Extrakce a selekce příznaků

Based on slides **Martina Bachlera** martin.bachler@igi.tugraz.at **, Makoto Miwa**

And paper Isabelle Guyon, André Elisseeff: An Introduction to variable and feature selection. *JMLR*, 3 (2003) 1157-1182

Osnova - plán přednášky

- Úvod/Motivace/PAC učení — Proč ?
- Základní definice, Terminologie — Co ?
- Metody hodnocení příznaků (Variable Ranking methods — Jak?
- Výběr podmnožiny příznaků

Které příznaky mají význam pro DM ?

- V případě prediktivních úloh musí jít především o příznaky, jejichž hodnota je známá v okamžiku, kdy chceme predikci provádět.

- Pozor na **anachronické příznaky** (anachronistic at.), tj. takové, že nesplňují výše uvedený požadavek.

- **Příklad.** Telefon.operátor a predikce těch, co přecházejí k jinému operátorovi. Mezi 500 použitými atributy se ukázal mít velkou prediktivní sílu atribut odpovídající jménu zaměstnance, který dělal s klientem poslední interview. Později se ukázalo, že jiný člověk měl na starosti klienty, kteří projevili zájem odejít.

Problém **zaostřování pozornosti**

- **Běžný a velmi častý problém inteligentních (učících se) agentů (subjektů).**
- Hledá se odpověď na otázku: Které aspekty řešeného problému jsou důležité/nezbytné pro vyřešení?
- **Je nutné rozlišit mezi relevantními a zbytnými částmi dostupné zkušenosti.**

Co je to **výběr (selekce) příznaků ?**

- **Cílem výběru příznaků**

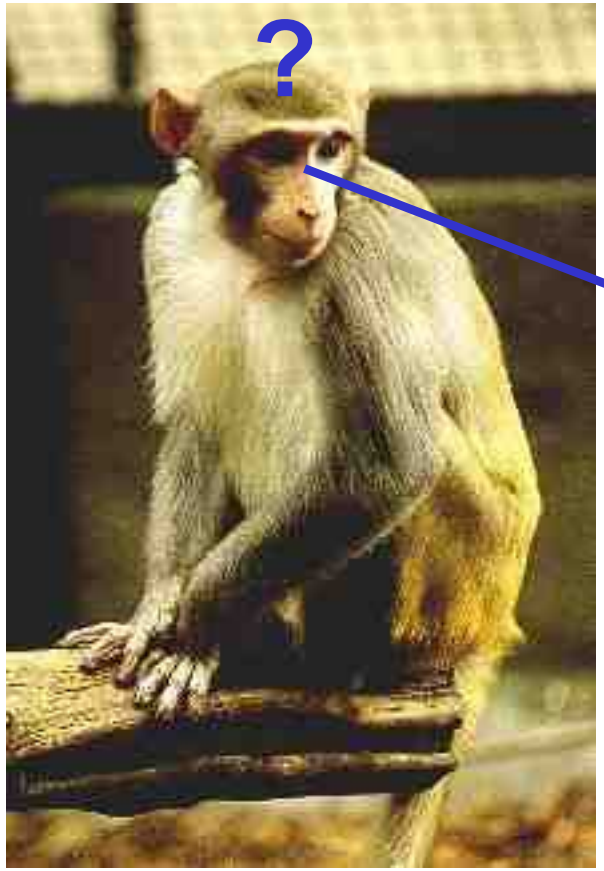
je nalézt takovou podmnožinu příznaků, která je dostatečná pro očekávané rozhodnutí – učící se algoritmus může ostatní příznaky ignorovat (výsledkem je REDUCE DIMENSIONALITY)

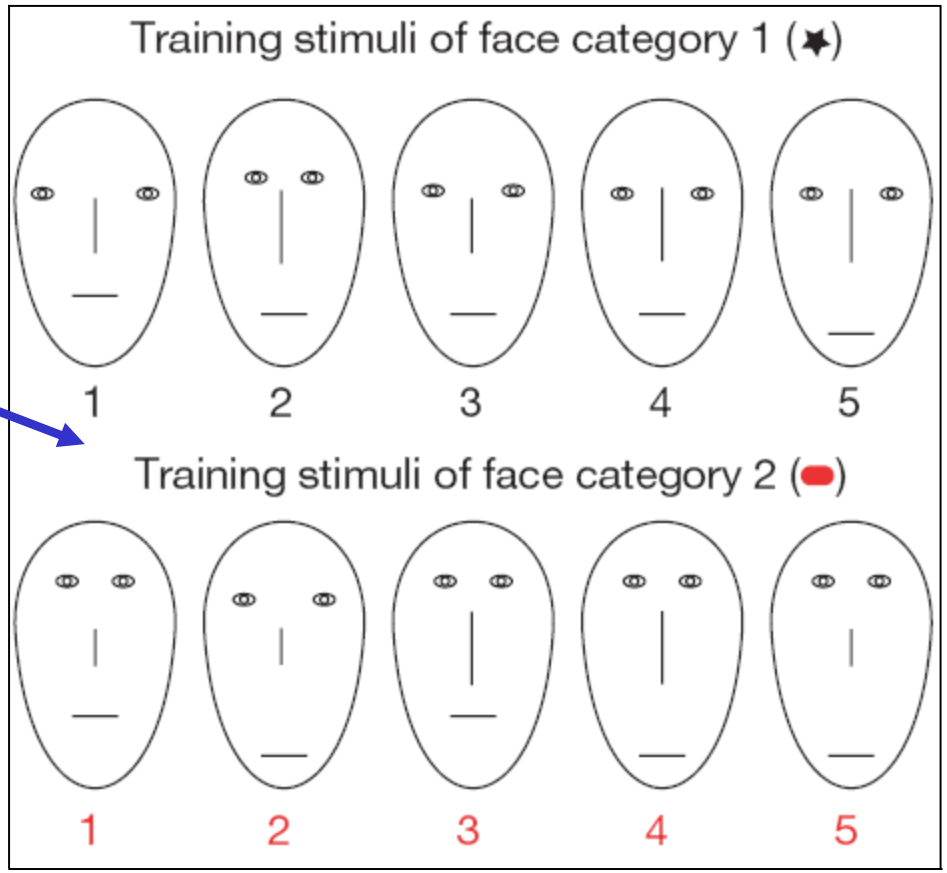
- **Živé organismy tento problém řeší stále a průběžně**!

