FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

PŘEDZPRACOVÁNÍ DAT

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

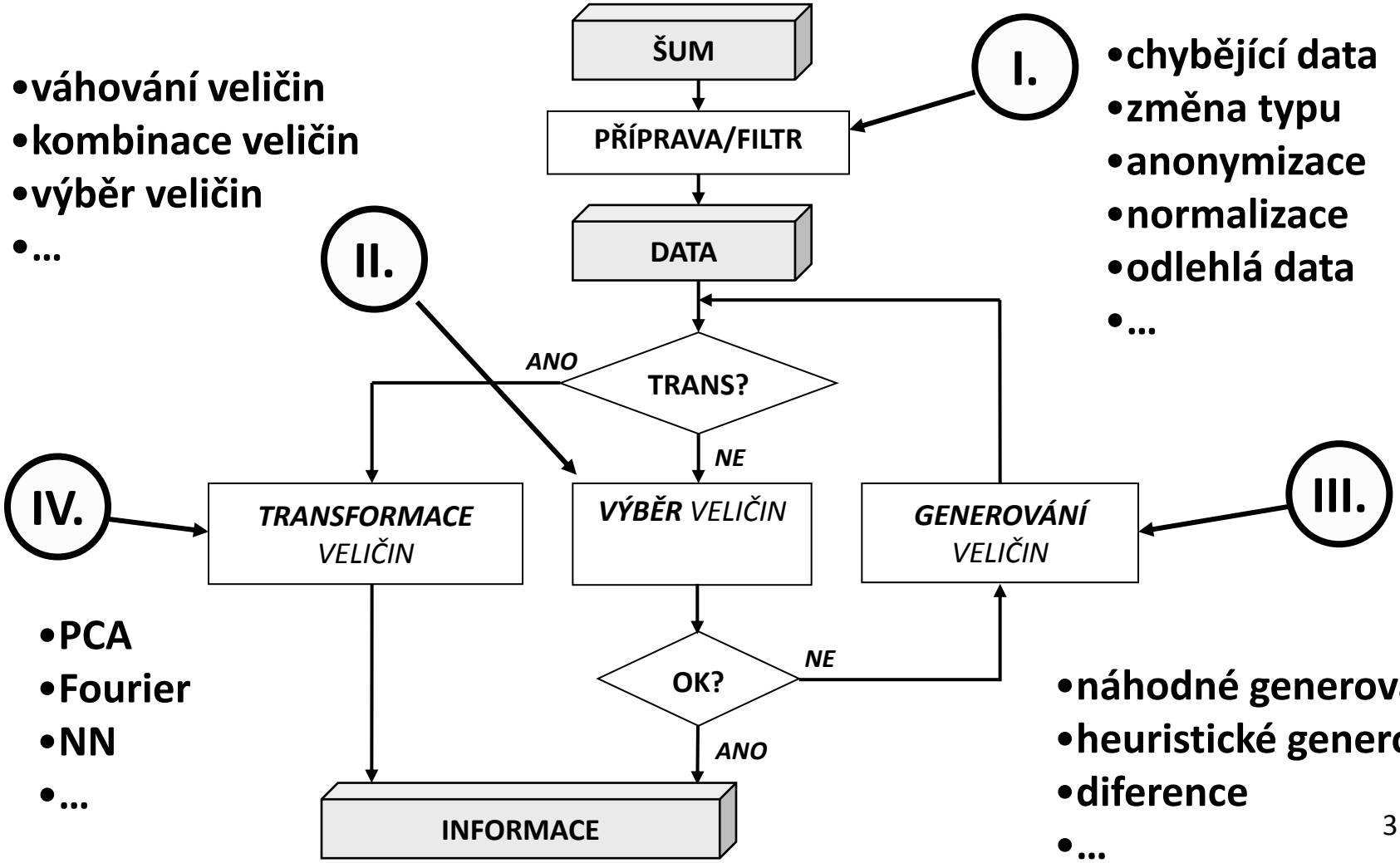
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Obecné principy

- generalizace je dosaženo **snižováním stupňů volnosti**, tedy také snižováním počtu vstupních proměnných
- ukazatelem je např. **vzájemná korelace** proměnných
- **minimální poměr** počtu záznamů *N* a počtu atributů *l* by měl být roven 20 ($N/l > 20$ orientační pomůcka)
- **odhad chyby modelu** je tím přesnější, čím je uvedený poměr větší
- důležitá je u jednotlivých veličin **velká odchylka průměrů** (mezi třídami) a **malý rozptyl** (uvnitř třídy)

?: *jak dosahujeme generalizace v fázi předzpracování dat*

Ideové schéma předzpracování dat



Komponenty předzpracování dat

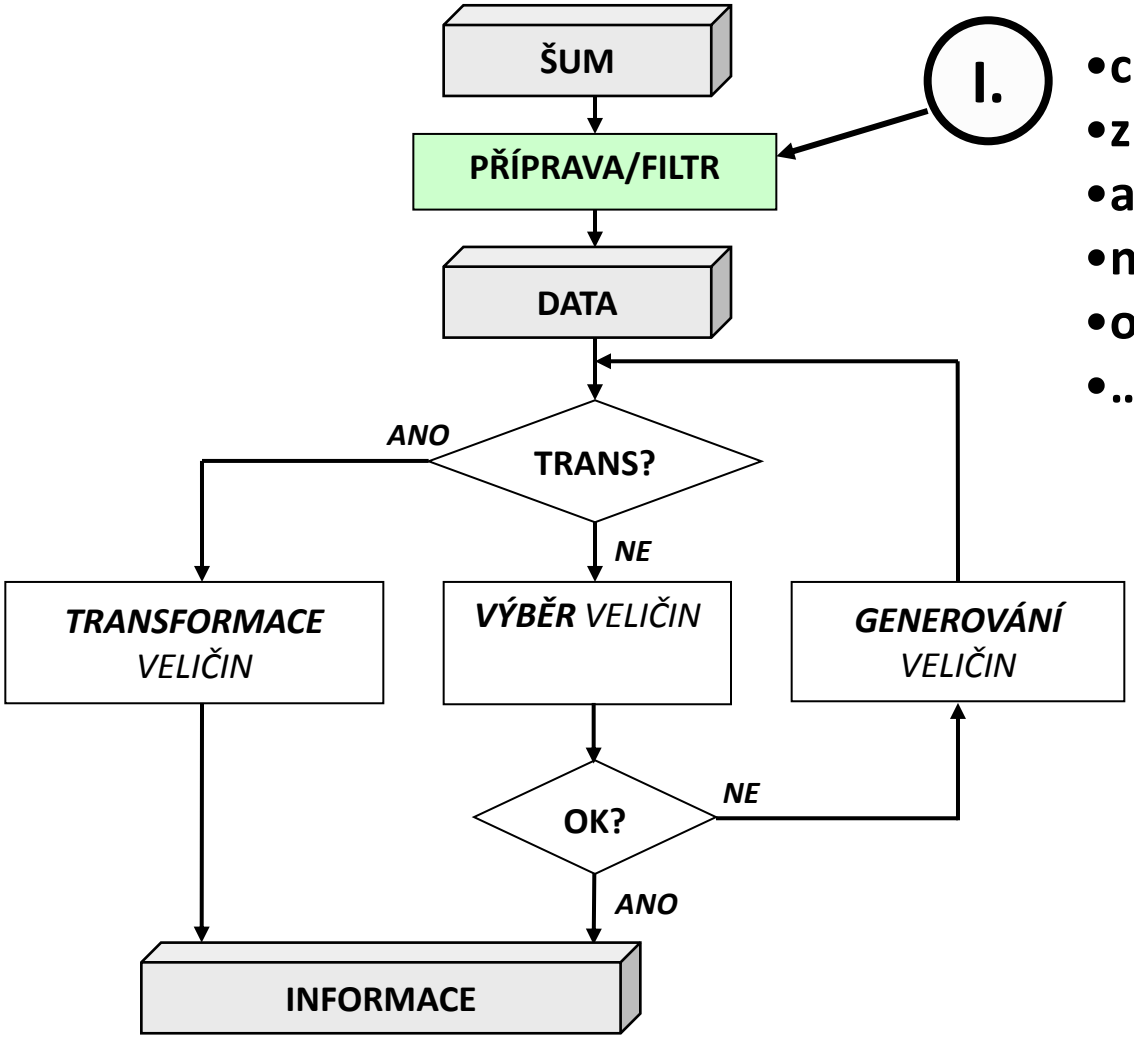
I. PŘÍPRAVA
   - připravená data lze přímo použít jako vstup do modelů (nezpůsobí funkční selhání)

II. VÝBĚR VELIČIN
   - z dostupných $|A|$ veličin vybrat ty, které nesou nejvíce informace (*skalární, vektorová*)

III. GENEROVÁNÍ
   - z dostupných dat vytvořit funkčními úpravami a vzájemnými kombinacemi nové veličiny (*náhodné, cílené, heuristické*)

IV. TRANSFORMACE
   - cílená redukce počtu veličin při maximálním možném zachování nesené informace (*PCA, NN*) nebo nová interpretace dat

I. PŘÍPRAVA DAT



- chybějící data
- změna typu
- anonymizace
- normalizace
- odlehlá data
- …

Základní operace

- Popis dat (in/out, typ + datum, text, …)
- Přetypování
  - kvalitativní na binární
  - kvalitativní na diskrétní (*dummy variables*)
  - kvantitativní na binární (práh, kritická hodnota)
- Zásah do dat
  - **normalizace**
  - **odlehlé hodnoty** (*ouliers*)
  - **chybějící hodnoty**
  - umělé zašumění
- Ostatní
  - anonymizace
  - ID

Typy normalizace

- **lineární** (typické intervaly ⟨0;1⟩, ⟨-1;1⟩)

$$x_{norm} = \frac{x - X_{\min}}{X_{\max} - X_{\min}}$$

- **střední hodnotou a rozptylem** (při normálním rozložení 99% v intervalu -3;3)

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{s} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$x_{norm} = \frac{x_i - \bar{x}}{\bar{s}} \qquad x_{norm} = \frac{x_i - \bar{x}}{r.\bar{s}}$$

- **logitovou funkcí**

  (interval (0;1) nebo (-1;1))

$$x_{norm} = \frac{1}{1 + e^{-x}}$$

?v jakém intervalu bude proměnná normalizovaná střední hodnotou a rozptylem;proč?

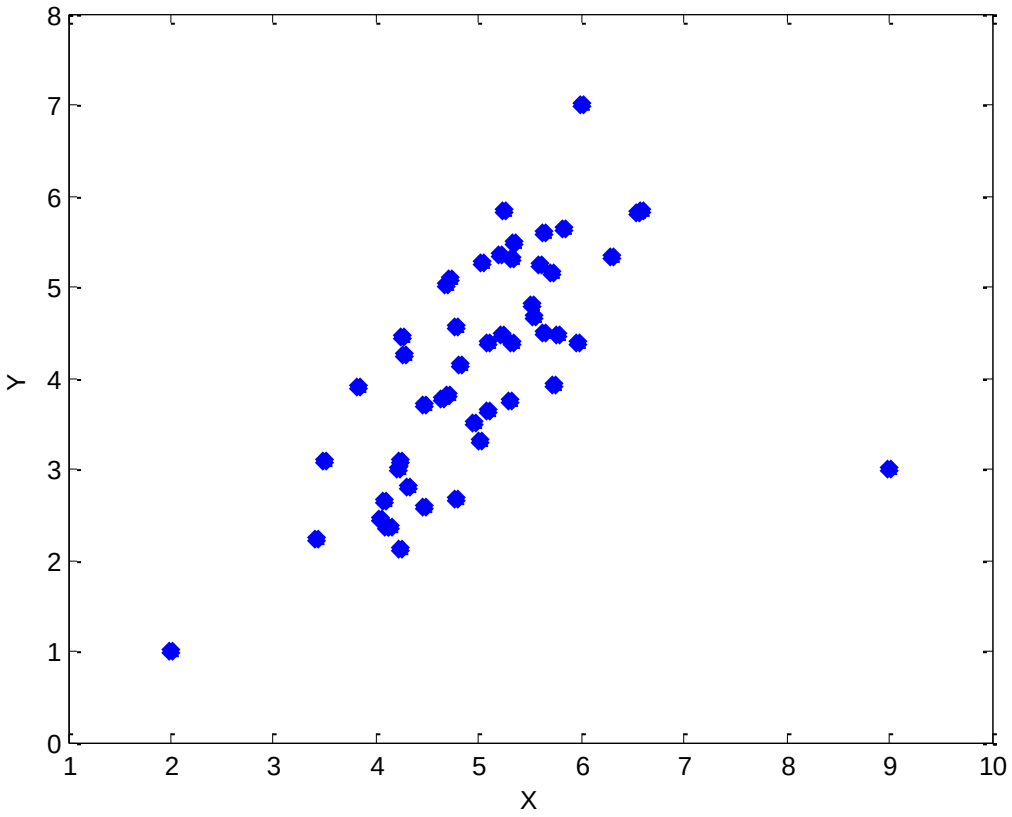
Chybějící data, odlehlé prvky

- **Chybějící data**
  - v případě dostatečného počtu dat **odstranit**
  - v případě nedostatečného počtu dat je základní variantou nahrazení **průměrem** z ostatních hodnot, případně **modusem**
  - existují lepší varianty vycházející z rozložení dat a podobnosti s existujícími kompletními záznamy, zde je však riziko vytvoření idealizované neexistující závislosti.
- **Outliers**
  - prvky ležící daleko od ostatních dat
  - zpravidla se určuje v násobcích rozptylu (např. dále než 3*s* od průměru)
  - na outliers je citlivá zejména MNČ

