EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Název projektu	Rozvoj vzdělávání na Slezské univerzitě v Opavě
Registrační číslo projektu	CZ.02.2.69/0.0./0.0/16_015/0002400

Dolování dat

Vyhodnocení výsledků – 2. část

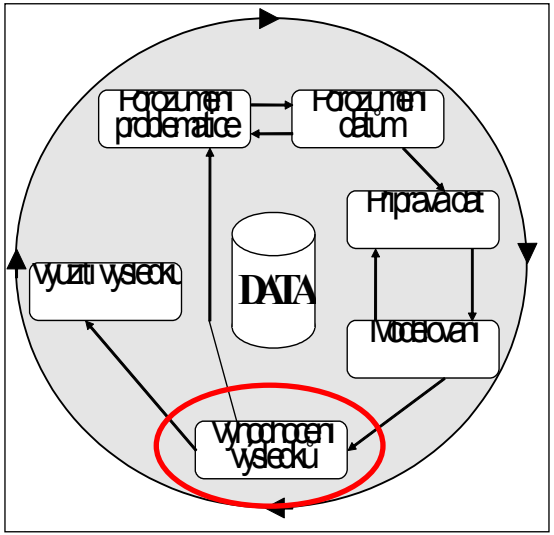
Jan Górecki

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Klasifikační úlohy
- Hodnocení v podobě grafů
- Regresní úlohy
- Vizualizace
- Volba nejvhodnějšího algoritmu



Klasifikační úlohy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- kritériem **úspěšnost klasifikace (predikce)** na datech

Cílem je zjistit v kolika případech došlo ke **shodě** resp. **neshodě** modelu (systému) s informací od učitele

Testování modelů

- testování na celých trénovacích datech
- náhodné rozdělení na část trénovací a testovací
- křížová validace (cross-validation)
- leave-one-out
- bootstrap (náhodný výběr s opakováním pro učení)
- testování na testovacích datech

Matice záměn (Confusion matrix)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Naivní Bayes	
Skutečnost	**Predikce**
ano	ano
ano	ano
ne	ano
ano	ano
ano	ano
ne	ne
ano	ano
ano	ano
ne	ne
ano	ano
ne	ne
ano	ano

Skutečnost \ **Predikce**	ano	ne
ano	TP	FN
ne	FP	TN

Skutečnost \ **Predikce**	ano	ne
ano	8	0
ne	1	3

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} = \frac{8+3}{8+3+1+0} = \frac{11}{12} = 0{,}917$$

Celková správnost (Acc) **nebere** v potaz:

1. Různé ceny chyb FP a FN (lze řešit maticí cen)
2. Míru přesvědčení klasifikátoru o správnosti klasifikace

Hodnocení v podobě grafů

- Většina klasifikačních metod dává k dispozici, kromě samotné klasifikace, i tzv. **skóre** (score) klasifikace, tz. jak moc je model „přesvědčen“ o tom, že daný objekt patří do dané třídy => lze taktéž využít pro další hodnocení modelů

	Naivní Bayes		
Skutečnost	**Predikce**	**Skóre(ano)**	**Skóre(ne)**
ano	ano	0,98	0,02
ano	ano	0,98	0,02
ne	ano	0,52	0,48
ano	ano	0,65	0,35
ano	ano	0,65	0,35
ne	ne	0,27	0,73
ano	ano	0,89	0,11
ano	ano	0,73	0,27
ne	ne	0,27	0,73
ano	ano	0,89	0,11
ne	ne	0,27	0,73
ano	ano	0,52	0,48

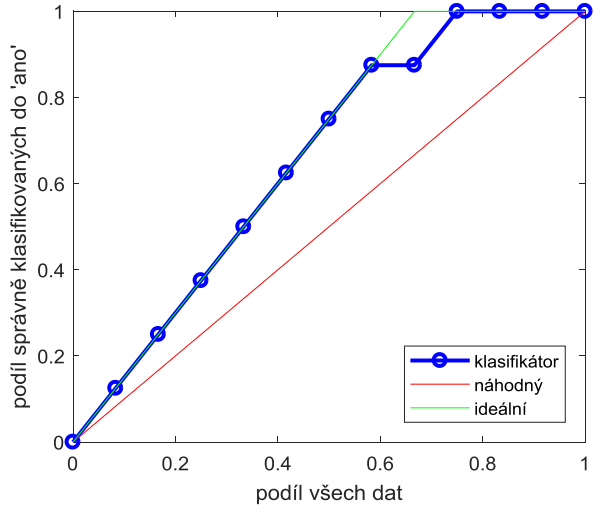
**Skutečné zařazení našich 12ti bankovních klientů do třídy,
klasifikace dle Naivního Bayese a skóre klasifikace pro jednotlivé třídy**

