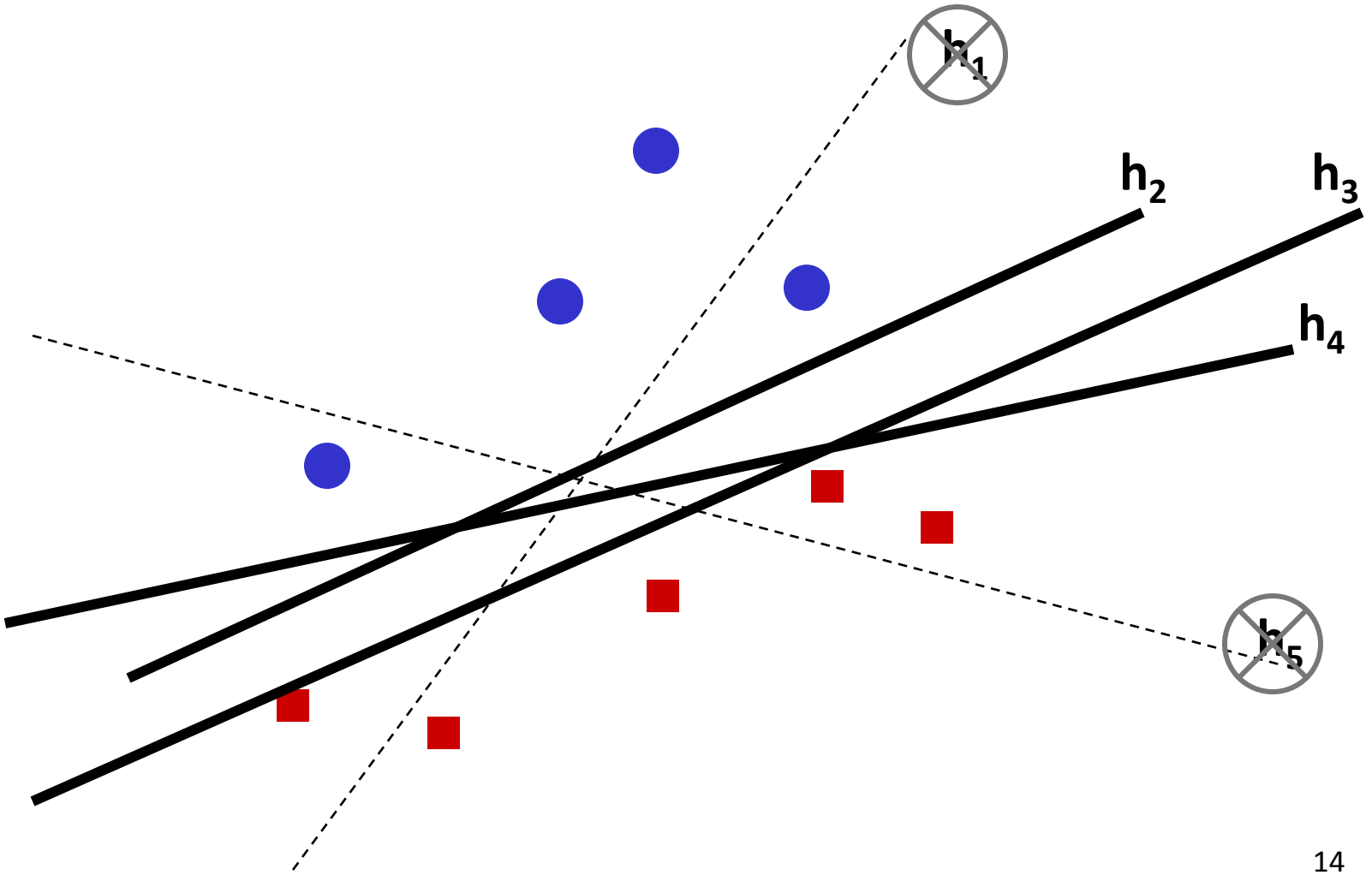
$$P(h_j) = \frac{1}{\sum_{j=1}^{hypotez} P(D|h_j)} = \frac{1}{|H_{konz}|} = konst$$

$$\Rightarrow g_{out} = \arg\max_{j=1\ldots C} \sum_{h_i \in H} P(g_j|h_i)$$

BBK – příklad



BBK – příklad



Maximální aposteriorní pravděpodobnost

- **Nejpravděpodobněji platná hypotéza** je ta s největší aposteriorní pravděpodobností.
- Podoba s maximální věrohodností a MNČ

