



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 – 1. část

Jan Górecki

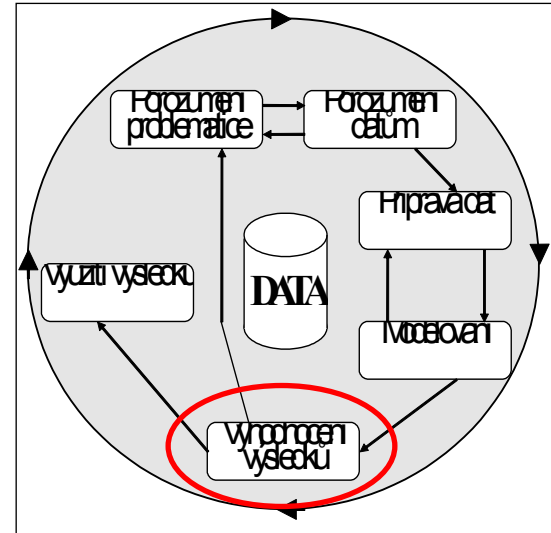


**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

Obsah přednášky



- Deskriptivní úlohy
- Klasifikační úlohy
- Hodnocení jedním/dvěma čísly



- kritériem **novost, zajímavost, užitečnost a srozumitelnost**

Expert = odborník na danou oblast, např. lékař nebo bankéř
(nemusí vědět nic o Dolování dat)

Kvalitativní hodnocení

- zřejmé znalosti, které jsou ve shodě se „zdravým selským rozumem“
- zřejmé znalosti, které jsou ve shodě se znalostmi experta z dané oblasti
- nové, zajímavé znalosti, které přinášejí nový pohled
- znalosti, které musí expert podrobit bližší analýze, neboť není zcela jasné co znamenají
- „znalosti“, které jsou v rozporu se znalostmi experta

Kvantitativní hodnocení

- Např. spolehlivost a podpora u pravidel

Pozor, ne vše co je statisticky významné je i zajímavé!

- kritériem úspěšnost klasifikace (predikce) na datech

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

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

Matice záměn (Confusion matrix)



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

Matice záměn (Confusion matrix)



Rozhodovací stromy	
Skutečnost	Predikce
ano	ano
ano	ano
ne	ne
ano	ne
ano	ne
ne	ne
ano	ano
ano	ano
ne	ne
ano	ano
ne	ne
ano	ne

Skutečnost	Predikce	
	ano	ne
ano	TP	FN
ne	FP	TN



Skutečnost	Predikce	
	ano	ne
ano	5	3
ne	0	4

Hodnocení jedním/dvěma čísly

Celková správnost resp. celková chyba (overall accuracy a error)

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \text{Err} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Celková správnost $\in [\text{Acc}_{\text{def}}, \text{Acc}_{\text{max}}] \subseteq [0, 1]$, kde

Acc_{def} ... správnost při klasifikaci všech příkladů do majoritní třídy (8/12 pro bankovní klienty)

Acc_{max} ... maximální možná správnost pro daná data (1 pro pro bankovní klienty)

	Predikce	
Skutečnost	ano	ne
ano	TP	FN
ne	FP	TN

Správnost pro jednotlivé třídy

- V případě, že třídy jsou v datech rozloženy výrazně nerovnoměrně (např. pouze 5% klientů banky je podezřelých, zbylých 95% je v pořádku)

$$Acc_{ano} = \frac{TP}{TP+FP} \quad Acc_{ne} = \frac{TN}{TN+FN}$$

Interpretace: Z těch co jsou predikováni jako *ano* (*ne*), kolik jich je skutečně *ano* (*ne*).

Skutečnost	Predikce	
	ano	ne
ano	TP	FN
ne	FP	TN

Skutečnost	Predikce	
	ano	ne
ano	8	0
ne	1	3

Skutečnost	Predikce	
	ano	ne
ano	5	3
ne	0	4

Přesnost a úplnost

- Vyhledávání informací - Přesnost nám říká, kolik nalezených dokumentů se skutečně týká daného tématu a úplnost nám říká, kolik dokumentů týkajících se tématu jsme našli

$$\text{Přesnost} = \frac{TP}{TP + FP} \quad \text{Úplnost} = \frac{TP}{TP + FN}$$

Interpretace úplnosti: Z těch, co jsou *ano*, kolik z nich predikujeme, že jsou *ano*.