Motivační příklad z biologie

[1]

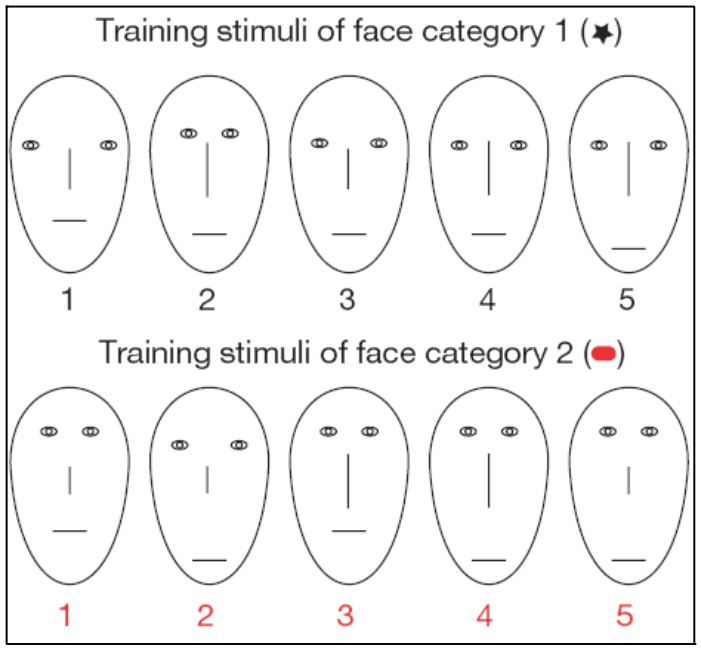
Opice, které se mají naučit rozlišovat mezi 2 třídami obličejů označenými jako ★ a ⬬

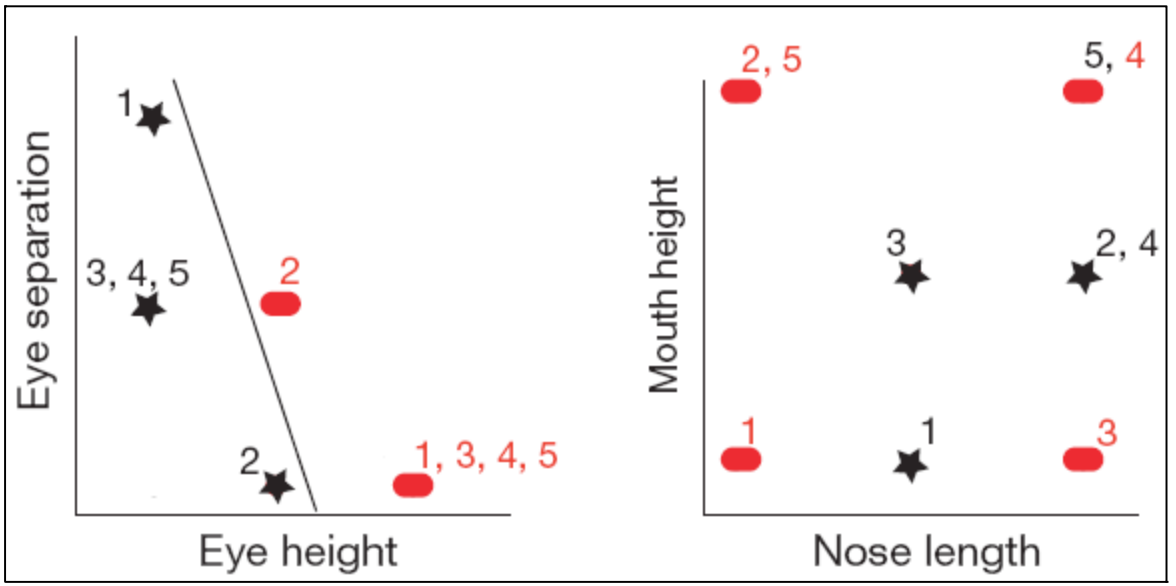




[1] Nathasha Sigala, Nikos Logothetis: Visual categorization shapes feature selectivity in the primate visual cortex. *Nature* Vol. 415(2002)

Motivační příklad z biologie





Všechny uvažované příznaky:

- Oči: výška umístění
- Oči: vzdálenost mezi nimi
- Nos: délka
- Ústa: výška umístění

Kolik lze vytvořit párů?

Diagnostické příznaky:

- Oči: výška umístění
- Oči: vzdálenost mezi nimi

Ne-diagnostické příznaky:

- Nos: délka
- Ústa: výška umístění

Motivační příklad z biologie

V průběhu sledovaného procesu učení u opic byla sledována aktivita 150 neuronů v příslušné části mozkové kůry (anterior inferior temporal cortex)

Výsledky:

- Na začátku pokusu bylo identifikováno 44 neuronů, jejichž chování se měnilo v souvislosti se výrazně měnilo v souvislosti se změnou alespoň jednoho pozorovaného příznaku

- Poté, co opice úlohy zvládly (naučily se rozpoznávat určené 2 třídy): 72% (32/44) neuronů reagovalo pouze v případě změny jednoho nebo obou diagnostických příznaků (a nikoliv v případě změny ne-diagnostických příznaků)

Proč je výběr příznaků důležitý?

Vztah mezi výběrem příznaků a strojovým učení (ML) nebo dobýváním znalostí?

- Předpokládáme-li, že informace o cílové třídě je **implicitně zahrnuta v hodnotách příznaků**, pak

- Můžeme učinit naivní závěr, že mít více příznaků
- je výhodné, neboť tím získáváme
  => víc informací
  => větší rozlišovací schopnost.

- Praktická zkušenost upozorňuje, že **často tomu tak není!**

- Další doplňkový argument:
  Optimalizace je (obvykle) výhodná. Proč se tedy nepokusit o optimalizaci kódování vstupu ?

Věta o PAC učení rozhodovacího stromu

Nechť objekty jsou charakterizovány pomocí ***n*** binárních atributů a nechť připouštíme jen hypotézy ve tvaru rozhodovacího stromu s maximální délkou větve ***k***. Dále nechť δ, ε jsou malá pevně zvolená kladná čísla blízká 0. Pokud algoritmus strojového učení vygeneruje hypotézu φ, která je konzistentní se všemi **m** příklady trénovací množiny a platí

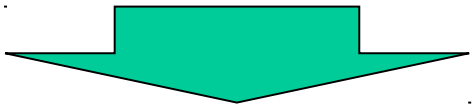
$$m \geq m_{k\text{-DT}}(n) \geq c\,(\,n^k + \ln(1/\delta))\,/\,\varepsilon$$

pak φ je ε-skoro správná hypotéza s pravděpodobností větší než $(1-\delta)$, t.j. chyba hypotézy φ na celém definičním oboru konceptu je menší než ε s pravděpodobností větší než $(1-\delta)$.

Podobný vztah jako ten popsaný větou o PAC učení rozhodovacímu stromu platí o všech ostatních metodách strojového učení!