Hodnocení v podobě grafů

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Křivka navýšení (lift curve)

	Naivní Bayes		
Skutečnost	Predikce	Skóre(ano)	Skóre(ne)
ano	ano	0,98	0,02
ano	ano	0,98	0,02
ano	ano	0,89	0,11
ano	ano	0,89	0,11
ano	ano	0,73	0,27
ano	ano	0,65	0,35
ano	ano	0,65	0,35
ne	ano	0,52	0,48
ano	ano	0,52	0,48
ne	ne	0,27	0,73
ne	ne	0,27	0,73
ne	ne	0,27	0,73



Tečkovaná čára – výkon „modelu“, když pošleme dopisy náhodně
Plná čára – výkon modelu, když pošleme dopisy adresátům s nejvyšším skoré (dle modelu)

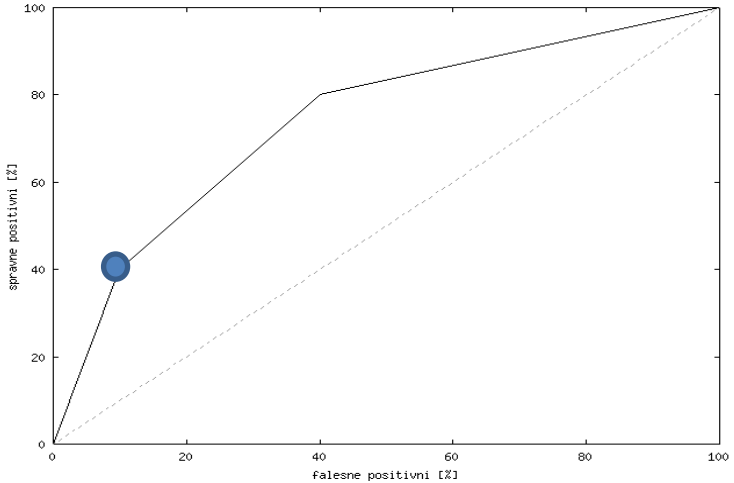
Tvorba křivky se **neřídí** predikovanou třídou!
Řídí se pouze podle skóre a skutečné třídy!

Vztah mezi počtem úspěšných klasifikací a váhou klasifikace

Hodnocení v podobě grafů

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Křivka ROC



Vztah mezi TP a FP pro různá nastavení klasifikátoru

Interpretace: Každý bod křivky říká, kolik falešně pozitivních objektů budu mít, pokud budu chtít mít určité procento skutečně pozitivních objektů (modrý bod - 10% falešných při 40% skutečných)

$TP_{\%}$ = Senzitivita (Úplnost),
$1 - FP_{\%}$ = Specificita

$$TP_{\%} = \frac{TP}{TP + FN} \qquad FP_{\%} = \frac{FP}{FP + TN}$$

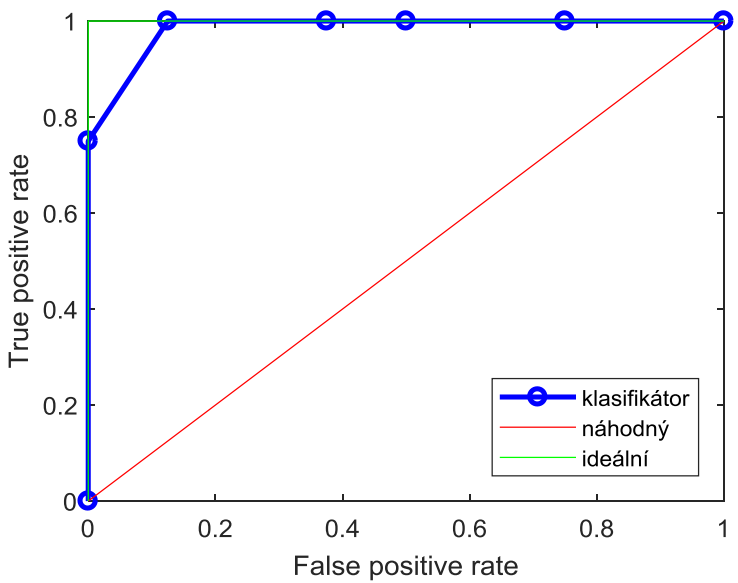
	Predikce	
Skutečnost	ano	ne
ano	TP	FN
ne	FP	TN

Tvorba křivky se **neřídí** predikovanou třídou!
Řídí se pouze podle skóre a skutečné třídy!