?: co jsou to outliers

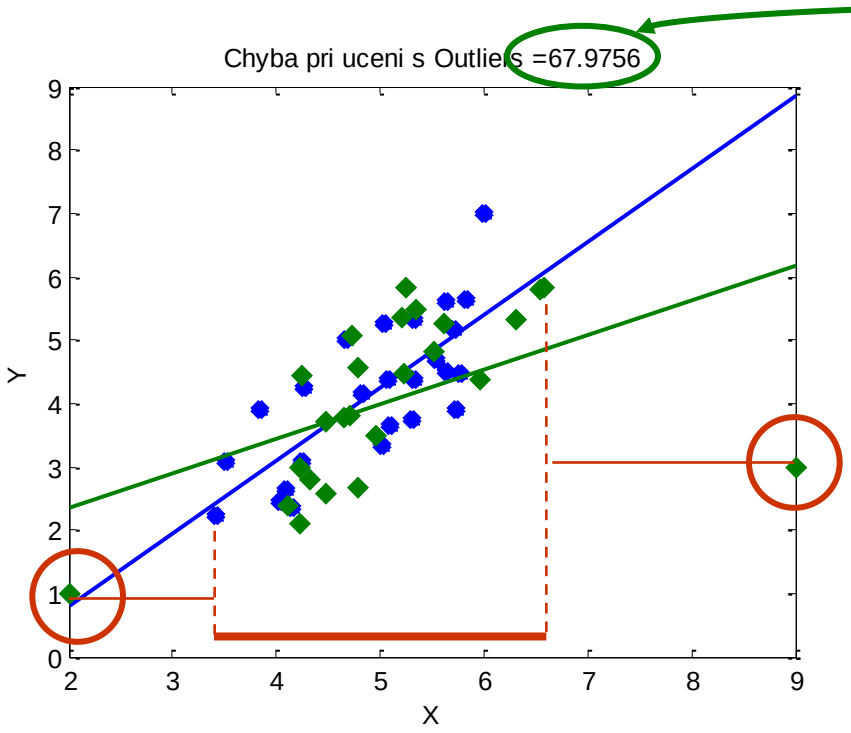
Outliers princip – příklad 1/2

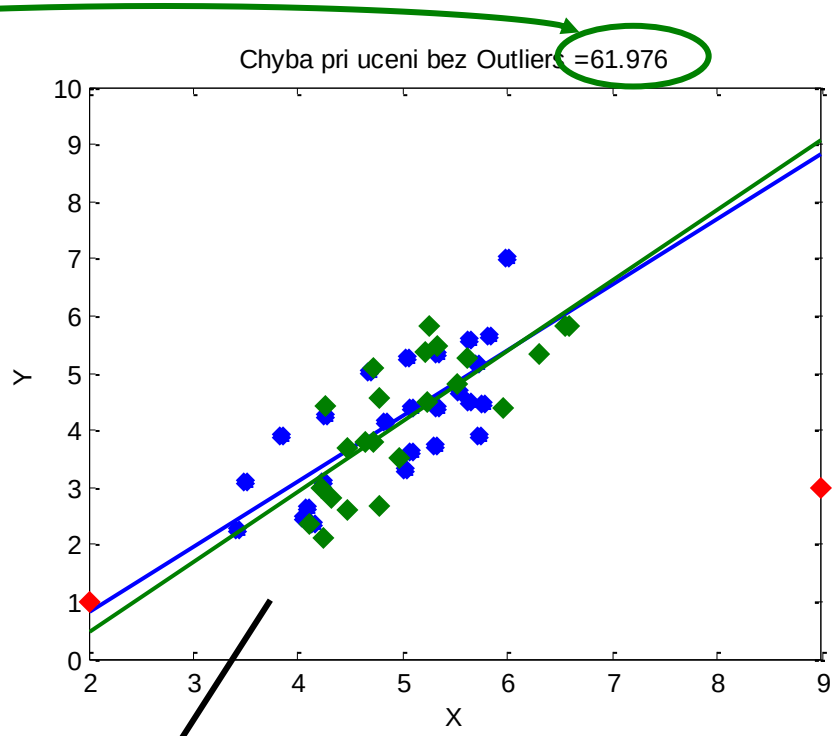
Proložte následující data lineární funkcí



Outliers princip – příklad 2/2

Výsledky při Cross Validation = 2 (modrá, zelená)

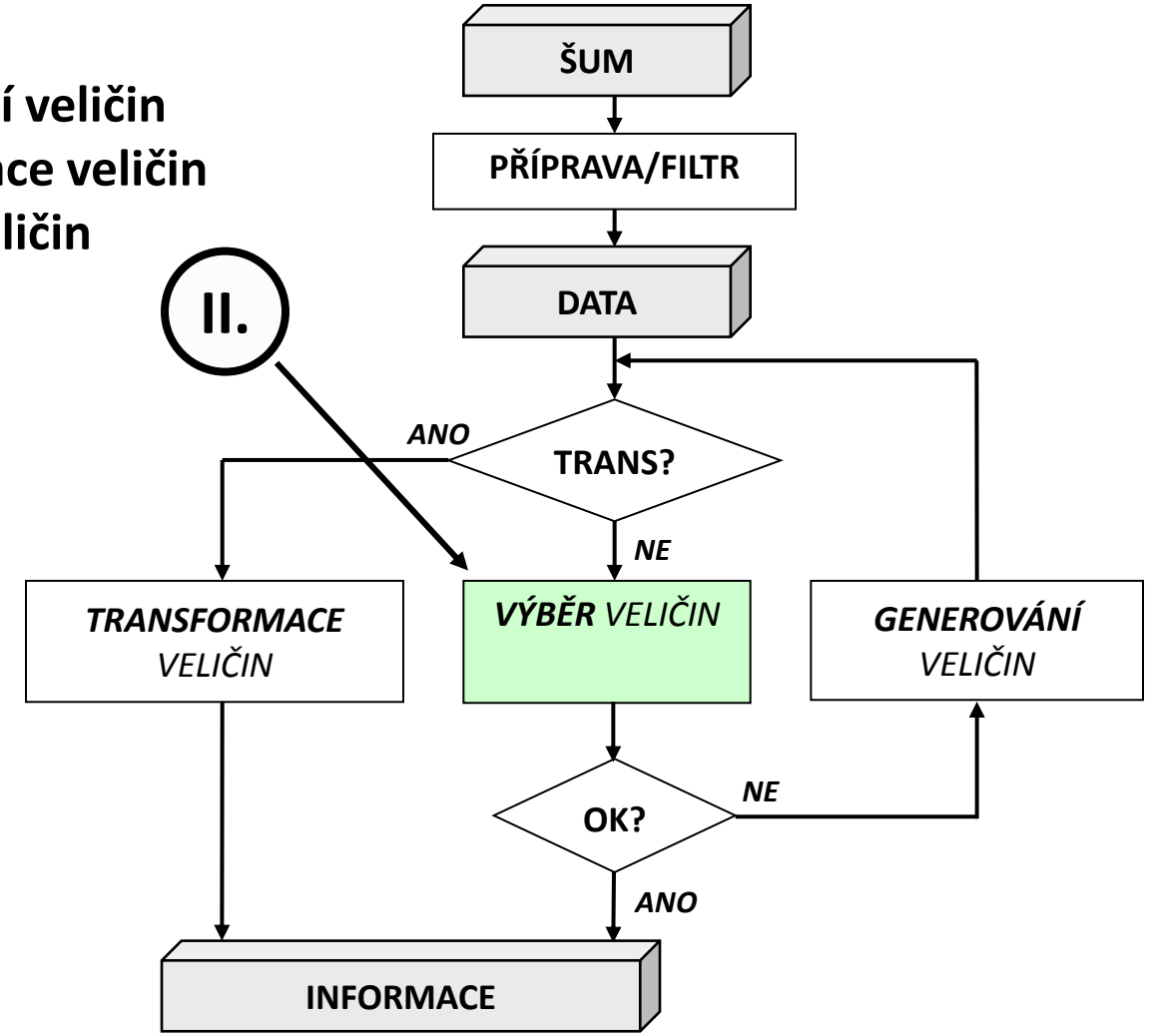




! Outliers **NEPOUŽITY** při trénování, **POUŽITY** při testování **!**

II. VÝBĚR VELIČIN

- váhování veličin
- kombinace veličin
- výběr veličin
- …



Problém výběru příznaků

- Mějme M příznakových vektorů $x_1, \ldots, x_M$
- Cílem je najít podmnožinu S příznaků, pro kterou model dosahuje nejvyšší přesnosti
- Počet takových řešení jest 2^M
- Nalezněme pro dané k optimální podmnožinu S_k (nesoucí v k příznacích maximální možnou informaci o výstupní veličině)
- Potom **ne vždy platí**, že $S_k \subset S_{k+1}$
- V důsledku toho gradientní metody negarantují nalezení optimálního řešení, ale prohledávaný prostor jest M^2 místo 2^M
- Není znám algoritmus jiný než brutal-force garantující nalezení optimální podmnožiny příznaků

Metody výběru (selekce) příznaků

- Zjednodušení prohledávání vstupního prostoru (hledání optimální množiny příznaků) gradientní (či „hladovou“, greedy search) metodou lze zapsat následujícím způsobem:

$$\max_{\forall \boldsymbol{f} \in \mathbb{U}} Eval(c; f \cup \mathbb{S}) = \max_{\forall \boldsymbol{f} \in \mathbb{U}} (Eval(c; f|\mathbb{S}) + Eval(c; \mathbb{S}))$$

Jedno/více rozměrné metody	Míra kvality	Aproximace $Eval(c; f\|\mathbb{S})$	Metoda hledání	Speciální označení	Konrkétní metody
Univariate	Evaluation Function	Unconditional	Greedy Search	FILTER	IG, MI (IGR), AUC, Relief, ChiSquare, FDR, …
		Conditional	Greedy Search		mRMR, MIFS, DMIFS, CMIM, …
Multivariate	Evaluation Function		Greedy Search		Relief, Scatter…
	Performance Estimation		Brutal Force Search		odhad "performance" pomocí modelu a křížové validace
			Population Based Search		
			Greedy Search	WRAPPER	

Typy výběru veličin

- **JEDNOROZMĚRNÉ (univariate)**
  - **Váhování** jednotlivých veličin
    - Gini Index, Information Gain, Chi-Square, Relief
    - t-Test
    - ROC analýza (AUC)
    - scatter matrix
  - Výběr konkrétních veličin
- **VÍCEROZMĚRNÉ (mutlivariate)**
  - skalární
    - kritérium + korelace
    - mRMR
  - vektorová
    - dobředná/zpětná selekce

Metody selekce – orientační dělení

Filter

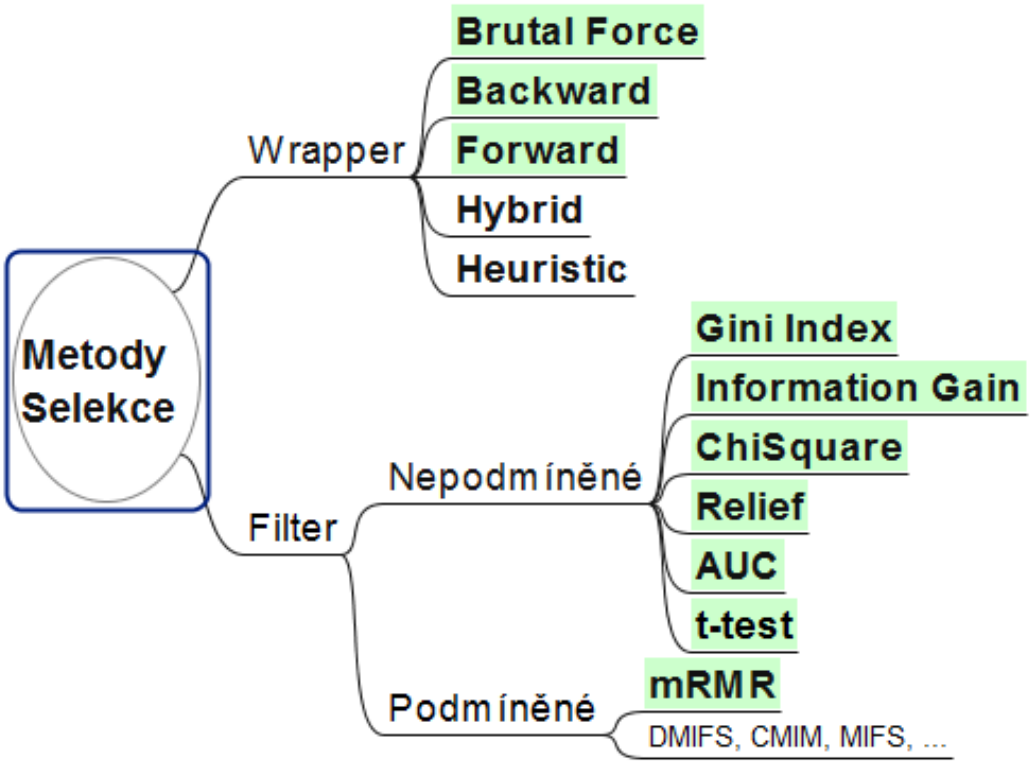
- váhuje veličiny zvlášť
- výpočetně méně náročné
- pro rozsáhlé databáze