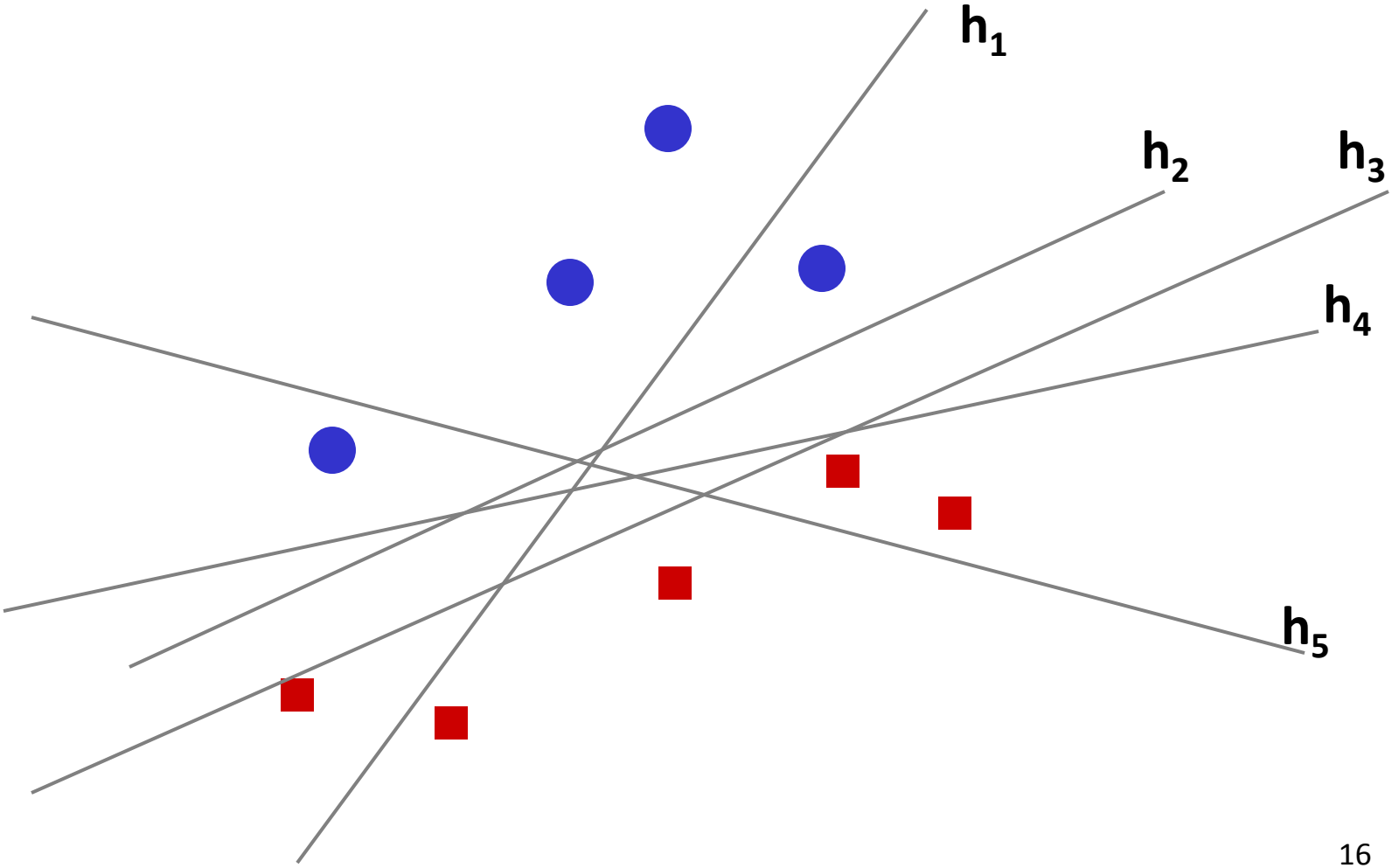
$$h_{MAP} = \arg\max_{j=1\ldots|H|} P(h_j|D) =$$