- Příklady objemných dat o vysoké dimenzi
  - Textové dokumenty, etc…
  - Situace, kdy je třeba doplnit další informace jako apriorní znalosti (třeba 3D struktura molekuly DNA)

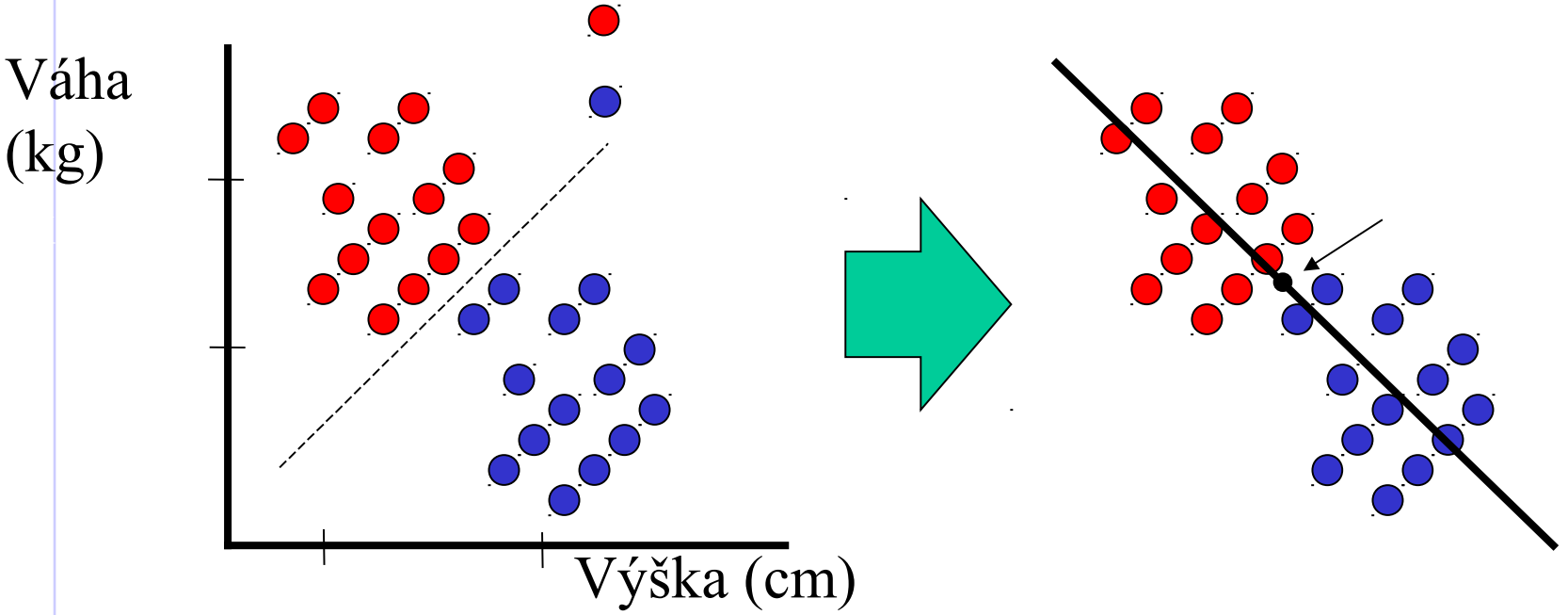
  v úlohách typu
  - Získávání informací
  - Klasifikace
  - etc…



Redukce dimenze je velmi důležitá

Příklad redukce dimenze

V případě klasifikace na lidi „štíhlé" a „s nadváhou" podle 2 atributů



Redukce dimenze pomocí přechodu na nový příznak

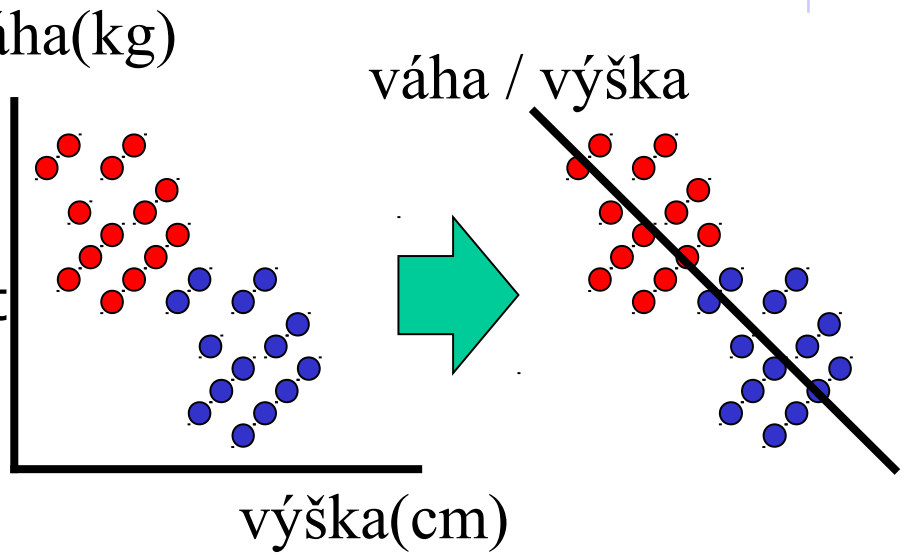
váha/(výška – 100)

- zachová informaci o klasifikaci na „štíhlé" a „s nadváhou"
- zjednodušuje klasifikaci
- Redukuje velikost dat (2 příznaky → 1 příznak)

2 cesty k redukci dimenze

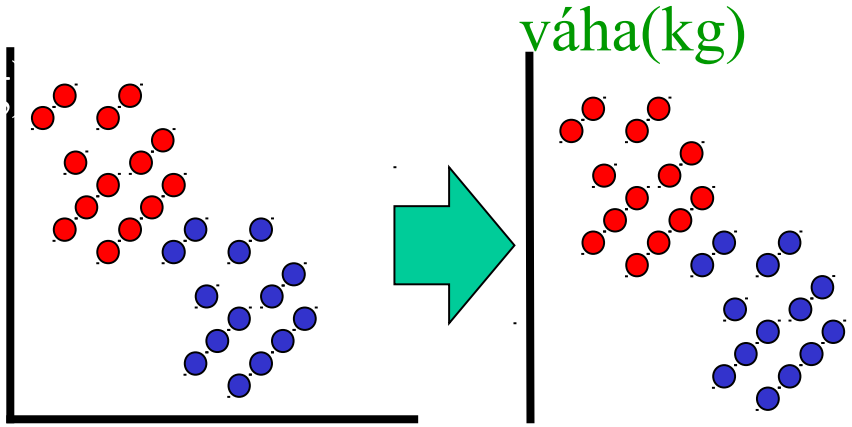
Extrakce příznaků (*Feature Extraction*)