Podmíněný Filter

- výpočetní náročnost různá
- vhodnost příznaku závisí na množině již vybraných

Wrapper

- používá během selekce cílový model, přesnější
- přesnost pomocí *accuracy*
- pořadí výběru veličiny odpovídá pořadí její relevance
- výpočetně náročný



Filtery – 3 základní principy

AUC, Relief

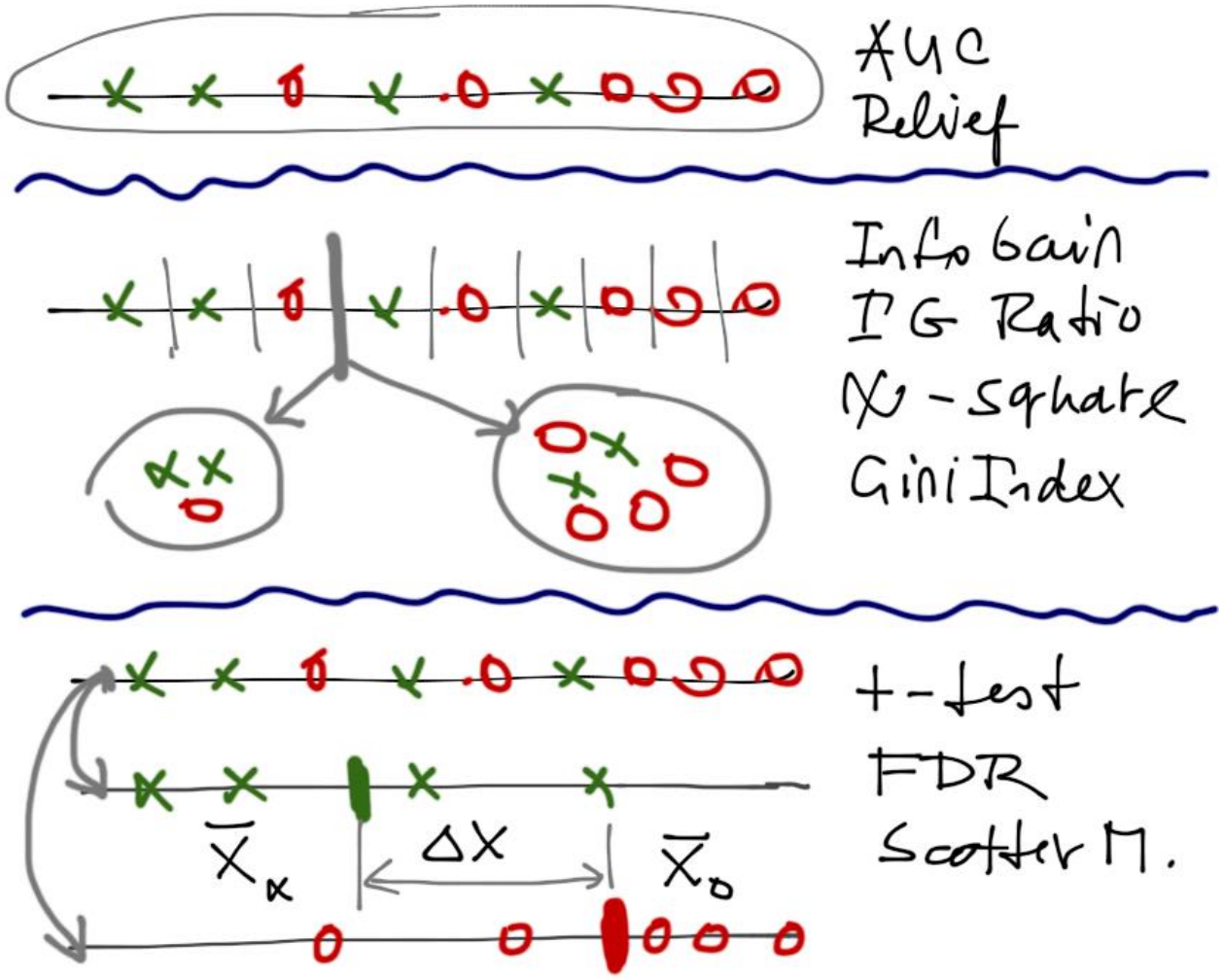
- vypočítají vlastnost veličiny jako celku

InfoGain, GiniIndex, …

- rozdělí veličinu na *N*-1 podmnožin
- nejlepší rozdělení charakterizuje celou veličinu

T-test, FDR, …

- míra rozdílu mezi středními hodnotami tříd normovaná rozptyly



Gini Index, Information Gain, Chi-Square

- veličina rozdělena na 2 části = N-1 způsobů (pokud všechna měření různá), míra separability odpovídá největší (Information Gain, Chi-Square) nebo nejmenší (Gini Index) hodnotě

- **Information Gain (Mutual Information)**

$$I(S,A) = H(S) - H(S|A)$$

- **Gini Index**

$$GI = \frac{N_L}{N}\left(1-\sum_{i=1}^{c}\left(\frac{N_{Li}}{N_L}\right)^2\right)+\frac{N_R}{N}\left(1-\sum_{i=1}^{c}\left(\frac{N_{Ri}}{N_R}\right)^2\right)= p_L\left(1-\sum_{i=1}^{c} {p_{Li}}^2\right)+ p_R\left(1-\sum_{i=1}^{c} {p_{Ri}}^2\right)$$

- **Chi-Square**

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{c}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}}$$

Relief

- princip podobný metodě k-NN
- předpoklad: ve vhodné veličině bude mít vybraný prvek dané třídy blíž k prvku stejné třídy než třídy jiné
- vzdálenost nejbližšího stejné třídy (nearest hit) $\left|x_i - x_i^{HIT}\right|$
- vzdálenost nejbližšího různé třídy (nearest miss) $\left|x_i - x_i^{MISS}\right|$
- relief R je definován:
  - čím větší, tím lepší

$$R = \frac{1}{N}\sum_{i=1}^{N}\left(\left|x_i - x_i^{MISS}\right| - \left|x_i - x_i^{HIT}\right|\right)$$

- Existuje obrovské množství variant a rozšíření, např.:
  - různé metriky (zde Manhattanská vzdálenost)
  - provádí se přes M náhodně vybraných prvků
  - MISS a HIT se počítá přes více sousedů
  - vzdálenost lze počítat i přes více veličin (multivariate)

ROC křivka (1/5)

- **ROC** (Receiver Operating Characteristic) je diskrétní charakteristika. Každý její bod je dán dvěma hodnotami – FPR (1-*senzitivita*) a TPR (*specificita*).
- **Čtyřpolní tabulka** zobrazuje počty úspěšně (TP, TN) a neúspěšně (FP, FN) klasifikovaných prvků (binární klasifikace).

$$TPR = senzitivita = \frac{TP}{TP + FN}$$

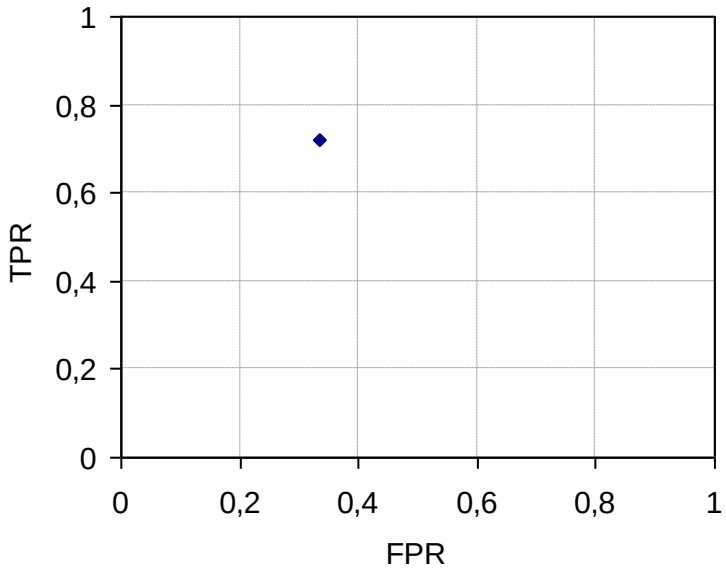
$$FPR = 1 - specificita = \frac{FP}{FP + TN}$$

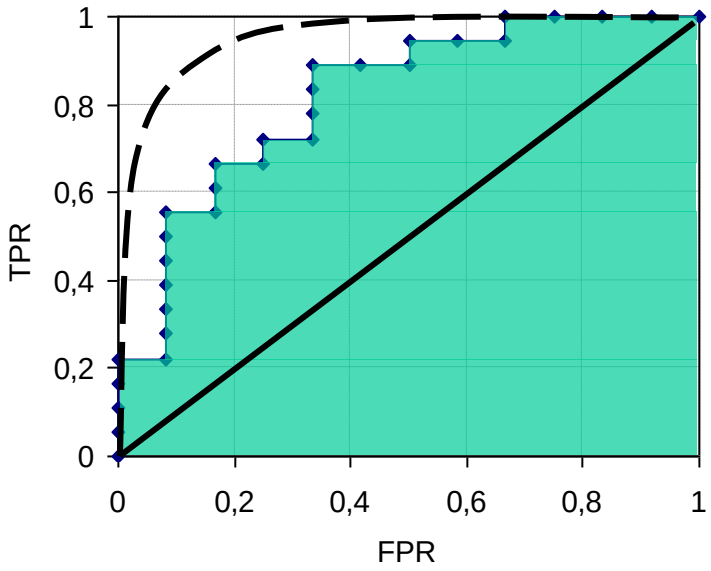
$$specificita = \frac{TN}{FP + TN}$$

Predikce	Skutečný výstup	
	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

AUC – plocha pod ROC (2/5)

- **AUC** (Area Under ROC) je plocha pod ROC křivkou.
  - AUC = 0,5 jedná se o jev náhodný
  - AUC >0,8 (<0,2) hovoříme o signifikantní míře asociace (separability)





Jeden bod v grafu odpovídá jednomu nastavení modelu nebo jedné kritické hodnotě příznaku. Soubor všech možných kritických hodnot vytváří křivku ROC.

?: čemu odpovídá jeden bod v ROC grafu?

AUC – základní vlastnosti (3/5)

- **AUC** vyjadřuje neparametrickou míru asociace (separability), tedy jakou měrou asociuje veličina *X* veličinu *Y*.
- **Nezávislá na rozložení** veličin, její hodnota však nenese absolutní informaci, je to *míra* (pracuje s veličinami jako s ordinálními, je-li tedy *X* kvantitativní, ztrácí informaci).
- **Binární klasifikace** – AUC lze snadno vypočítat, představuje plochu pod dvourozměrnou křivkou.
- **Vícerozměrná klasifikace** – výpočet vzájemných AUC mezi všemi dvojicemi výstupních tříd (4 třídy = 6 výpočtů AUC)
- **Obdobné parametry** – AUC lze získat úpravou charakteristik: Somersovo D_{xy}, Gini index, Mann-Whitney U

?: co vyjadřuje hodnota AUC? Jaký je rozdíl mezi daty mající AUC=0,8 a AUC=0,2

AUC – výpočet (4/5)

- **Binární klasifikace** – vypočítáme např. podle následujícího vztahu (existuje jich více)

$$AUC = \frac{1}{n^+ n^-} \sum_{j=1}^{n^+} \sum_{k=1}^{n^-} g\left(x_j^+ - x_k^-\right)$$

Kde n^+/n^- odpovídá počtu prvků klasifikovaných jako pozitivní/negativní, x_j^+/x_k^- určuje velikost *j*-tého/*k*-tého prvku vstupní veličiny a *g(x)* je heavisidova funkce (pro: $x<0$ je $g(x)=0$; $x=0$ je $g(x)=0{,}5$; $x>0$ je $g(x)=1$)

- **Vícerozměrná klasifikace** – vytvoříme kombinace všech párů výstupních tříd, určíme jejich AUC a ty nakonec zprůměrujeme

$$AUC = \frac{2}{C(C-1)} \sum_{\forall i,j:i\neq j}^{C} AUC\left(c_i, c_j\right)$$

kde $AUC(c_i,c_j)$ je hodnota *AUC* z výběru prvků spadajících pouze do tříd c_i a c_j

AUC – příklad (5/5)

Rozhodni, zda je uvedený vektor vhodný pro další modelování.

G	×	× o ×	o	× ×	o	× o o
X	1	2	3	4	5	6

vektor $x_\times$ pro třídu × : (1,2,2,4,4,6); $|x_\times| = 6 = n^+$

vektor x_o pro třídu o : (2,3,5,6,6); $|x_o| = 5 = n^-$

$$AUC = \frac{1}{5 \cdot 6} \sum_{j=1}^{5} \sum_{k=1}^{6} g\left(x_k^o - x_j^\times\right) = \frac{21}{30} = 0{,}7$$

Z pohledu míry separability (asociace) tedy nebudeme uvedenou veličinu považovat za vhodnou.