$$= \arg\max_{j=1\ldots|H|} P(h_j)\frac{P(D|h_j)}{P(D)} =$$

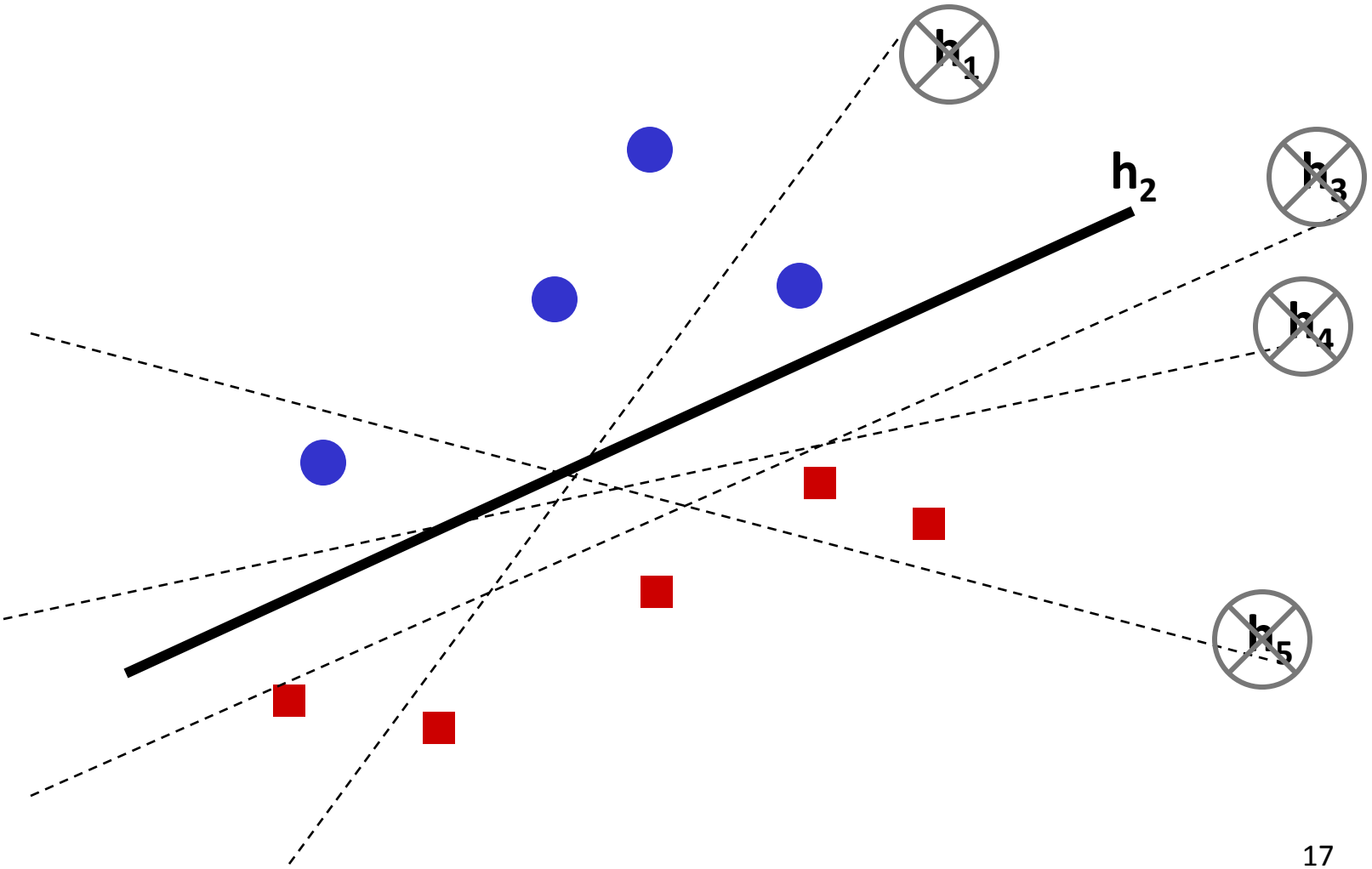
$$= \arg\max_{j=1\ldots|H|} P(h_j)\cdot P(D|h_j)$$