- Vytvoří nový příznak, který může skupinu jiných nahradit
- Např. váha/výška

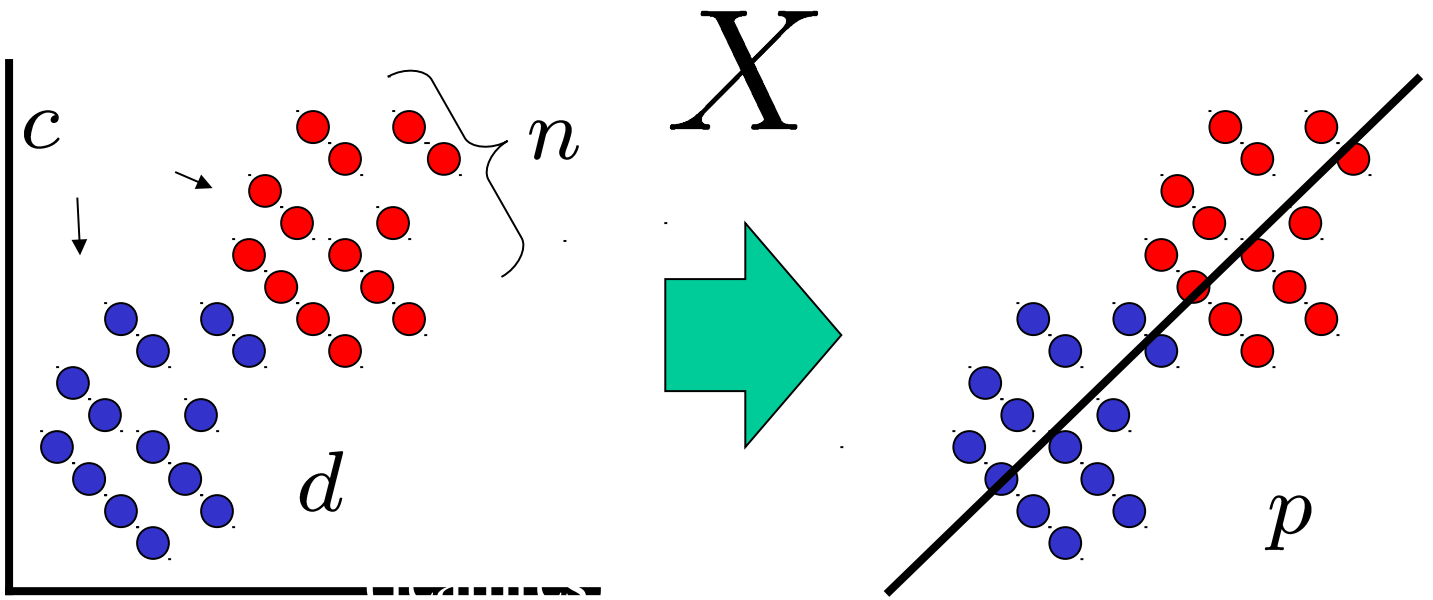


Výběr příznaků (Feature Selection)

- Vybere 1 nebo více příznaků, na které se sosutředí
- např. zachová p. váha (používá příslušný průmět)
- V tomto příkladě není klasifikace jednoznačná



Východiska



- Nechť matice X reprezentuje trénovací data odpovídající n vzorkům, z nichž každý je popsán pomocí d vlastností (atributů)
- V případě učení s učitelem víme dále, že data patří do c tříd
- Cílem redukce dimenze je získat (výběrem či extrakcí) p atributů tak, aby se podařilo zachovat co nejvíc z původní informace vzhledem k nějakému kritériu. Zhruba by mělo platit:

$$1 < p \simeq c \ll d < n$$

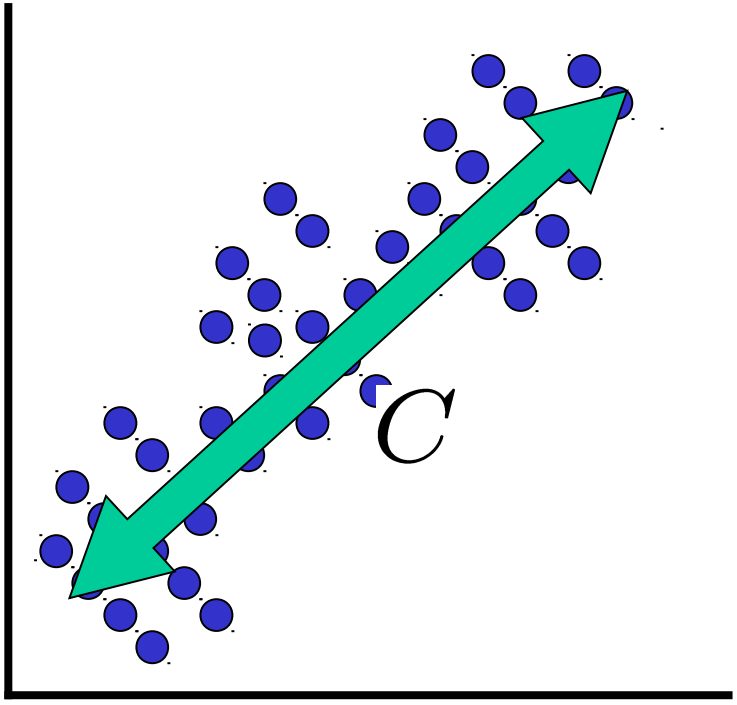
Metody pro extrakci příznaků

- postupují tak, že provádějí projekci dat do prostoru o nižší dimenzi
  - Metody bez učitele
    - Principal Component Analysis (PCA)
    - Independent Component Analysis (ICA)
  - Metody s učitelem
    - Linear Discriminant Analysis (LDA)
    - Maximum Margin Criterion (MMC)
    - Orthogonal Centroid algorithm (OC)
- a hledají co nejlepší matici projekce W, která zvýší výkon při řešení zpracovávané úlohy

Principal Component Analysis

- PCA se snaží maximalizovat $J(W) = trace(W^T C W)$

- pro výpočet PCA je potřeba znát hodnotu Singular Value Decomposition (SVD), jejíž výpočet určuje složitost úlohy:
  - časová sl.: $O(n^2 d)$
  - prostorová sl.: $O(n\ d)$



C : kovarianční matice

trace představuje celkovou varianci studovaných dat (počítá se jako součet diagonál v variance-kovarianční matici)

Linear Discriminant Analysis

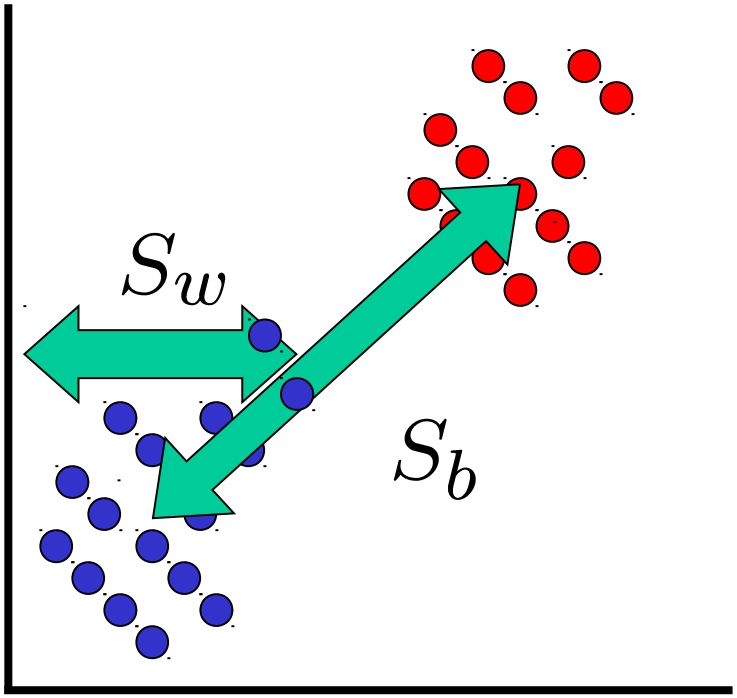
LDA (podobně jako PCA) hledá projekci (lineární kombinace původních atributů), která maximalizuje $\boldsymbol{S}_b$ a minimalizuje $\boldsymbol{S}_w$

Časová složitost

$$O((n+c)^2 d)$$

Prostorová složitost

$$O(nd)$$



S_b : „scatter matrix“ mezi různými třídami

S_w : „scatter matrix“ uvnitř jednotlivých tříd

Metoda používaná pro učení s učitelem

Potřebujeme selekci příznaků?