T-test: je veličina *X* separabilní? (1/5)

- **Dáno**: vstupní veličina ***X***, výstupní ***G***, klasifikuje se do dvou tříd A, B
- **Testujeme** hypotézu o nulovém rozdílu skutečných středních hodnot μ_A a μ_B veličin $\boldsymbol{X}_A$ a $\boldsymbol{X}_B$ (notací $\boldsymbol{X}_A$ je myšleno každé ***X***(i), pro které ***G***(i)=A)
  - H_0: $\Delta\bar{x} : \bar{x}_A - \bar{x}_B = 0$
- **Předpokládáme**
  - normální rozložení veličin $\boldsymbol{X}_A$ a $\boldsymbol{X}_B$
  - rozptyl není statisticky významně odlišný (F-test, H_0: s_A=s_B)
- **Vyhodnocení** testu na dané hladině významnosti α
  - zamítnutí H_0 a přijetí H_1, střední hodnoty se významně liší, veličina může mít dobré separabilní vlastnosti, lze dále využít
  - nezamítnutí H_0 a tedy i její přijetí, zřejmě nevhodné separabilní vlastnosti, veličinu nepoužijeme

?:co testujeme (jakou nulovou hypotézu) při použití t-testu?

T-test: výpočet (2/5)

- **Výpočet:** stanoví se koeficient *q*, jehož hodnota je porovnána s hodnotou v tabulce t-rozložení při stupni volnosti N_A+N_B-2 na zvolené hladině významnosti α.

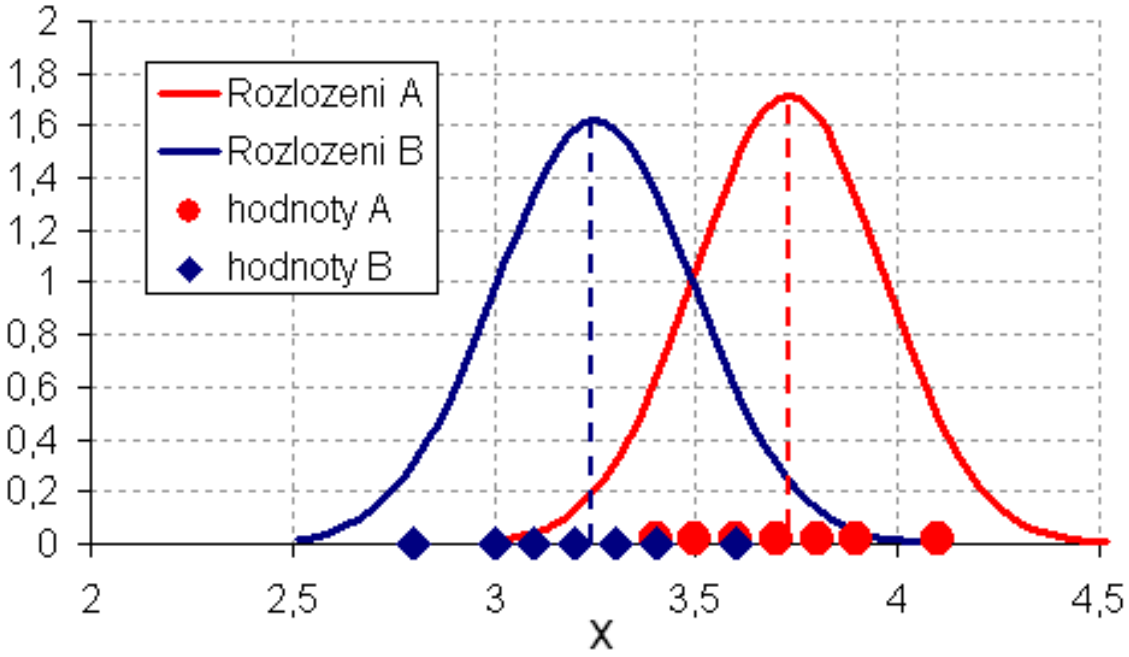
$$s_z = \sqrt{\frac{\left(\sum_{i=1}^{N_A}(x_i-\bar{x}_A)^2+\sum_{i=1}^{N_B}(x_i-\bar{x}_B)^2\right)}{N_A+N_B-2}} \qquad q = \frac{(\bar{x}_A-\bar{x}_B)}{s_z\sqrt{\frac{1}{N_A}+\frac{1}{N_B}}}$$

- **Interpretace**: pokud je hodnota *q* vyšší než příslušná hodnota v tabulce, tvrdíme, že hypotézu H_0 lze zamítnout na hladině významnosti α a přijímáme alternativní hypotézu H_1.

T-test: příklad (3/5)

- **Rozhodni**, zda je veličina *X* vhodná (separabilní z pohledu klasifikace) pro predikci výstupní binární veličiny *G*.

X	3,5	3,7	3,9	4,1	3,4	3,5	4,1	3,8	3,6	3,7	3,2	3,6	3,1	3,4	3	3,4	2,8	3,1	3,3	3,6
G	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B



T-test: příklad (4/5)

- **F-test**: H_0: $s_A=s_B$ nezamítnuta, p(H_0)=0,87; (Excel, *FTEST*)
- **T-test** (Matlab):

```
[H,P,CI,STATS] = TTEST2(xA,xB,0.05,0)
%prijimame alternativni hypotezu H1
 H   = 1
%pravdepodobnost vyberu pri platnosti H0
 P   = 4.7769e-004
%tstat - tabelovana hodnota, df - stupne volnosti
 STATS      = tstat: 4.2537    df: 18
```

- **Hypotézu H_0** na hladině významnosti α=0,05 zamítáme, veličina má nadějné separabilní vlastnosti (na základě výsledku P lze zamítnout dokonce na α=0,0005)

T-test: další informace (5/5)

- **Veličina nemá normální rozložení**, pak lze použít neparametrické testy - Wald-Wolfofiz runs test, Mann-Whitney U test, Kolmogorov-Smirnov two-sample test
- **Vícerozměrná klasifikace** – používá se ANOVA test, neparametrickou alternativou je Kruskal Wallis analysis of ranks, Median test
- **Závislé vs. nezávislé vzorky**
  - o nezávislém vzorku hovoříme, pokud test provádíme v rámci jednoho příznaku, což jest případ míry separability proměnné (t-test for independet samples)
  - o závislém vzorku hovoříme, pokud posuzujeme rozdíl průměrů mezi různými příznaky, jejichž hodnoty byly získány měřením na totožných objektech (t-test for dependent samples)

Scatter matrix (1/4)

- **míra separability**
- **výhodou** metody je nezávislost na typu rozložení
- **princip metody** spočívá v porovnávání rozptylů uvnitř jednotlivých tříd a rozptylu globálního (v rámci celého definičního oboru)
- **jednorozměrný prostor** odpovídá situaci, kdy chceme posoudit míru separability jedné veličiny; pak pracujeme s rozptyly a hlavním ukazatelem je FDR (Fisher's discriminant ratio)
- **vícerozměrný vstupní prostor** nepoužívá rozptylů ale kovariančních matic (značeny S_i), pak se používá ukazatelů J_1, J_2 a J_3

Fisherův diskriminační poměr FDR (2/4)

- **FDR** (Fisher's discriminant ratio) je
  - míra separability odvozená z metody scatter matrix
  - lze použít v případě jedné vstupní veličiny *x* a klasifikace do libovolného počtu tříd.

- **hlavní výhodou** je nezávislost na typu rozložení veličiny *x*

$$FDR = \sum_{i=1}^{M} \sum_{j \neq i}^{M} \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}$$

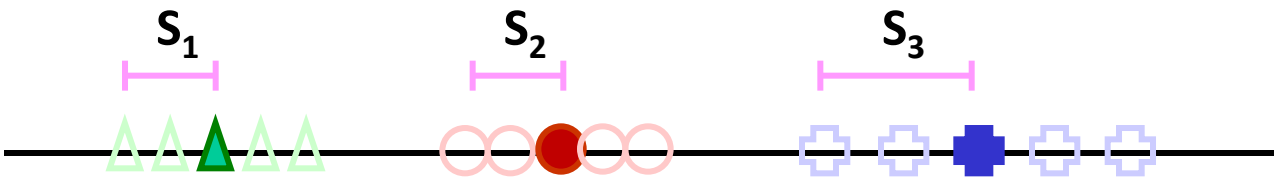
- **míra separability** je úměrná velikosti FDR (čím větší, tím lepší)

?: co je to FDR (Fisher's discriminant ratio)?

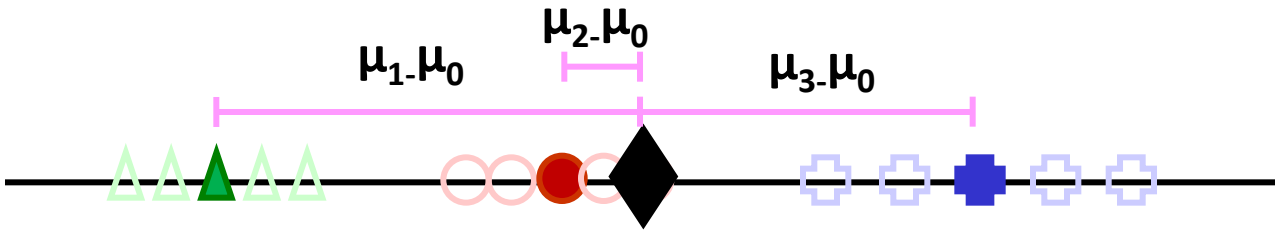
Vícerozměrný scatter matrix (3/4)

(zjednodušená 1-rozměrná ilustrace principu)

- **rozptyl (kovariance) uvnitř jednotlivých tříd** (S_w)



- **rozptyl (kovariance) mezi jednotlivými třídami** (S_b)



- **Definice matic S_w, S_b a S_m** $\quad S_m = S_w + S_b$

$$S_w = \sum_{i=1}^{M} P_i S_i \qquad S_b = \sum_{i=1}^{M} P_i (\mu_i - \mu_0)(\mu_i - \mu_0)^T \qquad \mu_0 = \sum_{i=1}^{M} P_i \mu_i$$

Míry separability (4/4)

- **Míry separability** se odvozují na základě poměrů jednotlivých kovariančních matic. Typické jsou následující parametry:

$$J_1 = \frac{trace\{S_m\}}{trace\{S_w\}} \qquad J_2 = \frac{|S_m|}{|S_w|} = |S_w^{-1} S_m| \qquad J_3 = trace\{S_w^{-1} S_m\}$$

- **funkce trace** je součet prvků na hlavní diagonále matice (*determinant matice je roven součinu vlastních čísel matice, součet prvků na hlavní diagonále pak součtu vlastních čísel matice*)

- **výhodou** ukazatelů J_2 a J_3 je jejich nezávislost na typu rozložení vstupní veličiny

- **interpretace**: čím je ukazatel větší, tím je separabilita větší; nenese absolutní informaci

Výběr veličin podle vah

- Výše uvedené metody umožňují **váhování** jednotlivých veličin
- **Výběr** se provádí jako
  - předem daný počet nejlepších veličin (řazení podle váhy)
  - selekce všech s váhou větší než daná mez
  - tvorba modelu s postupně 1..*k* vstupními veličinami (veličiny se přidávají se podle váhy)
- Existují algoritmy, které takto vytvořené váhy využijí při tvorbě modelu

Malá rekapitulace

- AUC, Relief
  - *ReliefF je verze pro klasifikaci do více tříd*
- Information Gain, Gini Index, Chi-square
  - *Mutual Information – ekvivalentní InfoGain*
  - *Symetrical Uncertainty – normovaná Mutual Information*
- t-test, FDR, Scatter Matrix

…a na čem to všechno může selhat…

Příklad: XOR Data

Relief pořadí:

x1, x3, x2

Rel(x1) = 0,188

Rel(x2) = -0,220

Rel(x3) = 0,013

AUC pořadí:

x3, x2, x1

AUC(x1) = 0,547

AUC(x2) = 0,562

AUC(x3) = 0,625

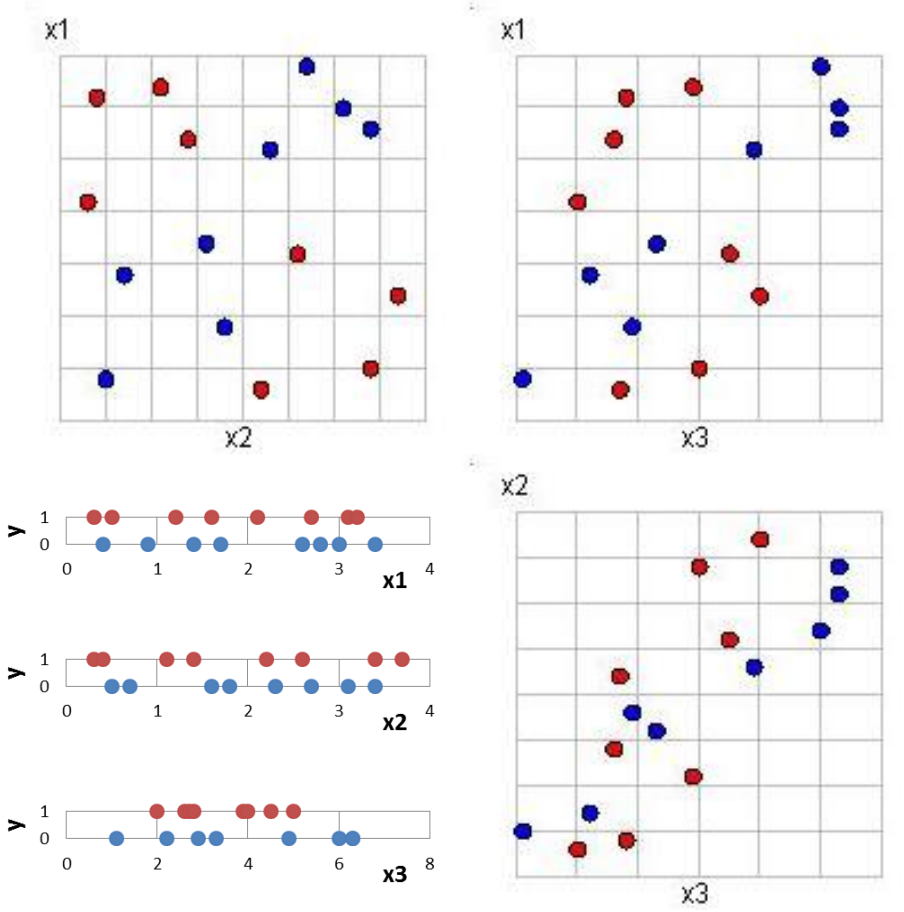
InfoGain pořadí:

x3, x2, x1

IG(x1)= 0,07

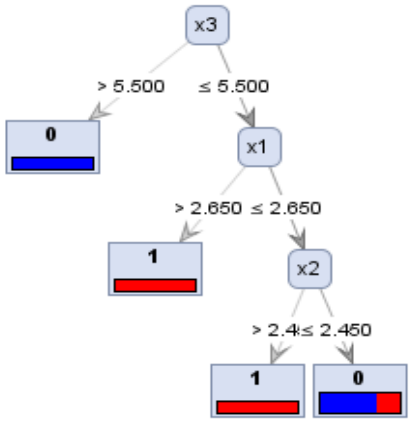
IG(x2)= 0,14

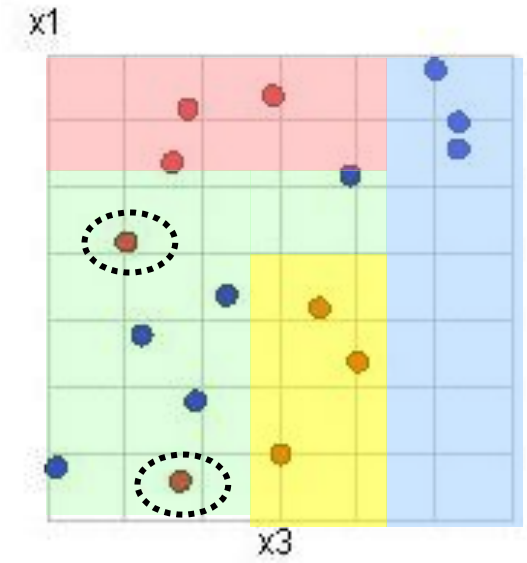
IG(x3)= 0,22

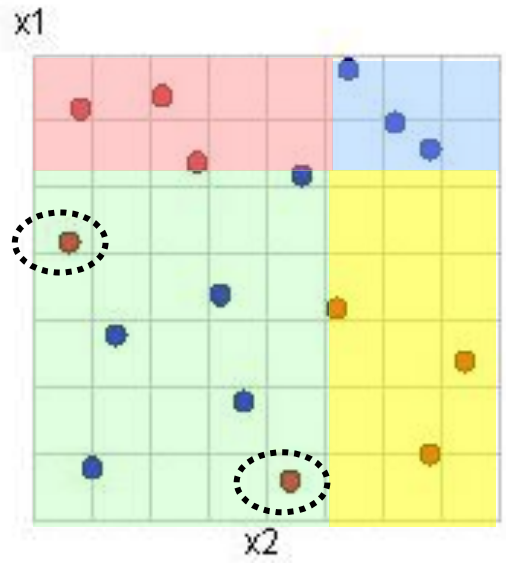


Řešitelné pomocí RS. Jak ale bude takový RS vypadat?

Příklad: XOR Data – řešení pomocí IG

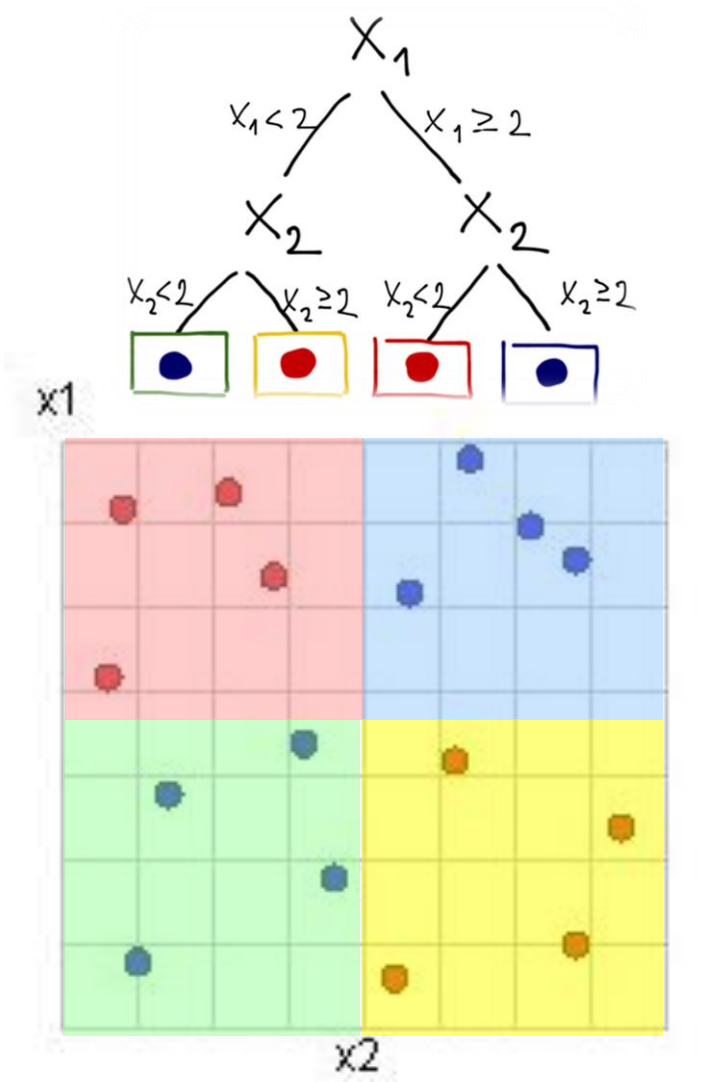






x1	x2	x3	y
3	3,1	6,3	0
2,8	3,4	6,3	0
3,4	2,7	6	0
3,2	1,1	3,9	1
3,1	0,4	2,8	1
2,7	1,4	2,6	1
1,2	3,7	5	1
0,5	3,4	4	1
1,6	2,6	4,5	1
2,6	2,3	4,9	0
0,3	**2,2**	**2,7**	**1**
0,9	1,8	2,9	0
1,7	1,6	3,3	0
1,4	0,7	2,2	0
0,4	0,5	1,1	0
2,1	**0,3**	**2**	**1**

Příklad: XOR Data – optimální řešení



x1	x2	x3	y
0,4	0,5	1,1	0
0,9	1,8	2,9	0
1,4	0,7	2,2	0
1,7	1,6	3,3	0
0,3	2,2	2,7	1
0,5	3,4	4	1
1,2	3,7	5	1
1,6	2,6	4,5	1
2,1	0,3	2	1
2,7	1,4	2,6	1
3,1	0,4	2,8	1
3,2	1,1	3,9	1
2,6	2,3	4,9	0
2,8	3,4	6,3	0
3	3,1	6,3	0
3,4	2,7	6	0

Příklad: XOR Data – závěr

- V úloze **binární klasifikace** byly dány dvě vstupní veličiny x_1 a x_2 rozdělující 2D prostor jako funkce XOR, dále pak veličina x_3
- Z pohledu nesené informace je optimální pracovat s veličinami x_1 a x_2
- Žádná z metod určených pro výběr veličin pomocí řazení však toto **řešení nenalezne**, protože **jednotlivě** se veličiny x_1 a x_2 z pohledu použitelnosti pro klasifikaci jeví jako takřka náhodné (x_3 je jen nepatrně „lepší“, obecně žádnou z veličin jednotlivě nelze doporučit, všechny se jeví jako náhodné; přesto kombinace x_1 a x_2 umožňuje přesné řešení)

Typy výběru veličin

- JEDNOROZMĚRNÉ
  - **Váhování** jednotlivých veličin
    - Gini Index, Information Gain, Chi-Square
    - ROC analýza (AUC), Relief
    - t-Test, scatter matrix
  - Výběr konkrétních veličin

- **VÍCEROZMĚRNÉ**
  - Filter feature selection
    - **kritérium + korelace (křížová, Pearsonova, Spearmanova)**
    - **minimální redundance maximální relevance**
  - Wrappers (používají během selekce skutečný model)
    - **brutal force**
    - **dobředná/zpětná selekce**
    - plovoucí prohledávání, …

Selekce pomocí křížové korelace

- **křížová korelace** ρ_{ij} mezi veličinami x_i a x_j

$$\rho_{ij} = \frac{\sum_{n=1}^{N} x_{ni} x_{nj}}{\sqrt{\sum_{n=1}^{N} x_{ni}^{\,2} \cdot \sum_{n=1}^{N} x_{nj}^{\,2}}}$$

- **výběr veličin**
  - podle kritéria C vyber nejlepší veličinu $C(i_1)$ (např. dle AUC)
  - podle následujícího vztahu k ní vyber do páru další veličinu

$$i_2 = \arg\max_j \left\{ \alpha_1 C(j) - \alpha_2 \left| \rho_{i_1 j} \right| \right\}$$

  - ke dvojici, trojici… vyber další veličinu podle vztahu

$$i_k = \arg\max_j \left\{ \alpha_1 C(j) - \frac{\alpha_2}{k-1} \sum_{r=1}^{k-1} \left| \rho_{i_r j} \right| \right\}$$

kde α_i jsou váhy vyjadřující relativní míru důležitosti

Minimální redundance maximální relevance (mRMR)

- Do RapidMineru lze stáhnout plug-in obsahující metodu mRMR (http://sourceforge.net/projects/rm-featselext/).
- Mějme dvě množiny již vybraných (*S* – selected) a nevybraných (*U* – unselected) vstupních veličin, přičemž S∪U dává všechny vstupní veličiny.
- V iteračním procesu je postupně vybírána veličina $x \in U$ s největší vzájemnou informací (MI - mutual information) snížená o MI mezi *x* a již vybranými veličinami z množiny *S*.

$$\max_{\forall x \in U} \left[MI(x,c) - \frac{1}{|S|-1} \sum_{\forall xi \in S} MI(x, x_i) \right]$$

- Existuje více metod, zejména na bázi teorie informace, pracujících takto s konceptem „relevance – redundance“. Míra vhodnosti *x* se mění v závislosti na veličinách v množině *S*, je tedy v každém kroku jiná.

Brutal force selekce

- je dáno $|A|$ atributů a kritérium kvality predikce $C(A_i,\ldots,A_k)$
- cílem je vybrat nejlepší kombinaci k atributů
- kritériem kvality může být např. nějaký typ modelu společně s metodou odhadu chyby (např. 10fold-Cross-validation)
- počet kombinací bez opakování je dán vztahem:

$$kombinací = \binom{|A|}{k} = \frac{|A|!}{k!(|A|-k)!}$$

- máme-li např. 40 veličin a chceme vybrat nejlepších 10, je zapotřebí provést cca. **9.10^9** výpočtů…
- nalezení optimální kombinace je garantováno

Zpětná (*backward*) selekce

- je dáno $|A|$ atributů a kritérium kvality predikce $C(A_i,\ldots,A_k)$
- kritériem kvality může být např. nějaký typ modelu společně s metodou odhadu chyby (např. 10fold-Cross-validation)
- Vyber *k* veličin:
  - vypočti kritérium C pro všechny atributy $|A|$
  - vypočti kritérium pro všechny kombinace s $|A|$-1 veličinami a vyber tu s největší hodnotu
  - pokračuj odebíráním až zbude *k* veličin
- máme-li např. 40 veličin a chceme vybrat nejlepších 10, je zapotřebí provést pouhých **7.760** výpočtů…
- nalezení optimální kombinace není garantováno