$$g_{out} = \arg\max_{j=1\ldots C} P(g_j|h_{MAP})$$

MAP – příklad



MAP – příklad

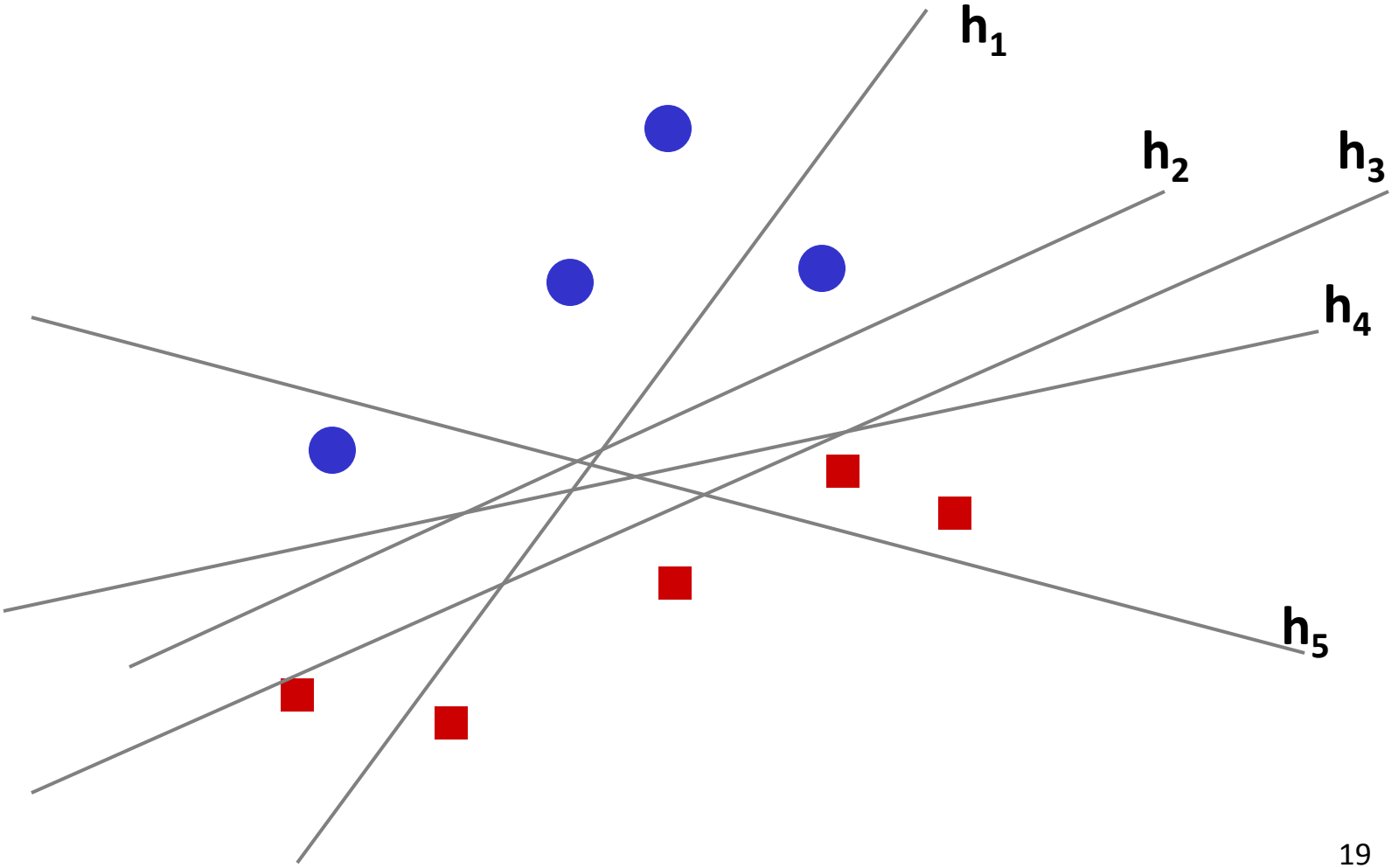


Optimální Bayesovský klasifikátor

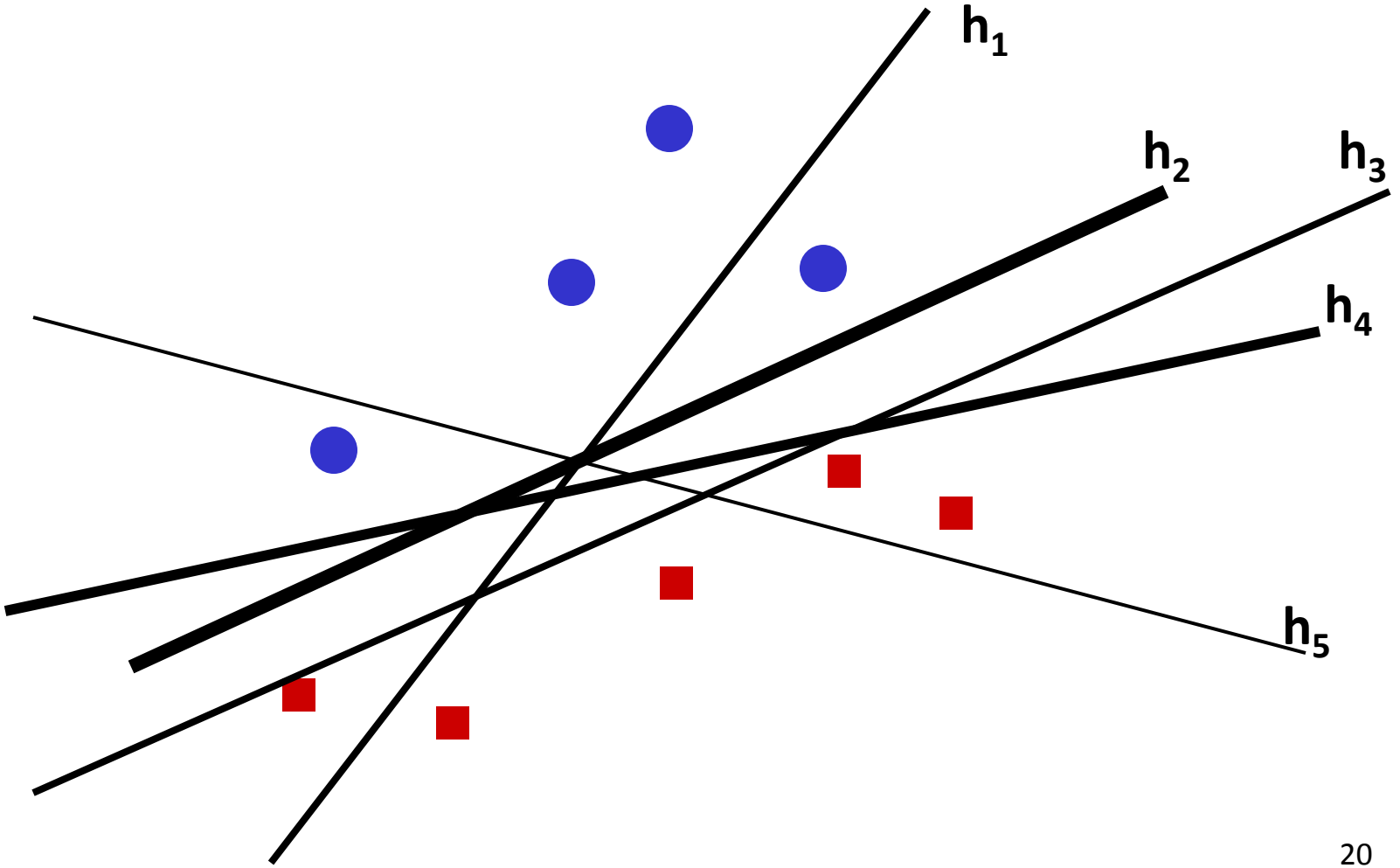
- „Jaká je **nejpravděpodobnější klasifikace g_j** nové instance za předpokladu informací založených na několika hypotézách h_i?“
- jinými slovy, optimální bayesovský klasifikátor vychází z hypotéz a jejich věrohodností; výpočty sejné jako u MAP, nicméně není vybrán 1 nejlepší, ale **všechny hypotézy se úměrně své kvalitě podílejí na predikci**
- Klasifikujme binárně do tříd +,– a mějme hypotézy s aposteriorní pravděpodobností $p(h_1|D)=0,4$ $p(h_2|D)=0,3$ a $p(h_3|D)=0,3$. Pokud podle h_1 je výsledek negativní a podle h_2 a h_3 je výsledek pozitivní, MAP považuje za výsledek predikci dle h_1, OBK však dle h_2 a h_3, protože $(0,3+0,3) > 0,4$.