Vytvoření ROC křivky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

	Naivní Bayes		
Skutečnost	Predikce	Skóre(ano)	Skóre(ne)
ne	ne	0,27	0,73
ne	ne	0,27	0,73
ne	ne	0,27	0,73
ne	ano	0,52	0,48
ano	ano	0,52	0,48
ano	ano	0,65	0,35
ano	ano	0,65	0,35
ano	ano	0,73	0,27
ano	ano	0,89	0,11
ano	ano	0,89	0,11
ano	ano	0,98	0,02
ano	ano	0,98	0,02

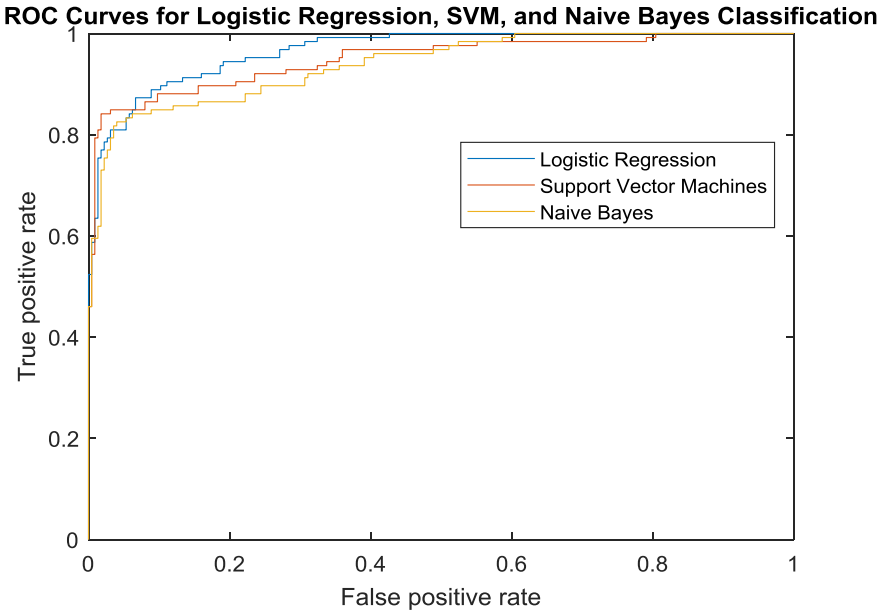


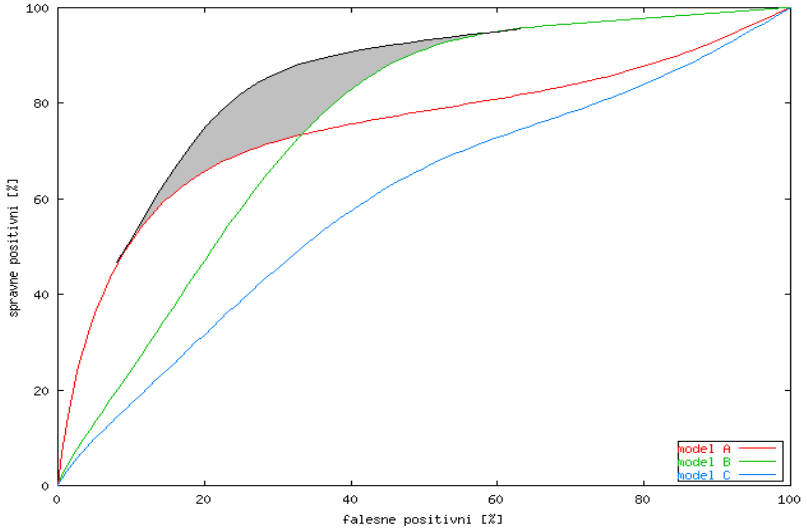
Tvorba křivky se **neřídí** predikovanou třídou!
Řídí se pouze podle skóre a skutečné třídy!