Nepochybně ano, protože

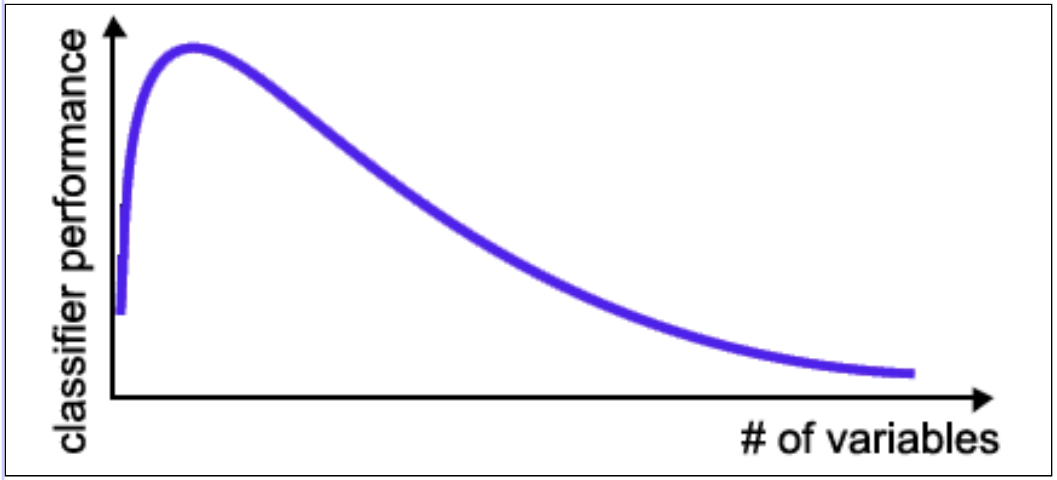
- Mnohdy pracujeme s daty charakterizovanými **stovkami** či dokonce **desítkami tisíc** atributů, z nichž mnohé jsou irelevantní nebo redundantní!
- Pro taková data je velmi obtížné zjistit (nebo odhadnout) jejich pravděpodobnostní distribuční funkci (např. kvůli vzájemné závislosti mezi některými atributy) !
- Informace od hodnotách irelevantních a redundantních atributů komplikují proces učení, neboť učící algoritmy „matou či zavádějí" !
- Teorie PAC v souvislosti s omezeným počtem trénovacích dat!
- Omezené výpočetní prostředky!
- **Prokletí dimensionality**!

Prokletí dimensionality

- Počet trénovacích příkladů ***m*** potřebných proto, aby byla vytvořena hypotéza o dostatečné přesnosti roste exponenciálně s počtem atributů! PAC: ***m*** > Počet_prvků (Prostor_hypotéz)

- V praktických úlohách bývá maximální počet trénovacích příkladů pevně dán!

  => výkon klasifikátoru (classifier performance) výrazně klesá s rostoucím počtem atributů (# of variables)!



Velmi často lze docílit toho, že

- ztráta informace vzniklá vynecháním některých atributů

je vyvážena

- daleko lepšími výsledky klasifikace v prostoru o nižší dimenzi !

Typický mnohodimenzionální problém

Kategorizace textů

Každý dokument je reprezentován prostřednictvím vektoru tvořeného údaji o frekvenci výskytu jednotlivých slov uvažovaného slovníku. Tedy délka vektoru je dána velikostí slovníku.

Slovník obvykle obsahuje asi 15.000 slov (pracujeme tedy s 15.000 atributy)

- Typické úlohy:
  - Automatická kategorizace dokumentů (podle tématu)
  - Identifikace spamu v emailech

Osnova

- Introduction/Motivation

- Základní definice, Terminologie

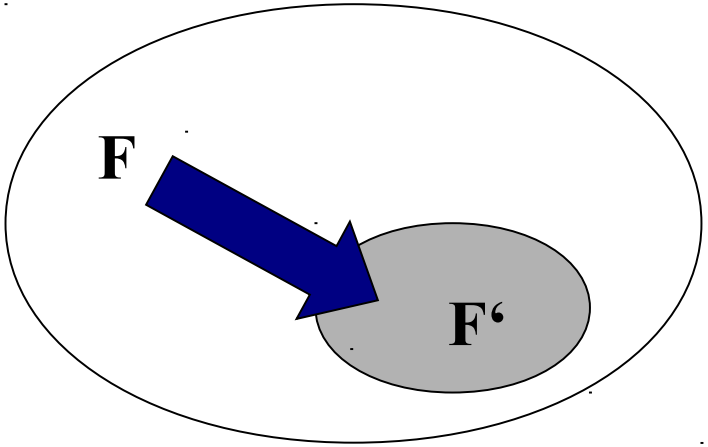
- Metody hodnocení příznaků (Variable Ranking methods

- Výběr podmnožiny příznaků

Selekce příznaků - definice

- Máme-li množinu příznaků $F=\{f_1,\ldots,f_i,\ldots,f_n\}$

pak cílem selekce příznaků je nalézt takovou podmnožinu $F'\subseteq F$ která "maximalizuje schopnost učicího algoritmu klasifikovat správně zpracovávaná data".



$$\{f_1,\ldots,f_i,\ldots,f_n\} \xrightarrow{f.selection} \{f_{i_1},\ldots,f_{i_j},\ldots,f_{i_m}\}$$

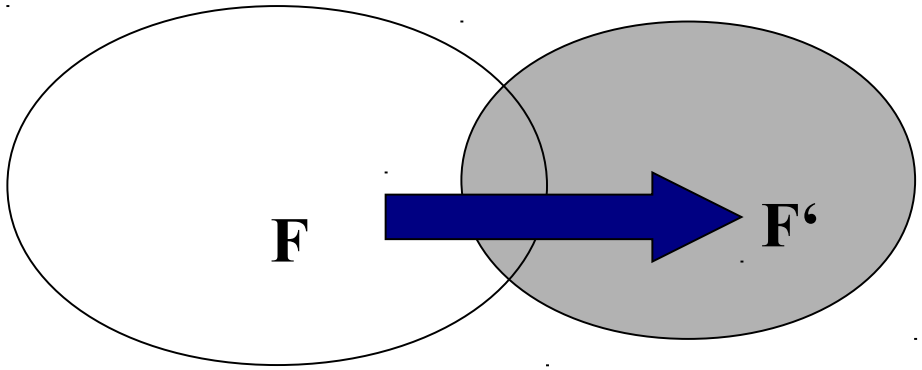
$$i_j \in \{1,\ldots,n\};\ j=1,\ldots,m$$

$$i_a = i_b \Rightarrow a=b;\ \ a,b \in \{1,\ldots,m\}$$

Extrakce příznaků - definice

- Máme-li množinu příznaků $F=\{f_1,\ldots,f_i,\ldots,f_n\}$,

  pak cílem extrakce (konstrukce) příznaků je nalézt takové zobrazení množiny F do množiny F', které "maximalizuje schopnost učicího algoritmu klasifikovat správně zpracovávaná data".



$$\{f_1,\ldots,f_i,\ldots,f_n\} \xrightarrow{f.extraction} \{g_1(f_1,\ldots,f_n),\ldots,g_j(f_1,\ldots,f_n),\ldots,g_m(f_1,\ldots,f_n)\}$$

Výběr optimální množiny příznaků ?