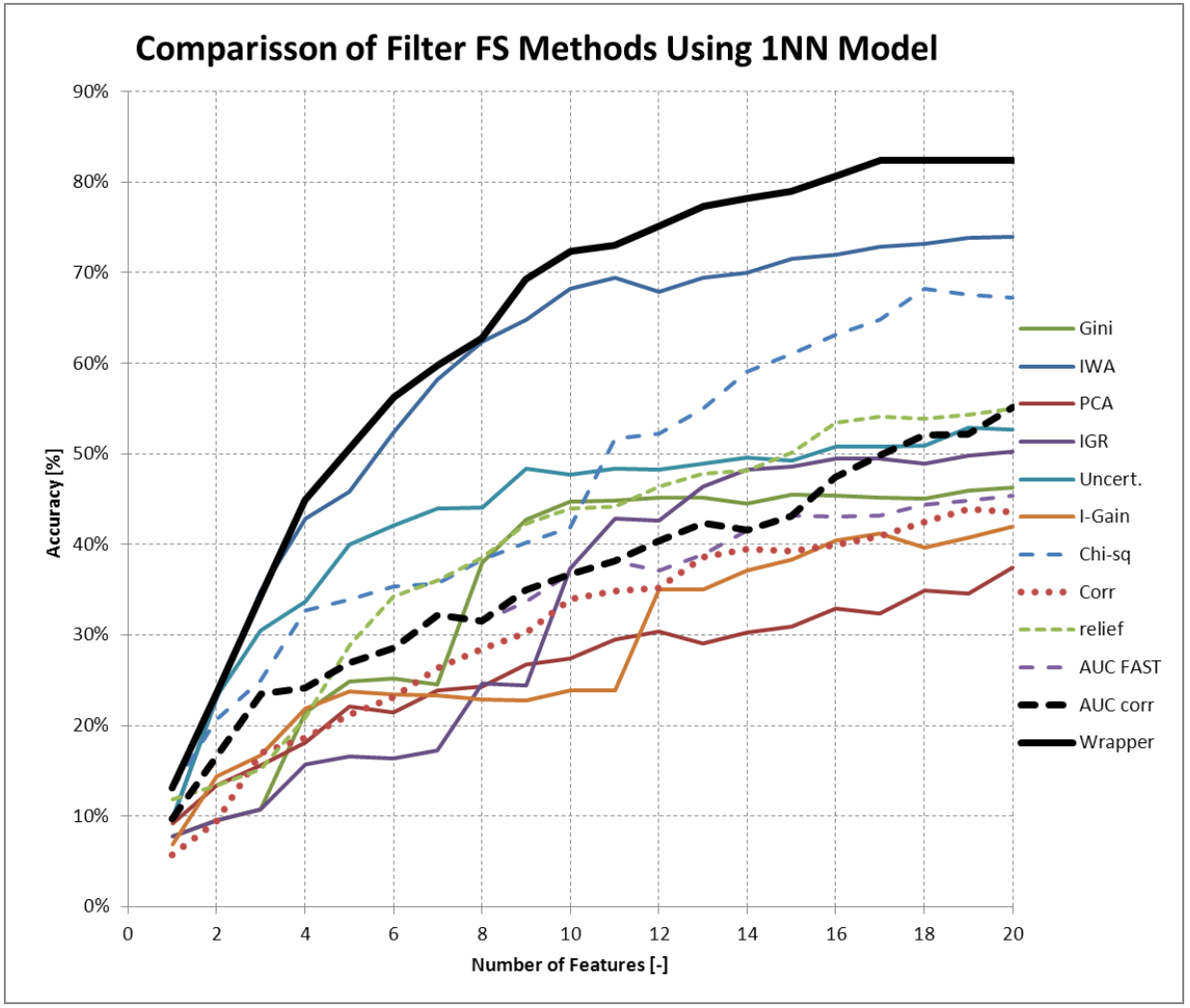
Dopředná (*forward*) selekce

- je dáno $|A|$ atributů a kritérium kvality predikce $C(A_i,\ldots,A_k)$
- kritériem kvality může být např. nějaký typ model společně s metodou odhadu chyby (např. 10fold-Cross-validation)
- Vyber *k* veličin:
  - vypočti kritérium C pro všechny jednotlivé atributy $|A|$ a vyber jeden nejlepší
  - vypočti kritérium *C* pro všechny dvojice tvořené nejlepším atributem z předešlého kroku a jedním dalším, vyber pár s největším *C*
  - pokračuj s přidáváním atributů až bude vybráno *k* veličin
- máme-li např. 40 veličin a chceme vybrat nejlepších 10, je zapotřebí provést pouhých **3.550** výpočtů…
- nalezení optimální kombinace není garantováno

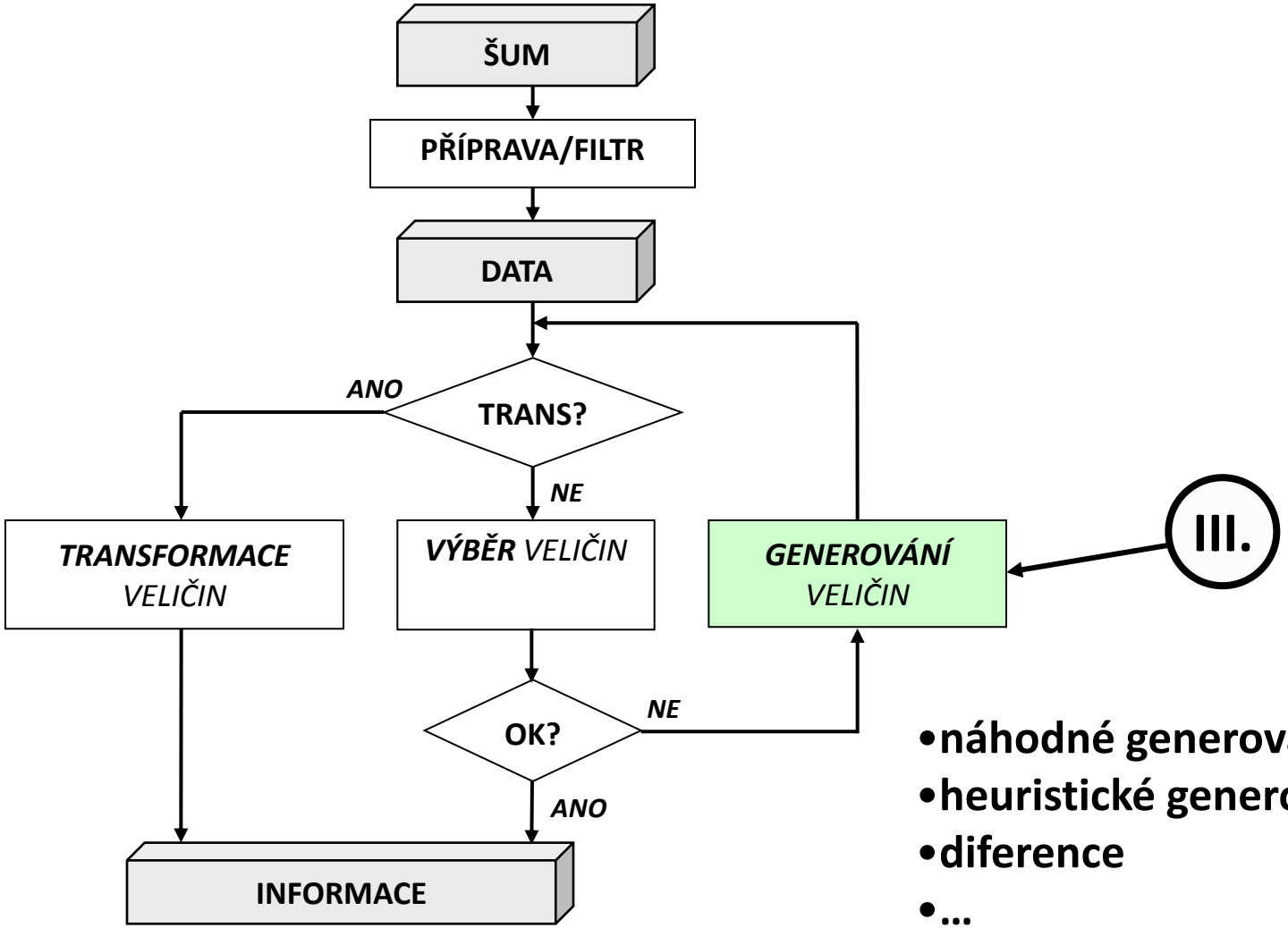
Population based

- použití optimalizačních metod typu genetické algoritmy, rojové algoritmy, atd.
- metody prohledávají celý prostor možných řešení
- jedincem je množina příznaků, kvalitou pak odhadnutá přesnost na konkrétním modelu
- **Výhoda**: v prohledávaném prostoru je i optimální řešení (což neplatí u greedy seach metod)
- **Nevýhoda**: časově náročné, stochastické (při opakovaném spuštění různá řešení)

Srovnání různých metod výběru příznaků



III. GENEROVÁNÍ VELIČIN

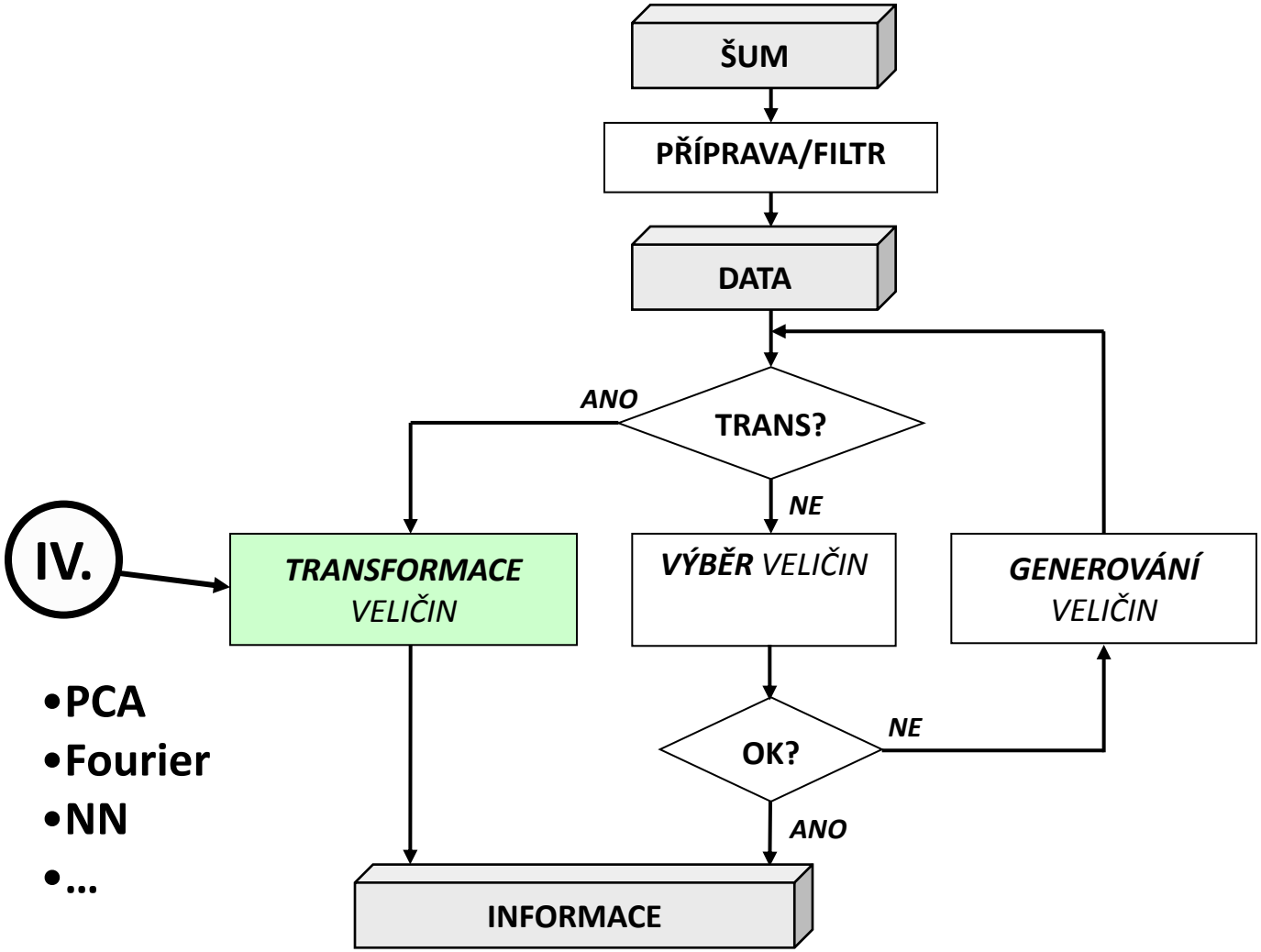


- náhodné generování
- heuristické generování
- diference
- …

Dva hlavní přístupy

- **Náhodné**
  - Lineární kombinace veličin
  - Funkční kombinace veličin (sin, cos, $[b_1 x_1 \hat{} x_2]^{-b2}$,…)
- **Řízené**
  - Pomocí heuristického prohledávání (např. pomocí genetických algoritmů)
  - v procesu generování je jako optimalizační kritérium volena některá z metod používaných pro stanovení váhy atributu (viz II. Výběr – např. AUC, …)