$$g_{\max} = \arg\max_{j=1\ldots C} \sum_{h_i \in H} P\left(g_j \middle| h_i\right) \cdot P\left(h_i \middle| D\right)$$

OBK – příklad



OBK – příklad



Gibbsův algoritmus

- řeší hlavní problém OBK – tím je **výpočetní náročnost** – použití všech váhovaných modelů
- Princip
  - při každé predikci vyber náhodně (avšak úměrně posteriorní pravděpodobnosti modelů) jednu hypotézu a podle té predikuj
  - Dosažená chyba je maximálně dvojnásobná oproti OBK

Příklad 2 – BBK, MAP, OBK

- V pytlíku je 7 hracích kostek. 4 jsou normální, na 1 padají víc „6“, na 2 víc „1“.
- Jaká je pravděpodobnost, že když po **experimetnu *E*** ve 20 hodech padla šestka 6-krát, mám kostku *Normální*, *Šestkovou*, *Jedničkovou*? (aposteriorní pravděpodobnost)
- (když $p(6|N)=1/6$, $p(6|Š)=1/3$, $p(6|J)=2/15$); $p(N|E)=0,52$ $p(Š|E)=0,36$ $p(J|E)=0,12$
- **S jakou pravděpodobností padne šestka ve 21. hodu?**

- **BBK:**
  - Konzistentní (možné) jsou všechny hypotézy, **p**=(1/6+1/3+2/15)/3**=0,21**
- **MAP:**
  - Nejpravděpodobnější byla hypotéza „normální“, tedy 1/6, **p=0,17**
- **OBK:**
  - Výsledná pravděpodobnost je dána součtem součinů pravděpodobnosti hypotézy a nastolení dané události, tedy
    p = 0,52·1/6 + 0,36·1/3 + 0,12·2/15 = **0,22**

Příklad 3

V pytlíku mohou být jehlany *j* a kostičky *k*. Vytáhli jsme 1 kostičku – **experiment *E***. S jakou pravděpodobností p_k vytáhneme další kostičku?

- Víme, že v pytlíku bylo na počátku právě 5 objektů
- Víme, že na počátku je jakákoliv kombinace jehlanů a kostiček v pytlíku stejně pravděpodobná

h_i = hypotéza, že v pytlíku bylo *i*-kostiček;
urči všechny $p(h_i|E)$ a pro jednotlivé metody $\boldsymbol{p_k}$ v dalším tahu