Skutečnost	Predikce	
	ano	ne
ano	TP	FN
ne	FP	TN

Skutečnost	Predikce	
	ano	ne
ano	8	0
ne	1	3

Skutečnost	Predikce	
	ano	ne
ano	5	3
ne	0	4

Sensitivita a specificita

Hodnocení kvality testu na nějakou nemoc:

- U kolika nemocných (*ano*) pacientů řekne test, že mají jsou nemocní (*ano*) – **sensitivita**
- U kolika zdravých (*ne*) pacientů řekne test, že mají jsou zdraví (*ne*) - **specificita**

$$\text{Sensitivita} = \frac{TP}{TP + FN} \quad \text{Specificita} = \frac{TN}{TN + FP}$$

Interpretace sensitivity: Úplnost pro třídu *ano*.

Interpretace specificity: Úplnost pro třídu *ne*.

Skutečnost	Predikce	
	ano	ne
ano	TP	FN
ne	FP	TN

Skutečnost	Predikce	
	ano	ne
ano	8	0
ne	1	3

Skutečnost	Predikce	
	ano	ne
ano	5	3
ne	0	4

Jen počet chyb nebo i ceny/náklady a výnosy



Chyba bez ceny

$$\text{Err} = 1 - \text{Acc}$$

Chyba s cenami

$$\text{Err} = \text{FP} * c_{\text{FP}} + \text{FN} * c_{\text{FN}}$$

c_{FP} – cena za chybné zařazení *ne* do *ano*

c_{FN} – cena za chybné zařazení *ano* do *ne*

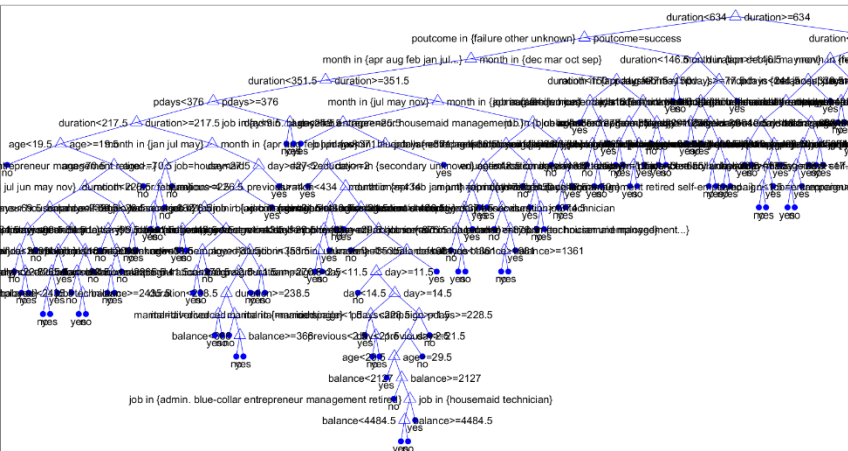
		Predikce	
		ano	ne
Skutečnost	ano	TP	FN
	ne	FP	TN

Příklad

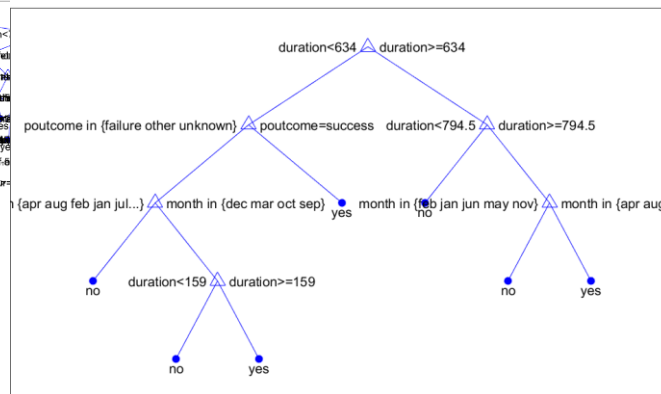


Evaluace tří modelů získaných pro data 4521 klientů portugalské banky
(L:\gorecki\Public\NPDOD-NKDOD\Data\bank.csv)

