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

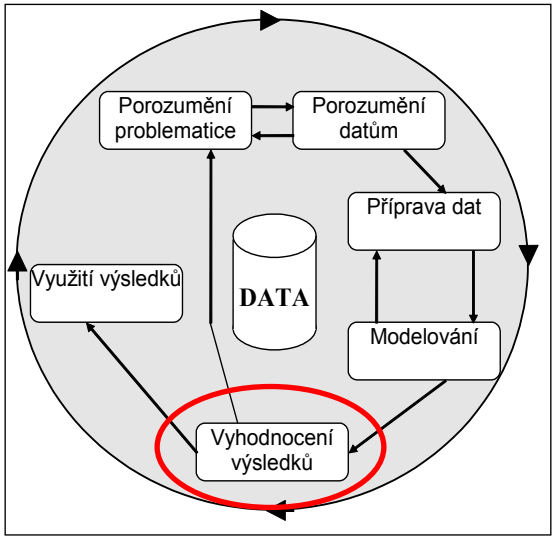
Jan Górecki

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Motivace
- Deskriptivní úlohy
- Klasifikační úlohy
- Hodnocení jedním/dvěma čísly
- Reálná úloha



Motivace

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Zadarmo ani kuře nehrabe.
(Nikde nelétají pečení holubi do huby!
Bez práce nejsou koláče,
…)

"No Free Lunch" :(

D. H. Wolpert. The supervised learning no-free-lunch theorems. In Soft Computing and Industry, pages 25–42. Springer, 2002.

Our model is a simplification of reality

⇩

Simplification is based on assumptions (model bias)

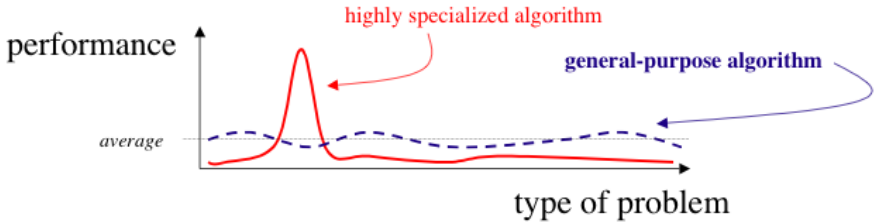
⇩

Assumptions fail in certain situations

Roughly speaking:

“No one model works best for all possible situations.”

No free lunch pro učení s učitelem



Deskriptivní úlohy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- 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!

Klasifikační úlohy – motivační příklad

SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINÉ

- Data (10 bodů metráž bytu vs cena, z kubické funkce + error
- Modely – lin, kvard, kub a x na 10
- Zobrazit modely a chyby
- Pak ilustrovat problém x na 10 na nových datech

Klasifikační úlohy

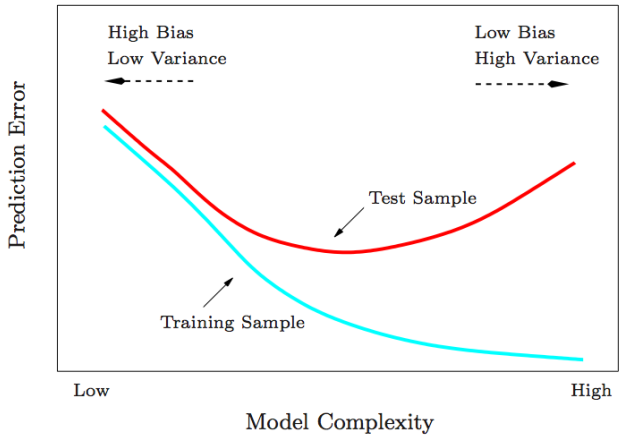
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- 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

Matice záměn (Confusion matrix)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

	Predikce	
Skutečnost	ano	ne
ano	TP	FN
ne	FP	TN

	Predikce	
Skutečnost	ano	ne
ano	5	3
ne	0	4

Hodnocení jedním/dvěma čísly

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

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

Celková správnost $\in [Acc_{def}, Acc_{max}] \subseteq [0, 1]$, 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)

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

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

	Predikce	
Skutečnost	ano	ne
ano	TP	FN
ne	FP	TN

	Predikce	
Skutečnost	ano	ne
ano	8	0
ne	1	3

	Predikce	
Skutečnost	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 nalezli

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

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

	Predikce	
Skutečnost	ano	ne
ano	TP	FN
ne	FP	TN

	Predikce	
Skutečnost	ano	ne
ano	8	0
ne	1	3

	Predikce	
Skutečnost	ano	ne
ano	5	3
ne	0	4

Sensitivita a specificita

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Hodnocení kvality testu na nějakou nemoc:

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

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

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

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Chyba bez ceny

$$Err = 1 - Acc$$

Chyba s cenami

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

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

		Predikce	
	Skutečnost	ano	ne
	ano	TP	FN
	ne	FP	TN

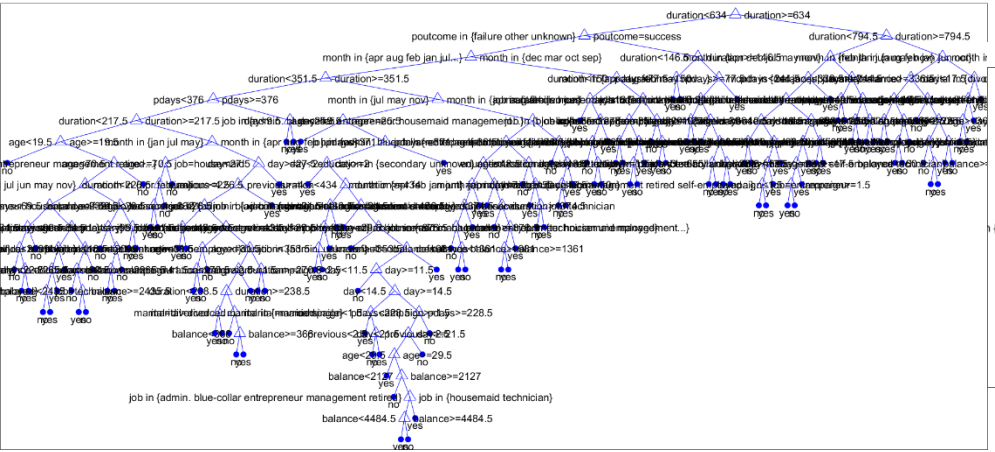
Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

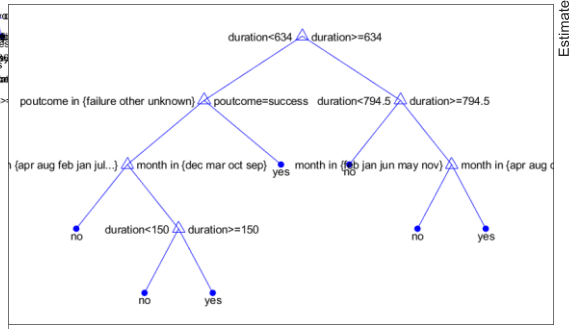
Evaluace tří modelů získaných pro data 4521 klientů portugalské banky (L:\gorecki\Public\NPDOD-NKDOD\Data\bank.csv)

1. Naive Bayes
2. Classification Tree
3. Classification Tree – Optimalizovaný parametr Minimální velikost listu (MinLeafSize)

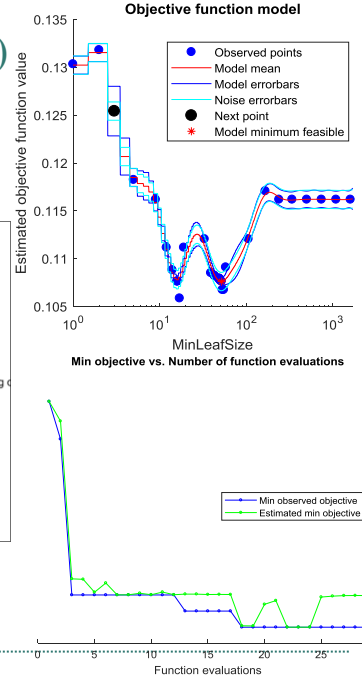
(75% trénovací, 25% testovací)