AUC = 0.9844
AUC = Area Under Curve (obsah plochy pod křivkou)

ROC křivky pro více modelů

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ





Do „šedé zóny“ se lze dostat **kombinací modelů**, např. *bagging* nebo *boosting*

Numerické predikce

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

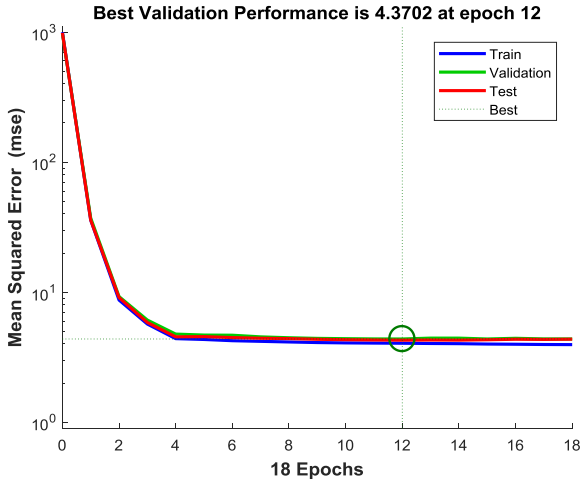
$$MSE = \frac{(p_1-s_1)^2 + \ldots + (p_n-s_n)^2}{n}$$

$$RMSE = \sqrt{\frac{(p_1-s_1)^2 + \ldots + (p_n-s_n)^2}{n}}$$

$$MAE = \frac{|p_1-s_1| + \ldots + |p_n-s_n|}{n}$$

$$RSE = \frac{(p_1 - s_1)^2 + \ldots + (p_n - s_n)^2}{(s_1 - \bar{s})^2 + \ldots + (s_n - \bar{s})^2}, \quad \text{kde} \quad \bar{s} = \frac{1}{n}\sum_i s_i$$

$$\rho = \frac{S_{ps}}{\sqrt{S_p^2\, S_s^2}}, \quad \text{kde } S_{ps} = \frac{\sum_i (p_i - \bar{p})(s_i - \bar{s})}{n-1}, \quad S_p^2 = \frac{\sum_i (p_i - \bar{p})^2}{n-1}, \quad S_s^2 = \frac{\sum_i (s_i - \bar{s})^2}{n-1}$$

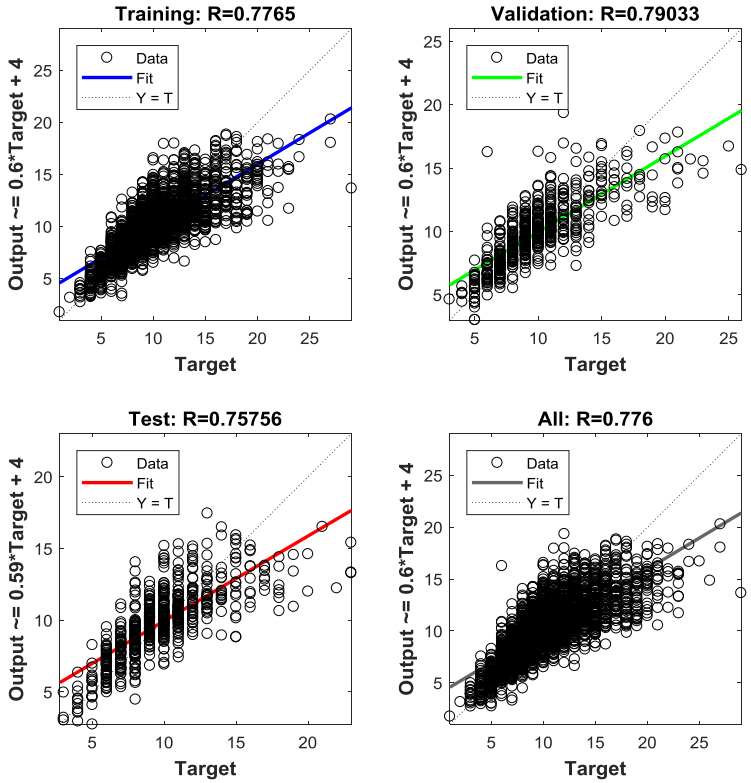


- p_i predikovaná hodnota
- s_i skutečná hodnota

Obě pro *i*-tý příklad ze soubor *n* příkladů tvořících trénovací data

Hodnocení pro numerický výstup – regresi

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



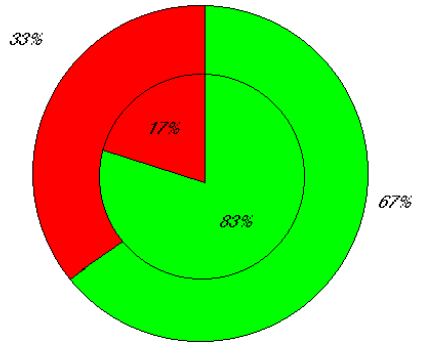
Analogie k matici záměn.