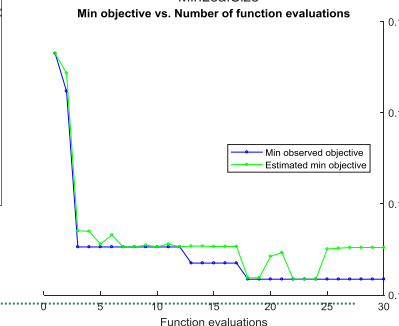
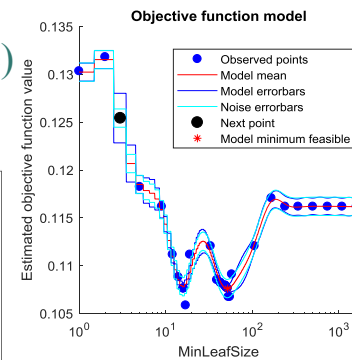
1. Naive Bayes
2. Classification Tree
3. Classification Tree – Optimalizovaný parametr Minimální velikost listu (MinLeafSize)
(75% trénovací, 25% testovací)



2. Classification Tree



3. Classification Tree



3. Průběh optimalizace

Příklad



**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

(train) Naive Bayes Confusion Matrix

	1	2	
1	2759 81.4%	201 5.9%	93.2% 6.8%
2	238 7.0%	193 5.7%	44.8% 55.2%
	92.1% 7.9%	49.0% 51.0%	87.1% 12.9%
	1	2	
	Target Class		

(train) Classification Tree Confusion Matrix

	1	2	
1	2943 86.8%	63 1.9%	97.9% 2.1%
2	54 1.6%	331 9.8%	86.0% 14.0%
	98.2% 1.8%	84.0% 16.0%	96.5% 3.5%
	1	2	
	Target Class		

(train) Optimized Classification Tree Confusion Matrix

	1	2	
1	2927 86.3%	260 7.7%	91.8% 8.2%
2	70 2.1%	134 4.0%	65.7% 34.3%
	97.7% 2.3%	34.0% 66.0%	90.3% 9.7%
	1	2	
	Target Class		

(test) Naive Bayes Confusion Matrix

	1	2	
1	916 81.1%	57 5.0%	94.1% 5.9%
2	87 7.7%	70 6.2%	44.6% 55.4%
	91.3% 8.7%	55.1% 44.9%	87.3% 12.7%
	1	2	
	Target Class		

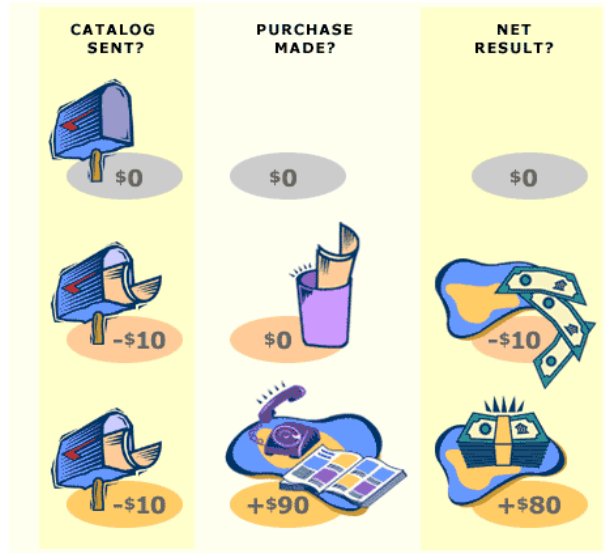
(test) Classification Tree Confusion Matrix

	1	2	
1	922 81.6%	63 5.6%	93.6% 6.4%
2	81 7.2%	64 5.7%	44.1% 55.9%
	91.9% 8.1%	50.4% 49.6%	87.3% 12.7%
	1	2	
	Target Class		

(test) Optimized Classification Tree Confusion Matrix

	1	2	
1	970 85.8%	83 7.3%	92.1% 7.9%
2	33 2.9%	44 3.9%	57.1% 42.9%
	96.7% 3.3%	34.6% 65.4%	89.7% 10.3%
	1	2	
	Target Class		