IV. TRANSFORMACE



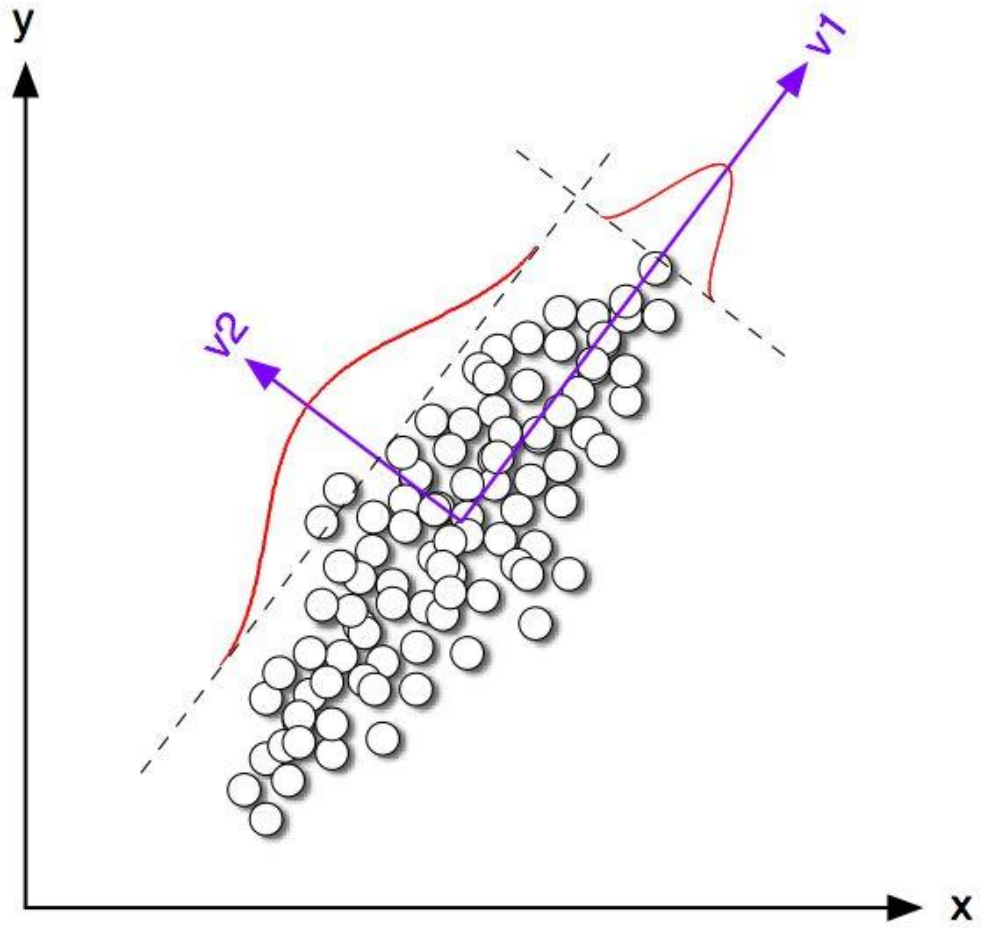
Některé transformace

- **PCA – principal component analysis**
- **Autoasociativní neuronové sítě**
- Selforganizing maps (SOM)
- Singular Value Decomposition
- Fourierova transformace
- …

Principal Component Analysis (PCA)

- česky *analýza hlavních komponent*
- nové **nekorelované veličiny** lineární kombinací veličin stávajících
- cílem je vytvořit **nový ortogonální souřadný systém** umožňující těsnější „box“ kolem bodů
- **učení bez učitele** (transformace vstupních veličin nezávisle na výstupní veličině)
- z pohledu SU je podstatné **pořadí nových os** (komponent)
- osy jsou vytvářeny tak, aby vysvětlovaly co největší podíl z celkového rozptylu (**suma rozptylů je konstantní**, tedy stejná v původním i novém souřadném systému)
- 3 nové osy (z původních 10) mohou vysvětlovat např. 90% rozptylu (a potenciálně tak nesou z pohledu separability většinu informace)

PCA – nový souřadný systém



PCA – rozlišení vína (1/3)

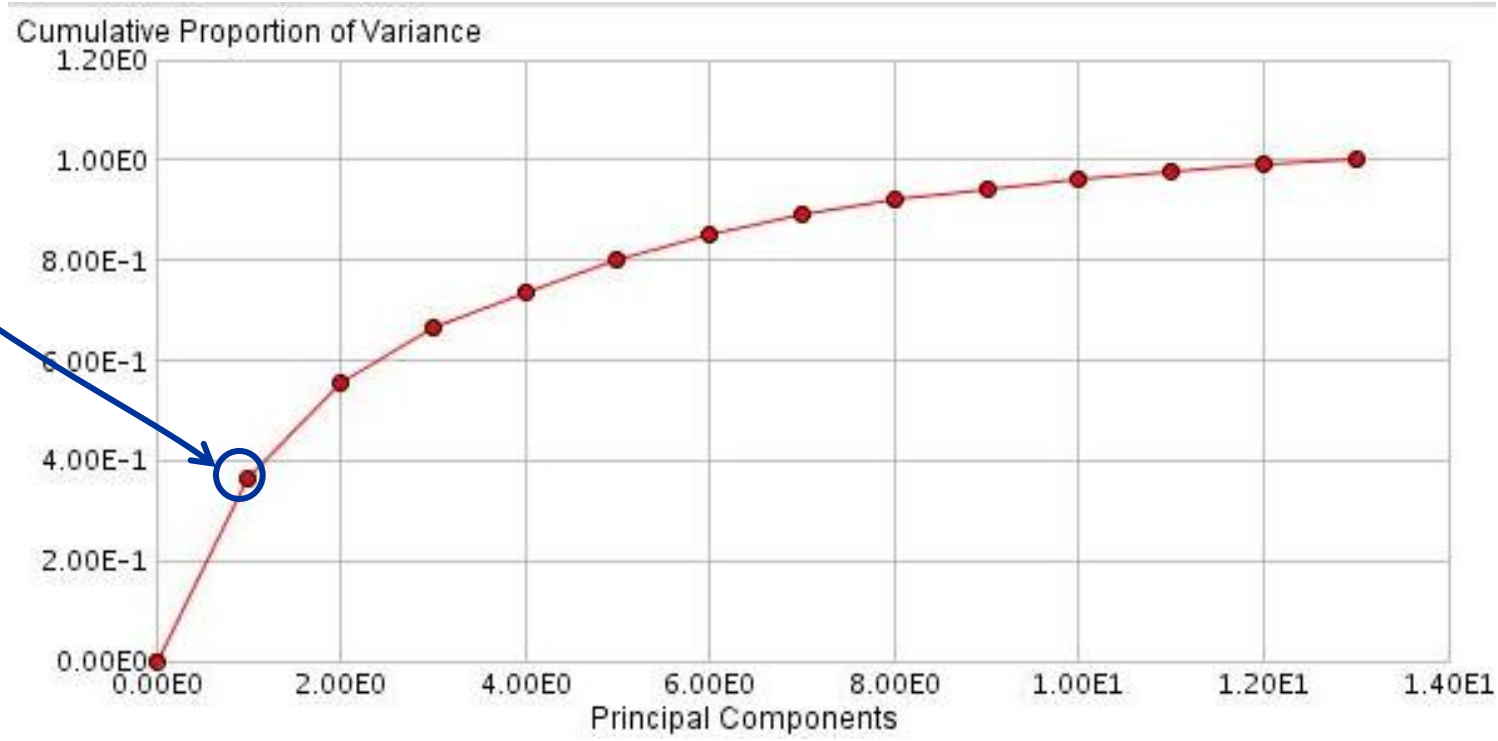
- Cílem je rozlišit tři druhy vína. Vstupní veličiny jsou obsah alkoholu, kyselost, …, celkem 13 veličin. K dispozici je 178 záznamů.

Wine type	Alcohol	Acid	Ash	Alcalinity	Mg	Phenol	Flavanoids	Nonflav phenol	Cyanin	Color intensity	Hue	OD280	Proline
1	14,23	1,71	2,43	15,60	127	2,80	3,06	0,28	2,29	5,64	1,04	3,92	1065
1	13,20	1,78	2,14	11,20	100	2,65	2,76	0,26	1,28	4,38	1,05	3,40	1050
1	13,16	2,36	2,67	18,60	101	2,80	3,24	0,30	2,81	5,68	1,03	3,17	1185
1	14,37	1,95	2,50	16,80	113	3,85	3,49	0,24	2,18	7,80	0,86	3,45	1480
1	13,24	2,59	2,87	21	118	2,80	2,69	0,39	1,82	4,32	1,04	2,93	735
1	14,20	1,76	2,45	15,20	112	3,27	3,39	0,34	1,97	6,75	1,05	2,85	1450
1	14,39	1,87	2,45	14,60	96	2,50	2,52	0,30	1,98	5,25	1,02	3,58	1290
1	14,06	2,15	2,61	17,60	121	2,60	2,51	0,31	1,25	5,05	1,06	3,58	1295
1	14,83	1,64	2,17	14	97	2,80	2,98	0,29	1,98	5,20	1,08	2,85	1045
1	13,86	1,35	2,27	16	98	2,98	3,15	0,22	1,85	7,22	1,01	3,55	1045
1	14,10	2,16	2,30	18	105	2,95	3,32	0,22	2,38	5,75	1,25	3,17	1510
1	14,12	1,48	2,32	16,80	95	2,20	2,43	0,26	1,57	5	1,17	2,82	1280
1	13,75	1,73	2,41	16	89	2,60	2,76	0,29	1,81	5,60	1,15	2,90	1320
1	14,75	1,73	2,39	11,40	91	3,10	3,69	0,43	2,81	5,40	1,25	2,73	1150
1	14,38	1,87	2,38	12	102	3,30	3,64	0,29	2,96	7,50	1,20	3	1547
1	13,63	1,81	2,70	17,20	112	2,85	2,91	0,30	1,46	7,30	1,28	2,88	1310
1	14,30	1,92	2,72	20	120	2,80	3,14	0,33	1,97	6,20	1,07	2,65	1280
1	13,83	1,57	2,62	20	115	2,95	3,40	0,40	1,72	6,60	1,13	2,57	1130
1	14,19	1,59	2,48	16,50	108	3,30	3,93	0,32	1,86	8,70	1,23	2,82	1680

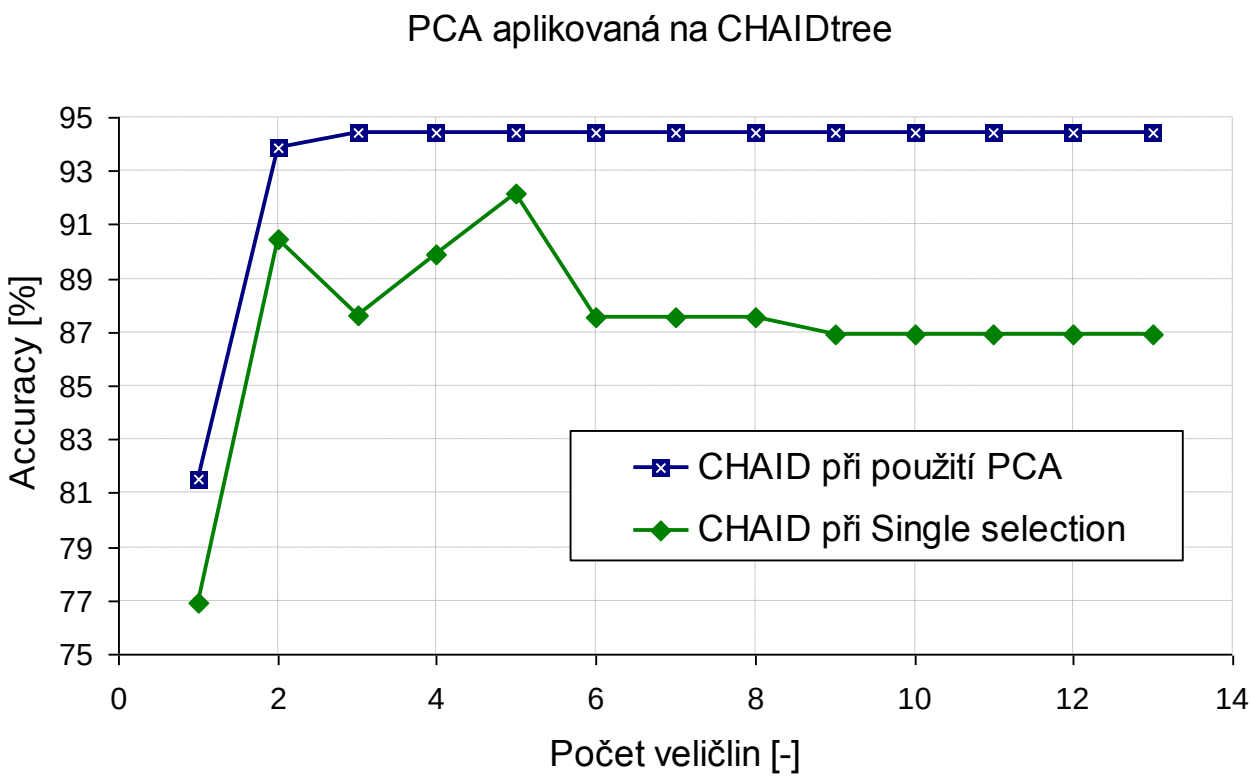
PCA – rozlišení vína (2/3)

PC 1=

0.144 * wine.dat (2) + 0.484 * wine.dat (3) - 0.207 * wine.dat (4) + 0.018 * wine.dat (5) + 0.266 * wine.dat (6) - 0.214 * wine.dat (7) + 0.056 * wine.dat (8) - 0.396 * wine.dat (9) - 0.509 * wine.dat (10) + 0.212 * wine.dat (11) + 0.226 * wine.dat (12) + 0.266 * wine.dat (13) + 0.015 * wine.dat (14)



PCA – rozlišení vína (3/3)



- z grafu je patrné, že k maximálnímu rozlišení stačí použít pouze 3 veličiny PCA oproti postupu *single selection.*

PCA – ionosféra (1/3)