- **BBK:** Až na h_0 jsou konzistentní; $p(k|h_1)=0$; $p(k|h_2)=0{,}25$ …
  $\boldsymbol{p_k}=\Sigma\, p(k|h_i)/5=(0+0{,}25+0{,}5+0{,}75+1)/5=\mathbf{0{,}5}$
- **MAP:** Nejpravděpodobnější je h_5, tedy $\boldsymbol{p_k=1}$
- **OBK:** $p(E|h_0)=0$; $p(E|h_1)=0{,}2$; … $p(E|h_5)=1$;
  protože $p(h_i)$=konst, *platí* $p(h_i|E)=\frac{p(E|h_i)\cdot p(h_i)}{\sum p(E|h_i)\cdot p(h_i)}=\frac{p(E|h_i)}{\sum p(E|h_i)}$
  $\mathbf{p_k}=\Sigma[\, p(k|h_i)\, p(E|h_i) / \Sigma\, p(E|h_i)\,]=\mathbf{0{,}67}$

Naivní Bayesovský klasifikátor - NBK

- navzdory předpokladu nezávislosti veličin (není většinou pravda) velice přesný; kvůli zjednodušení slovo *naivní*
- vychází z Bayesovy podmíněné pravděpodobnosti
- **klasifikace** na základě nejpravděpodobnější klasifikace g_{MAP} **za předpokladu vstupního vektoru** $(x_1,\ldots,x_n)$

$$g_{MAP} = \arg\max_{j=1\ldots C} P\left(g_j \middle| x_1,\ldots x_n\right)$$

C je počet tříd

n je počet atributů

Proč se NBK jmenuje naivní? V čem naivita spočívá?

Naivní Bayesovský klasifikátor - NBK

- což lze na základě bayesovy věty upravit následujícím způsobem:

$$g_{MAP} = \arg\max_{j=1\ldots C} P(g_j)\frac{P(x_1,\ldots,x_n|g_j)}{P(x_1,\ldots,x_n)} =$$

$$= \arg\max_{j=1\ldots C} P(g_j)\cdot P(x_1,\ldots,x_n|g_j)$$

- $P(g_j)$ se určí na základě četnosti výskytu v trénovacích datech
- $P(x_1,\ldots,x_n/g_j)$ pro větší n (desítky) prakticky nemožné zjistit (tolik dat nemáme…). Určí se proto na základě zjednodušeného předpokladu, že hodnoty vstupních veličin jsou na sobě podmíněně nezávislé (což není většinou pravda)

Naivní Bayesovský klasifikátor - NBK

- pro výpočet sdružené podmíněné pravděpodobnosti platí:

$$P(x_1,\ldots,x_n|g_j) = P(x_1|g_j)\cdot P(x_2,\ldots,x_n|g_j,x_1) = \ldots =$$
$$= P(x_1|g_j)\cdot P(x_2|g_j,x_1)\cdot P(x_3|g_j,x_1,x_2)\ldots$$

- to lze za předpokladu nezávislosti **zjednodušit** takto:

$$P(x_1,\ldots,x_n|g_j) = \prod_{i=1\ldots n} P(x_i|g_j)$$

- konečný vztah pro výslednou klasifikaci je dán vztahem

$$g_{MAP} = \arg\max_{j=1\ldots C} P(g_j)\cdot \prod_{i=1\ldots n} P(x_i|g_j)$$

? uveď rozdíl mezi optimálním a naivním Bayesovským klasifikátorem

NBK – okrajové podmínky

- **Problém:** situace, kdy konkrétní hodnota atributu pro danou třídu nikdy nenastala (žádná bruneta se nespálila); celý součin je pak kvůli jednomu členu roven 0, což není vhodné;

$$P(x_2|g_1)=0 \Rightarrow P(x_1,\ldots,x_n|g_1)=\prod_{i=1\ldots n} P(x_i|g_1)=0$$

- **Řešení**: výpočet podmíněné pravděpodobnosti dán vztahem:

$$P(x_1|g_1)=\frac{n_c+mp}{n+m} \qquad texty: \frac{n_c+1}{n+|slovník|}$$

kde n je počet prvků třídy g_1 (počet spálených), n_c počet prvků v g_1 s danou hodnotou atributu x_1 (spálených brunet), m váha (např. 1) a $p=1/k$, kde k je výčet hodnot atributu x_1 (3:bruneta,blond,zrzavá)