Příklad – Matice cen za chybu

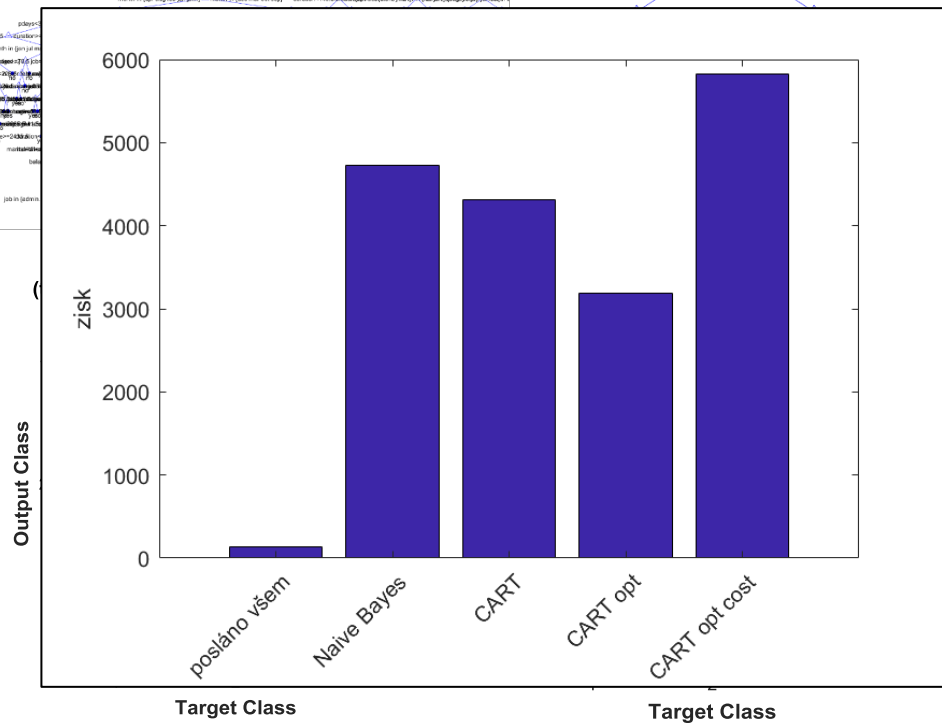
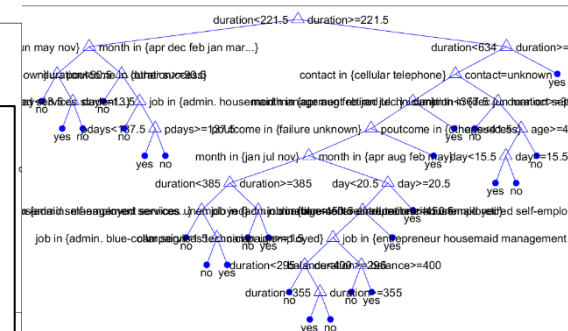
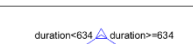
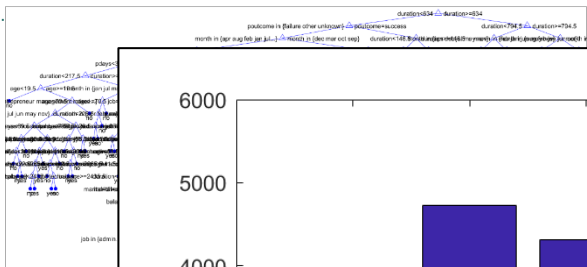


Matice cen za chybu
Predikce

Skutečnost	no	yes
no	0	1
yes	8	0

Pokud pošlu leták všem:
 $zisk = 90 * 127 - 10 * 1130 = 130$

Příklad



(test) Naive Bayes Confusion Matrix

Output Class \ Target Class	1	2	
1	916 81.1%	57 5.0%	94.1% 5.9%
2	87 7.7%	70 6.2%	44.6% 55.4%
	91.3% 8.7%	55.1% 44.9%	87.3% 12.7%

(test) Cost-Adjusted Classification Tree Confusion Matrix

Output Class \ Target Class	1	2	
1	785 69.5%	27 2.4%	96.7% 3.3%
2	218 19.3%	100 8.8%	31.4% 68.6%
	78.3% 21.7%	78.7% 21.3%	78.3% 21.7%

zisk

4730 4310 3190 5820

Děkuji za pozornost

Některé snímky převzaty od:
prof. Ing. Petr Berka, CSc. berka@vse.cz