- Teoretický úkol zní „nalézt optimální podmnožinu příznaků (která maximalizuje zvolenou hodnotící funkci)“
- V reálných aplikacích je nemožné tento úkol splnit, protože
  - Ve většině úloh je výpočetně nezvladatelné prohledat celý prostor všech možných podmnožin (exponenciální složitost)
- Hledají se proto vhodné aproximace optimální podmnožiny
  - Většina výzkumu v této oblasti spočívá v hledání efektivních heuristik.

Relevance příznaků

- Kdy je příznak relevantní?
  - Je třeba rozlišit více možností:
    - Relevantnost 1 příznaku,
    - Relevantnost příznaku v kontextu jiných příznaků,
    - Relevantnost pro pevně zvolený algoritmus strojového učení,
    - …
  - Většina definic je problematická, neboť se pro ně dají nalézt takové příklady, kdy by žádný z použitých příznaků nebyl považován za relevantní, i když je možné z nich vycházet při rozhodování
  - Rozlišují [2] se dva stupně relevantnosti: **slabě** a **silně relevantní příznak**.
  - Příznak je relevantní, pokud je slabě či silně relevantní, jinak je redundantní (irelevantní).

[2] Ron Kohavi, George H. John: Wrappers for Feature Subset Selection. *AIJ special issue on relevance* (1996)

Definice relevantního příznaku

- **Silná relevance** příznaku f_i :

Nechť $S_i = \{f_1, \ldots, f_{i-1}, f_{i+1}, \ldots f_n\}$ je množina příznaků, ve které byl vynechán příznak f_i.

Příznak f_i je silně relevantní právě tehdy, když jeho vynechání ve všech datech vede vždy ke snížení kvality klasifikace při použití optimálního Bayesova klasifikátoru.

- **Slabá relevance** příznaku f_i :

Příznak f_i je slabě relevantní, pokud není silně relevantní, ale existuje podmnožina S_i' množiny S_i , pro kterou platí, že výsledek použití optimálního Bayesova pro data s těmito příznaky je horší než při použití téhož klasifikátoru v případě práce se souborem příznaků S_i' rozšířeném o f_i .

Vlastnosti relevance příznaku

- Relevance $\not\Leftrightarrow$ Optimalita množiny příznaků
- Klasifikátory vytvořené nad trénovacími daty bývají suboptimální (nemají k dispozici informaci o skutečné distribuci dat)
- Relevance příznaku ještě neznamená, že tento příznak musí být v optimální podmnožině příznaků
- Dokonce "irelevantní" příznaky mohou zlepšit výkon klasifikátoru
- Při definici relevance by měl být brán v úvahu použitý klasifikátor (a tedy o obor možných hypotéz).

Osnova

- Introduction/Motivation
- Basic definitions, Terminology
- Metody hodnocení příznaků
- Výběr podmnožiny příznaků

Hodnocení příznaků

- Pro danou množinu příznaků F proces hodnocení příznaků probíhá tak, že se příznaky uspořádají podle nějaké hodnotící funkce $S: F \rightarrow \Omega$ (která vyjadřuje míru relevance jednotlivých příznaků „čím vyšší, tím relevantnější")
- Výsledek: taková permutace F' původní množiny F, že:

$$F' = \{f_{i_1}, \ldots, f_{i_j}, \ldots f_{i_n}\}$$

$$S(f_{i_j}) \geq S(f_{i_{j+1}}); \quad j = 1, \ldots, n-1;$$

Hodnoty $S(f_i)$ jsou vypočteny z trénovacích dat.

- Hodnotící funkce představuje vhodnou heuristiku – takto lze nalézt suboptimální řešení v rozumném čase