Vizualizace - pravidla

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

IF nezamestnany(ne) THEN uver(ano)

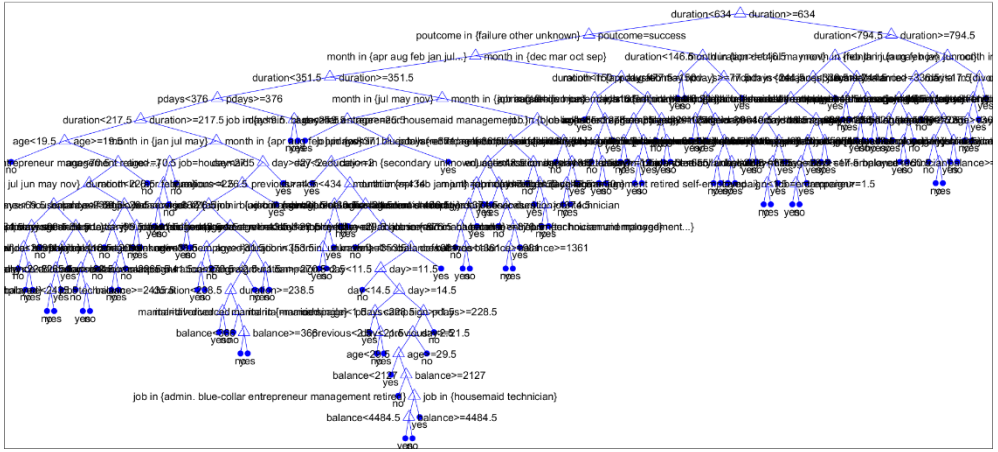
	uver(ano)	uver(ne)	
nezamestnany(ne)	5	1	6
nezamestnany (ano)	3	3	6
			12



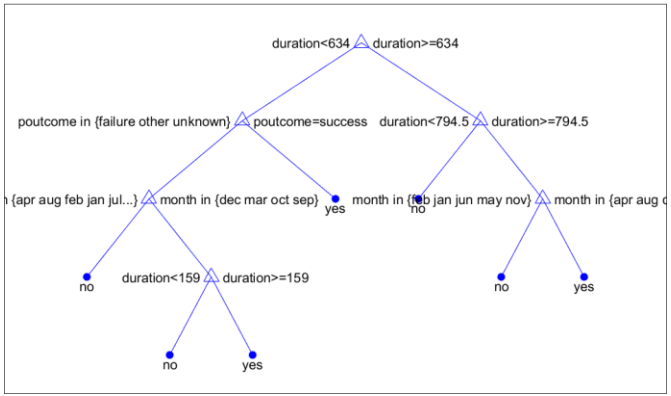
Uvedený graf názorněji ukazuje, že platnost pravidla 5/6 je větší než relativní četnost třídy *úvěr(ano)* v datech (ta je rovna 8/12), a že tedy toto pravidlo dobře charakterizuje bonitní klienty.

Vizualizace – Rozhodovací stromy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