? uveď rozdíl mezi optimálním a naivním Bayesovským klasifikátorem

NBK – příklad

den	předpověď	teplota	vlhkost	vítr	hrát tenis?
1.	slunečno	teplo	vysoká	slabý	**NE**
2.	slunečno	teplo	vysoká	silný	**NE**
3.	zataženo	teplo	vysoká	slabý	**ANO**
4.	déšť	středně	vysoká	slabý	**ANO**
5.	déšť	chladno	normální	slabý	**ANO**
6.	déšť	chladno	normální	silný	**NE**
7.	zataženo	chladno	normální	silný	**ANO**
8.	slunečno	středně	vysoká	slabý	**NE**
9.	slunečno	chladno	normální	slabý	**ANO**
10.	déšť	středně	normální	slabý	**ANO**
11.	slunečno	středně	normální	silný	**ANO**
12.	zataženo	středně	vysoká	silný	**ANO**
13.	zataženo	teplo	normální	slabý	**ANO**
14.	déšť	středně	vysoká	silný	**NE**

$$g_{MAP} = \arg\max_{j=1\ldots C} P(g_j) \cdot \prod_{i=1\ldots n} P(x_i | g_j)$$

Bude se hrát 15. den? **<slunečno,chladno,vysoká,silný>**

NBK – příklad

P(ANO)=9/14 P(NE)=5/14

P(slunečno/ANO)=2/9 P(slunečno/NE)=3/5

P(chladno/ANO)=3/9 P(chladno/NE)=1/5

P(vysoká/ANO)=3/9 P(vysoká/NE)=4/5

P(silný/ANO)=3/9 P(silný/NE)=3/5

P(ANO) . P(slunečno/ANO) P(chladno/ANO) . P(vysoká/ANO) . P(silný/ANO) = 0,0053

P(NE) . P(slunečno/NE) P(chladno/NE) . P(vysoká/NE) . P(silný/NE) = 0,0206

P(ANO|slunčeno,chladno,vysoká,silný) = 0,0053 / (0,0053 + 0,0206) = 0,20

P(NE|slunčeno,chladno,vysoká,silný) = 0,0206 / (0,0053 + 0,0206) = 0,80

Princip jednoduchého spamového filtru

- **Apriorní P**:
- P(spam) = 0,9
- P(ham) = 0,1

z databáze slov	spam (1000)	ham(300)
…	…	…
mamka	***1***	***12***
ahoj	***60***	***23***
oběd	***2***	***3***

- P(spam|mamka,ahoj,oběd – bez normalizace) = P(spam)*P(mamka|spam)*P(ahoj|spam)…

- P(spam|… – bez normalizace) = 0,9*1/1000*60/1000*2/1000= 1,08E-7
- P(ham|… – bez normalizace) = 0,1*12/300*38/300*3/300= 3,07E-6

- **P(spam|mamka,ahoj,oběd)**=

  P(spam|mamka,ahoj,oběd)/[P(spam|mamka,ahoj,oběd)+ P(ham|mamka,ahoj,oběd)]= 1,08E-7 / (1,08E-7+3,07E-6) = **0,034**
- **P(ham|mamka,ahoj,oběd)=0,966**

Shrnutí Bayese

- MAP (maximální aposteriorní pravděpodobnost) **vybírá nejpravděpodobnější hypotézu**. Podle této **jediné hypotézy** je následně predikováno.
- Brutální Bayesovské učení konceptů **vybírá** z množiny hypotéz (modelů) **hypotézy konzistentní** s trénovacími daty, které jsou si **rovnocenné**. Při predikci jsou si i **výstupy hypotéz rovnocenné**.
- Optimální Bayesovský klasifikátor stanoví **posteriorní pravděpodobnost všech hypotéz**. Při predikci jsou výstupy hypotéz **váhovány** posteriorními pravděpodobnostmi.

- Naivní Bayesovský klasifikátor **vytváří klasifikátor**, predikuje na základě nové instance a trénovacích dat.

? jaká je charakteristika : BBK, MAP, OBK, NBK