Vhodná hodnotící kritéria

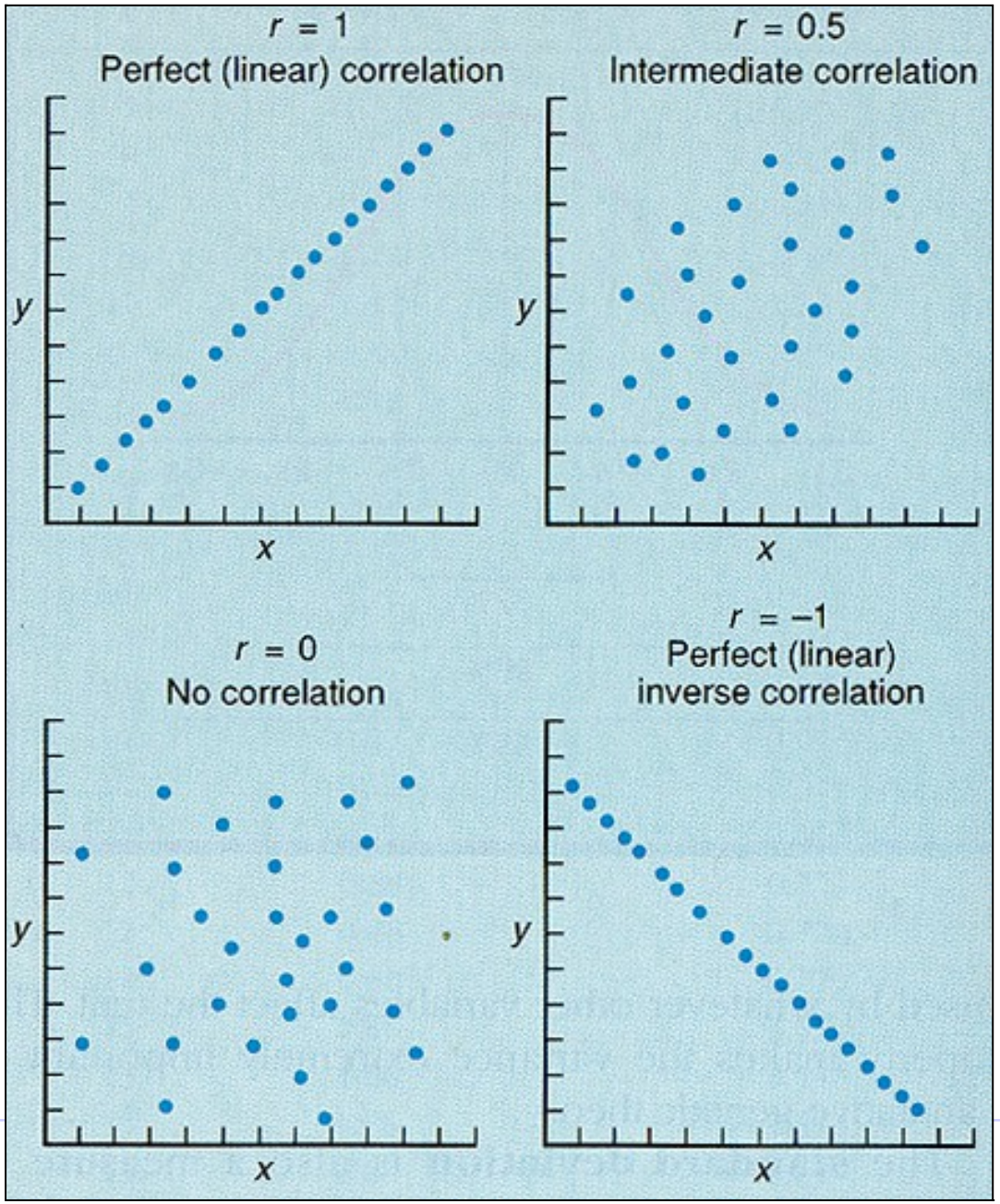
- Pearsonův korelační koeficient

$$\mathrm{R}(f_i, y) = \frac{\operatorname{cov}(f_i, y)}{\sqrt{\operatorname{var}(f_i)\operatorname{var}(y)}}$$

- Odhad pro m vzorků:

$$R(f_i, y) = \frac{\sum_{k=1}^{m}\left(f_{k,i} - \overline{f_i}\right)\left(y_k - \overline{y}\right)}{\sqrt{\sum_{k=1}^{m}\left(f_{k,i} - \overline{fi}\right)^2 \sum_{k=1}^{m}\left(y_k - \overline{y}\right)^2}}$$

Hodnotící kriterium – korelace



Nature Inspired
Technologies Group

Hodnotící kriterium – korelace

- $R(X_i, Y) \in [-1,1]$

  nejčastěji se proto používá $R(x_i,y)^2$ nebo $|R(x_i,y)|$

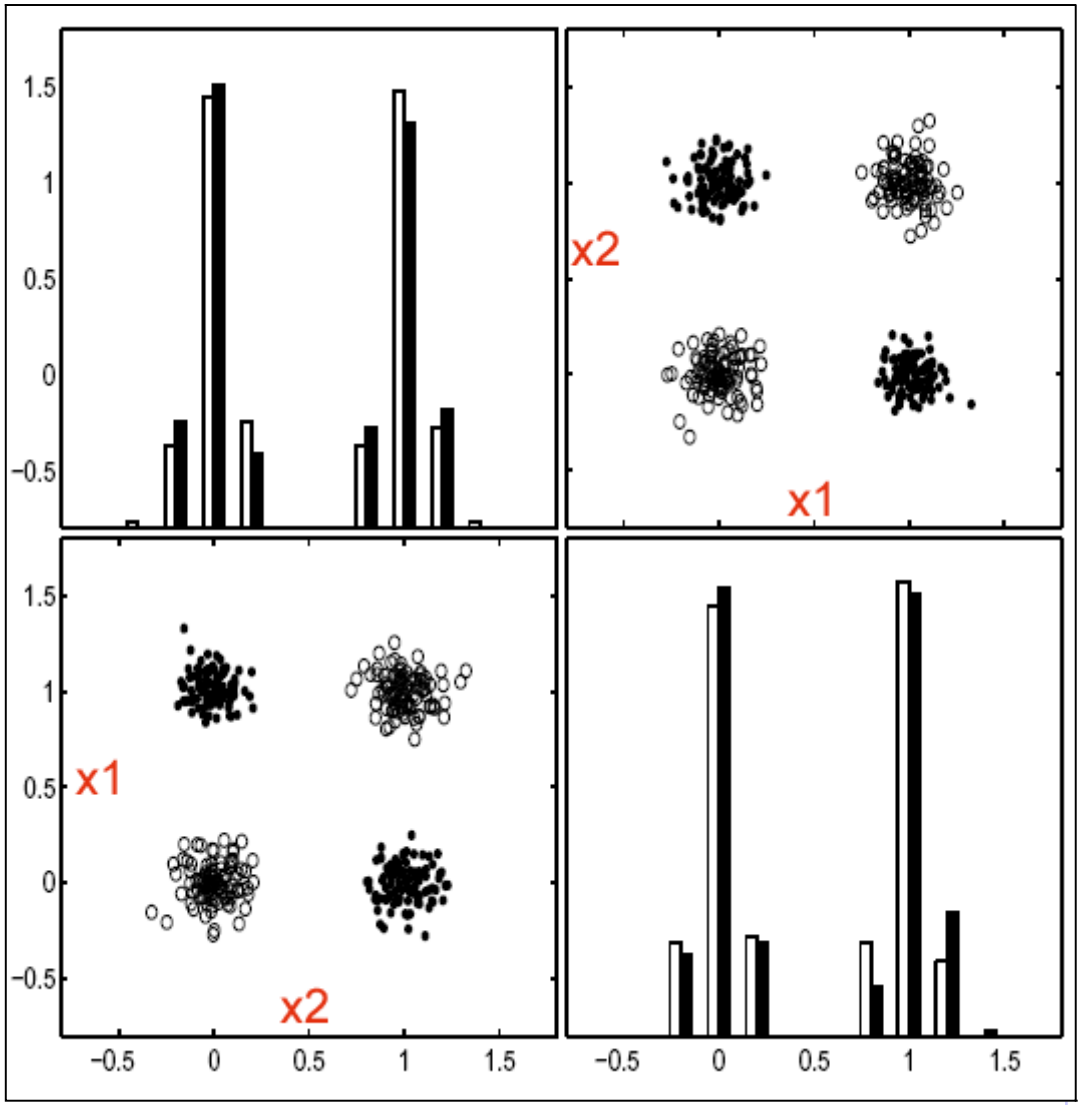
- Korelace odpovídá míře lineární závislosti mezi příznakem x_i a klasifikací y :
  - Může tedy odhalovat jen lineární vazby mezi příznakem a cílovou klasifikací .
  - To není případ $y = XOR(x1,x2)$?
- **Otázky** :
  - Je možné vždy vyloučit příznaky s nízkým hodnocením ?
  - Může být skupina (alespoň 2) příznaků s nízkým hodnocením užitečná?

Ranking Criteria – Correlation

- Může být skupina (alespoň 2) příznaků s nízkým hodnocením užitečná?

ANO!

Je nutné hledat další kriteria



Další hodnotící kritéria – Inf. Teorie

Informaticko-teoretická kriteria

- Většina postupů používá (empirické odhady) vzájmné informace mezi příznakem a cílovým hodnocením:

$$I(x_i, y) = \int_{x_i} \int_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)\,p(y)}\, dxdy$$

- Výpočet pro diskrétní příznaky:

$$I(x_i, y) = \sum_{x_i} \sum_{y} P(X = x_i, Y = y) \log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)}$$

(pravděpodobnost se odhaduje jako frekvence výskytu v trénovacích datech)

Osnova

- Introduction/Motivation
- Basic definitions, Terminology
- Variable Ranking methods
- Selekce podmnožiny příznaků

Výběr podmnožin příznaků

- Potřebujeme:
  - Míru pro měření kvality vybrané podmnožiny příznaků (hodnotící funkci)
  - Strategii prohledávání prostoru všech možných podmnožin

  => Good heuristics are needed!