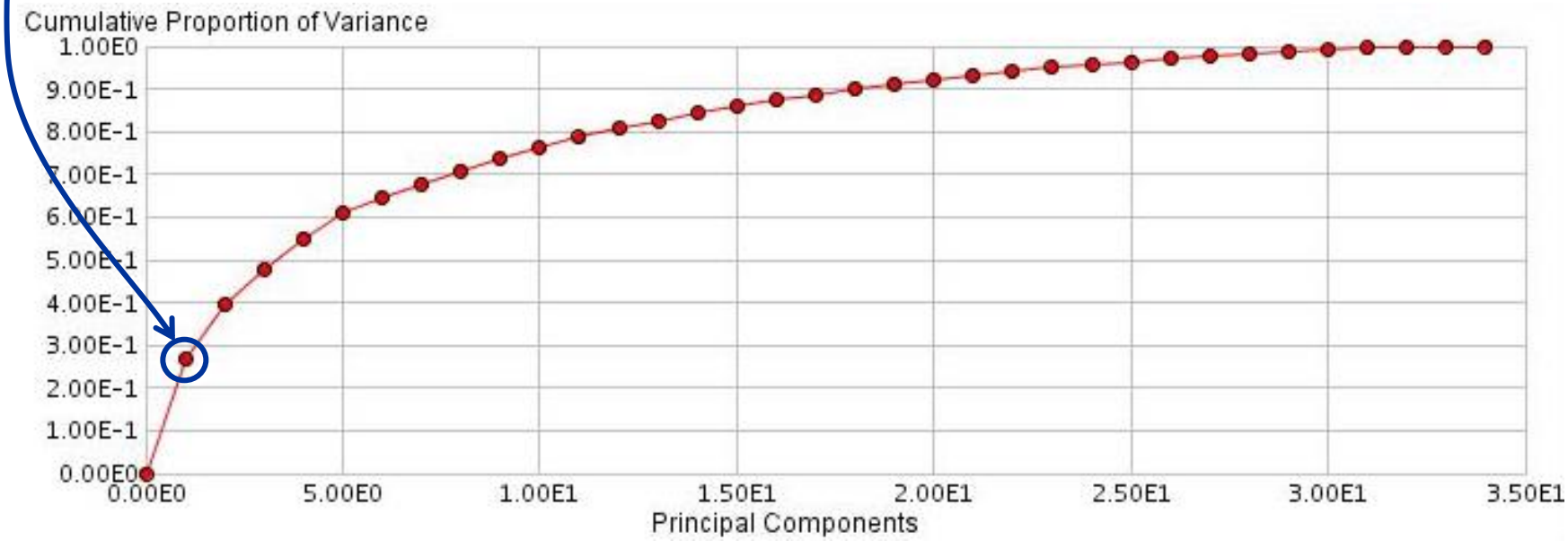
- Jsou dána vstupní data, 351 záznamů, 34 spojitých veličin (detekce odražených rádiových vln), klasifikujeme do dvou tříd. Cílem je vyhodnotit, jestli pro klasifikaci stačí pouze vhodně vybrané a normalizované veličiny, nebo je výhodné provést PCA.

Rx9	Ix9	Rx10	Ix10	Rx11	Ix11	Rx12	Ix12	Rx13	Ix13	Rx14	Ix14	Rx15	Ix15	Rx16	Ix16	Rx17	Ix17	C
0.84356	-0.38542	0.58212	-0.32192	0.56971	-0.29674	0.36946	-0.47357	0.56811	-0.51171	0.41078	-0.46168	0.21266	-0.3409	0.42267	-0.54487	0.18641	-0.453	g
0.05499	-0.62237	0.33109	-1	-0.13151	-0.453	-0.18056	-0.35734	-0.20332	-0.26569	-0.20468	-0.18401	-0.1904	-0.11593	-0.16626	-0.06288	-0.13738	-0.02447	b
0.83775	-0.13644	0.75535	-0.0854	0.70887	-0.27502	0.43385	-0.12062	0.57528	-0.4022	0.58984	-0.22145	0.431	-0.17365	0.60436	-0.2418	0.56045	-0.38238	g
0.54094	-0.3933	-1	-0.54467	-0.69975	1	0	0	1	0.90695	0.51613	1	1	-0.20099	0.25682	1	-0.32382	1	b
0.5294	-0.2178	0.45107	-0.17813	0.05982	-0.35575	0.02309	-0.52879	0.03286	-0.65158	0.1329	-0.53206	0.02431	-0.62197	-0.05707	-0.59573	-0.04608	-0.65697	g
-0.00343	-0.10196	-0.11575	-0.05414	0.01838	0.03669	0.01519	0.00888	0.03513	-0.01535	-0.0324	0.09223	-0.07859	0.00732	0	0	-0.00039	0.12011	b
0.57945	-0.68086	0.37852	-0.73641	0.36324	-0.76562	0.31898	-0.79753	0.22792	-0.81634	0.13659	-0.8251	0.04606	-0.82395	-0.04262	-0.81318	-0.13832	-0.80975	g
1	1	-1	-1	0	0	0	0	1	1	1	1	0	0	1	1	0	0	b
0.98305	-0.35257	0.84537	-0.6602	0.75346	-0.60589	0.69637	-0.64225	0.85106	-0.6544	0.57577	-0.69712	0.25435	-0.63919	0.45114	-0.72779	0.38895	-0.7342	g
-0.37133	0.15018	0.63728	0.22115	0	0	0	0	-0.14803	-0.01326	0.20645	-0.02294	0	0	0.16595	0.24086	-0.08208	0.38065	b
0.8923	-0.66474	0.69876	-0.70997	0.70645	-0.7632	0.63081	-0.80544	0.55867	-0.89128	0.47211	-0.865	0.40303	-0.83675	0.30996	-0.89093	0.22995	-0.89158	g
1	-0.29354	1	-0.93599	1	1	1	1	1	-0.40888	1	-0.62745	1	-1	1	-1	1	-1	b
0.94486	-0.28106	0.90137	-0.43383	0.86043	-0.47308	0.82987	-0.5122	0.8408	-0.47137	0.76224	-0.5837	0.65723	-0.68794	0.68714	-0.64537	0.64727	-0.67226	g
1	-1	1	-1	0.61831	0.15803	1	0.62349	1	-0.17012	1	0.35924	1	-0.66494	1	0.88428	1	-0.18826	b
1	0.10561	1	0.27087	1	0.44758	1	0.4175	1	0.20033	1	0.36743	0.95603	0.48641	1	0.32492	1	0.46712	g
-0.37681	0.03623	1	-1	0	0	0	0	-0.16253	0.92236	0.39752	0.26501	0	0	1	0.23188	0	0	b
0.88809	0.1112	0.86104	0.08631	0.81633	0.1183	0.83668	0.14442	0.81329	0.13412	0.79476	0.13638	0.7911	0.15379	0.77122	0.1593	0.70941	0.12015	g
-1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	b
0.58373	0.18151	0.14395	0.41224	0.53888	0.21326	0.5142	0.22625	0.48838	0.23724	0.46167	0.24618	0.43433	0.25306	0.40663	0.25792	1	0.33036	g
-1	1	1	1	-1	1	1	0.5625	-1	1	1	1	1	-1	1	1	1	1	b
0.92124	-0.31884	0.86473	-0.34534	0.91693	-0.44072	0.9606	-0.46866	0.81874	-0.40372	0.82681	-0.42231	0.75784	-0.38231	0.80448	-0.40575	0.74354	-0.45039	g
-1	-1	-1	1	-1	1	0	0	0	0	1	-1	-1	1	-1	1	-1	1	b

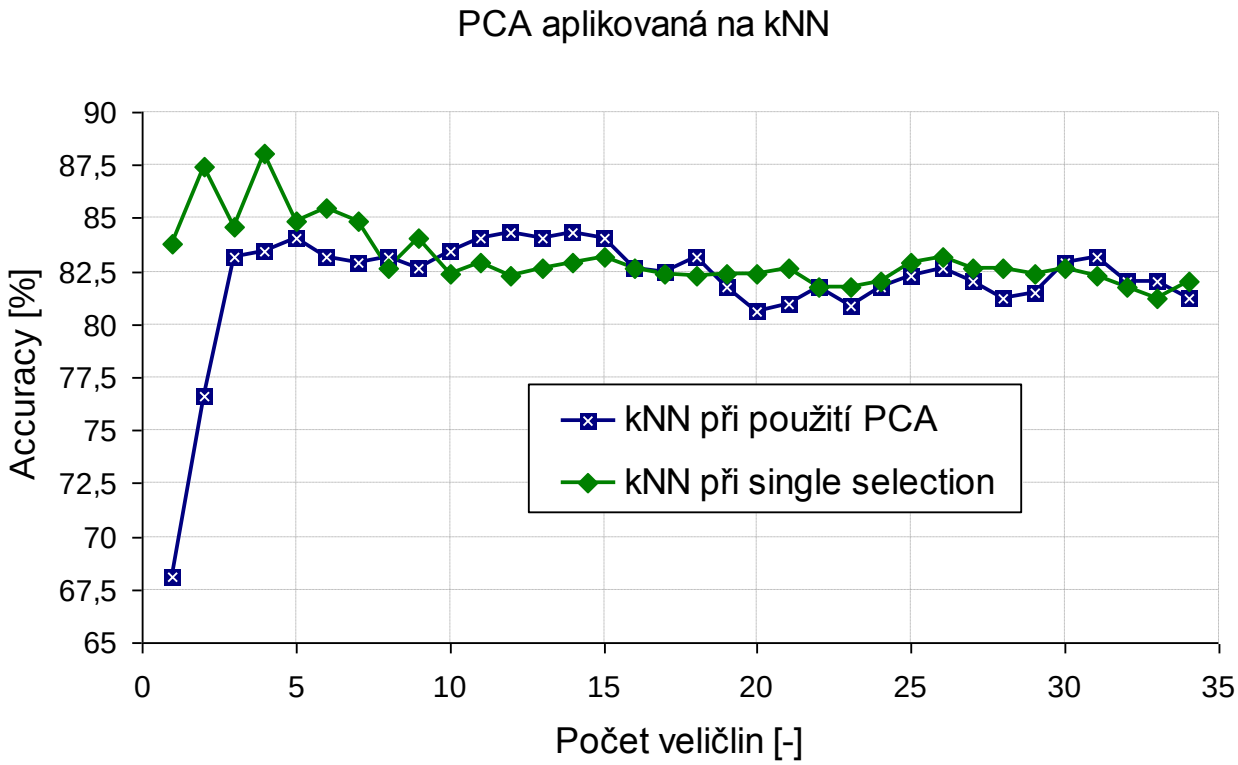
PCA – ionosféra (2/3)

PC 1=

- 0.045 * Rx1 - 0.032 * Ix1 - 0.261 * Rx2 - 0.102 * Ix2 + 0.218 * Rx3 + 0.425 * Ix3 - 0.451 * Rx4 + 0.163 * Ix4 + 0.084 * Rx5 - 0.153 * Ix5 - 0.308 * Rx6 - 0.181 * Ix6 + 0.063 * Rx7 + 0.096 * Ix7 + 0.096 * Rx8 + 0.038 * Ix8 - 0.024 * Rx9 - 0.032 * Ix9 + 0.294 * Rx10 + 0.046 * Ix10 + 0.046 * Rx11 + 0.022 * Ix11 + 0.197 * Rx12 + 0.206 * Ix12 + 0.121 * Rx13 - 0.043 * Ix13 + 0.168 * Rx14 + 0.157 * Ix14 + 0.021 * Rx15 + 0.153 * Ix15 - 0.058 * Rx16 - 0.045 * Ix16 + 0.075 * Rx17 - 0.000 * Ix17

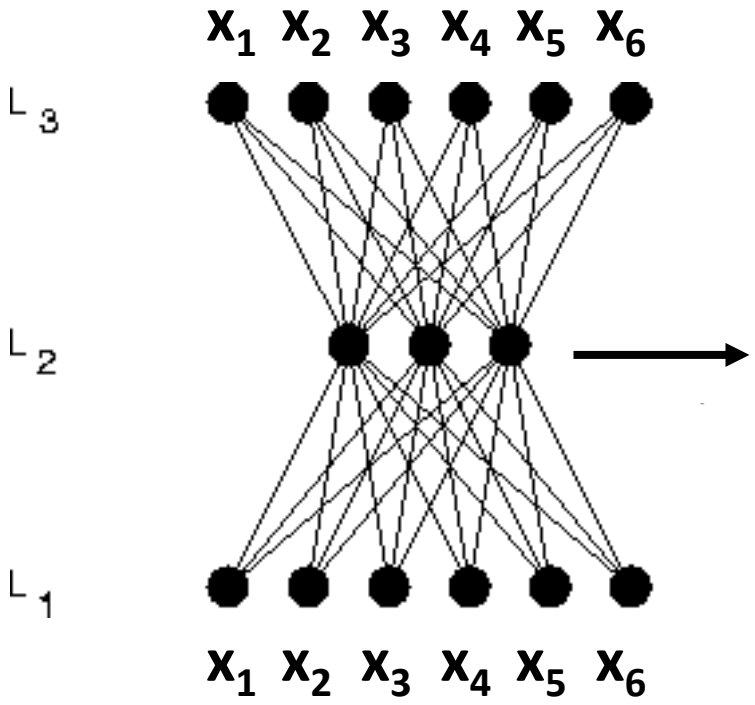


PCA – ionosféra (3/3)



- z grafu je patrné, že model vytvořený pomocí PCA představuje mírné zhoršení kvality predikce, lineární kombinace veličin nepřinesla zlepšení

Transformace autoasociativní sítí



Nové atributy A_1 A_2 A_3

www.bbsonline.org

Doporučená literatura

[1] Honzík, P.: *Strojové učení*, elektronická skripta VUT.

[2] Theodoridis, S. et.al.: *Pattern Recognition*, Elsevier 2003.

[3] Alpaydin, E.: *Introduction to Machine Learning*, MIT Press 2004.

[4] Hastie T.et.al.: *The Elements of Statistical Learning*. Springer, 2001.

[5] …nepřeberné množství materiálů na internetu…