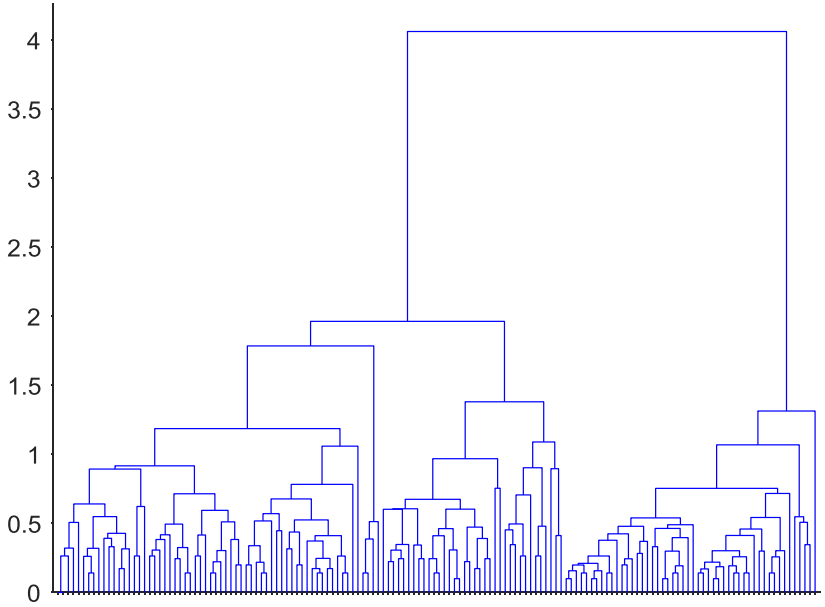
Přetrénovaný klasifikační strom



Optimalizovaný klasifikační strom

Vizualizace – Shlukování

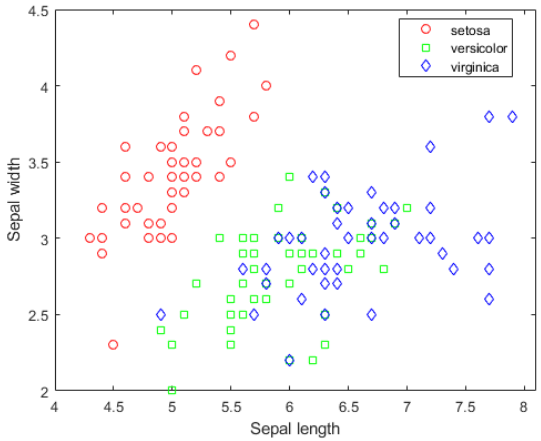
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

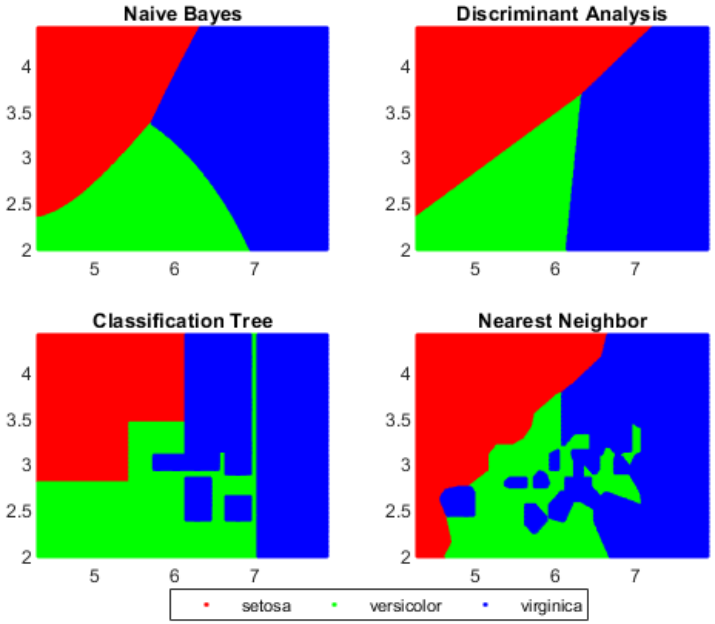


Dendrogram pro Fisher Iris data

Vizualizace – Rozhodovací povrchy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ





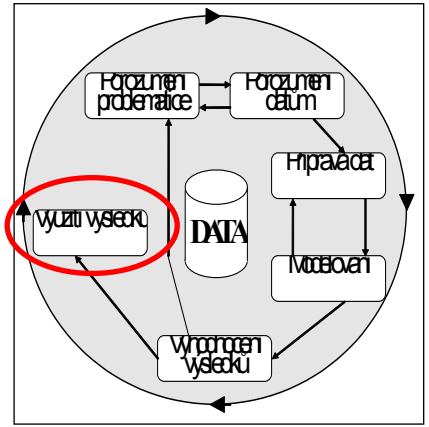
Jen pro dvou- či maximálně tří-rozměrná numerická data!

Volba nejvhodnějšího algoritmu

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

No Free Lunch Theorem:
„Neexistuje metoda, která by byla nejlepší na libovolných datech“

- charakteristiky algoritmů vs. charakteristiky dat
  - vyjadřovací síla,
  - schopnost práce s numerickými atributy,
  - schopnost práce se zašuměnými a chybějícími daty,
  - schopnost práce s maticí cen,
  - předpoklad nezávislosti mezi atributy,
  - ostrá vs. neostrá klasifikace



Děkuji za pozornost

Některé snímky převzaty od:

prof. Ing. Petr Berka, CSc. berka@vse.cz