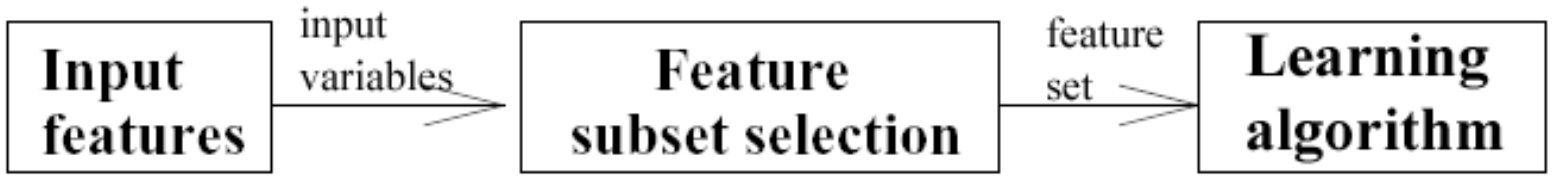
- Používané metody:
  - **Filtrační metody**
  - **Wrappers**
  - **Vnořené metody** jsou zabudované do jednotlivých algoritmů strojového učení (např. uvnitř ID3)

[9] E. Amaldi, V. Kann: The approximability of minimizing nonzero variables and unsatisfied relations in linear systems. (1997)

Výběr podmnožiny příznaků

Filtrační metody

- Vybírají podmnožinu příznaků obvykle v rámci předzpracování a nezávisle na tom, jaký bude použit klasifikátor!!



- Např. hodnocení jednotlivých příznaků je filtrační metoda
- **Výhody:** rychlé, universální (nezávisí na volbě následného použitého algoritmu)

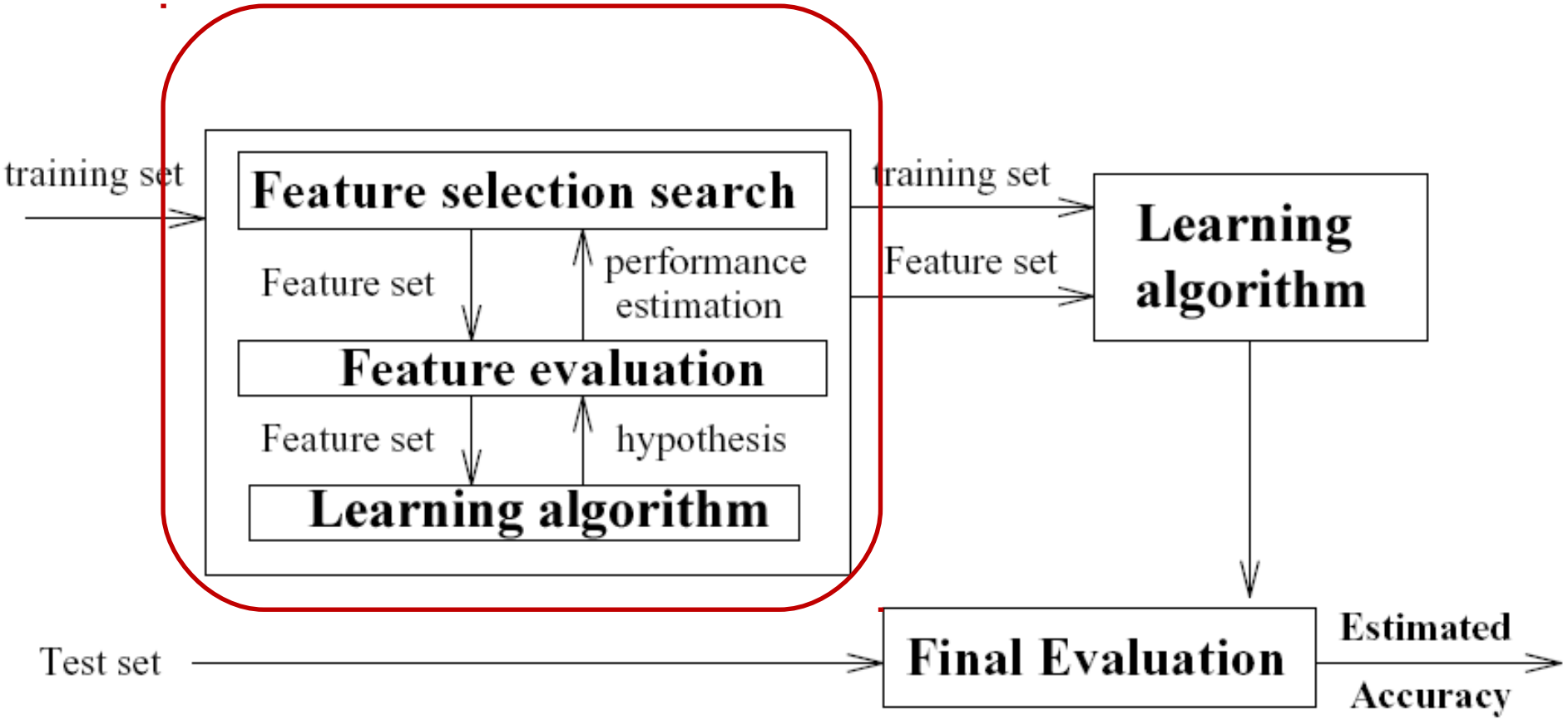
Výběr podmnožiny příznaků

❖ Wrapper Methods

- Wrapper prohledává prostor všech možných podmnožin příznaků a každou zvažovanou podmnožinu hodnotí tak, že její kvalitu otestuje na trénovacích datech s použitím nějakého učicího algoritmu (chápe se jako „černá skříňka")
- Výsledky se liší podle použitých metod strojového učení (učicích algoritmů)
- **Je nutné upřesnit** způsob :
  - prohledávání prostoru všech možných podmnožin (např. heuristické prohledávání dopředné nebo zpětné, hladové, ..)
  - Odhadu kvality výkonu dané množiny příznaků pro strojové učení (např. pomocí validační množiny, ...)

Výběr podmnožiny příznaků

Wrapper:



Závěrečné poznámky

- Vhodný výběr příznaků může významně zlepšit výkon při řešení problému strojového učení (přesnost i počítačová náročnost) – ale jedná se o náročnou úlohu!
- Je to cesta k řešení problémů s velmi mnoha atributy
- Pozor na vztah mezi relevancí a optimalitou (nelze automaticky ignorovat všechny příznaky s malým hodnocením – mohou mít význam v kombinaci!).
- **Prostor pro vylepšení ?**
  - Nový způsob prohledávání prostoru podmnožin příznaků
  - Odhad kvality aktuální množiny příznaků
- Malé množiny příznaků lze najít i při použití metody „boosting" (kombinace klasifikátorů) pro klasifikátory s jediným příznakem!