2. Classification Tree



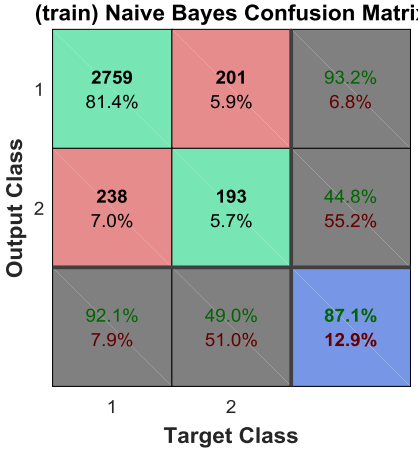
3. Classification Tree

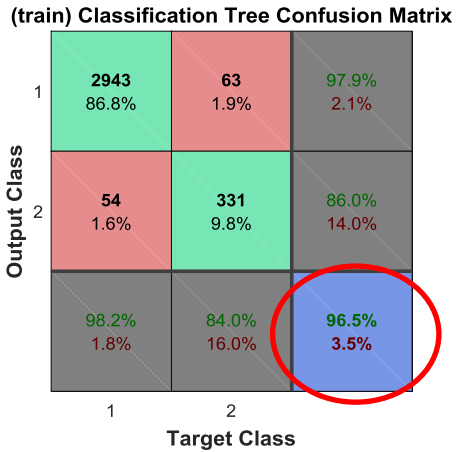


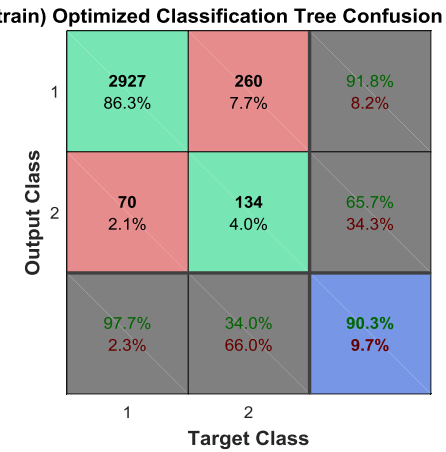
3. Průběh optimalizace

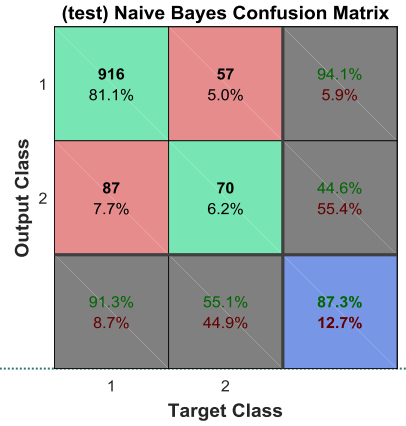
Příklad

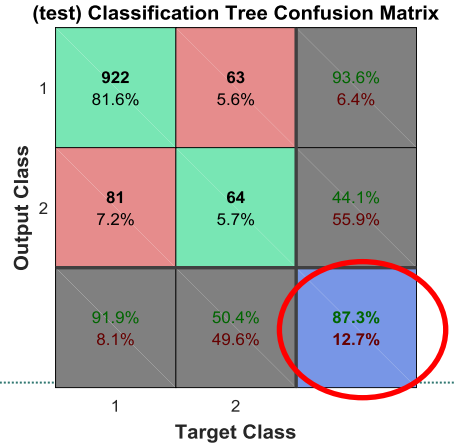
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

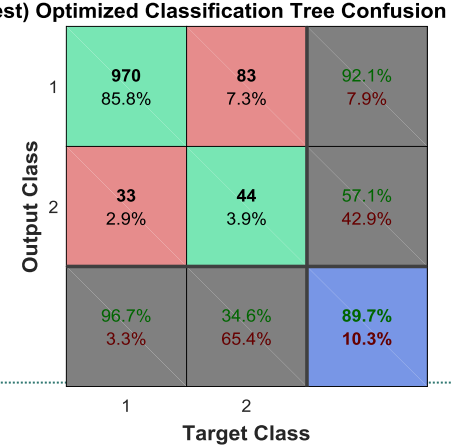












Příklad – Matice cen za chybu

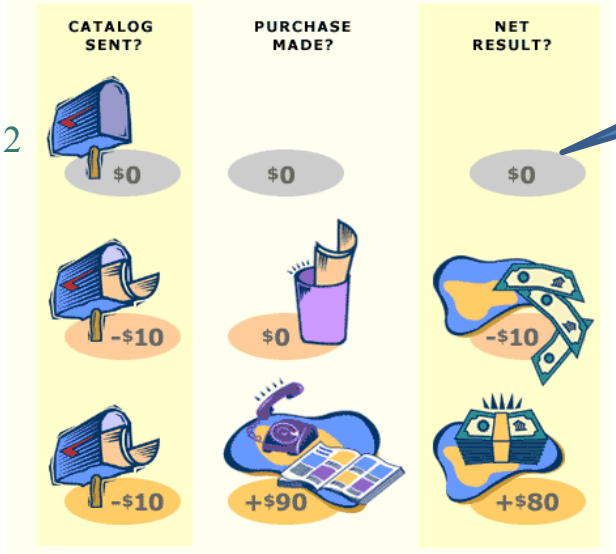
Zde je potřeba zvážit dvě teoretické situace (koupil by, kdybych poslal?):
1) Pokud ne, tak jsem šul nul
2) Pokud ano, tak jsem teoreticky přišel o zisk 80$!!

Interpretace:
Když predikuji 100% správně, vydělám maximum (=MAX)
1) Pokud udělám chybu "predikce yes, ale skutečnost no", pak prodělám **10$**.
2) Pokud udělám chybu "predikce no, ale skutečnost yes", pak přijdu o zisk **80$!!**.

Situace 1 a 2 (TN a FN)

Situace 3 (FP)

Situace 4 (TP)

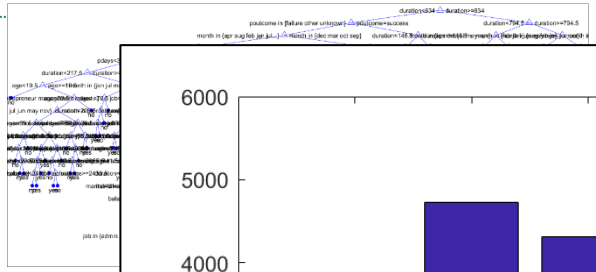


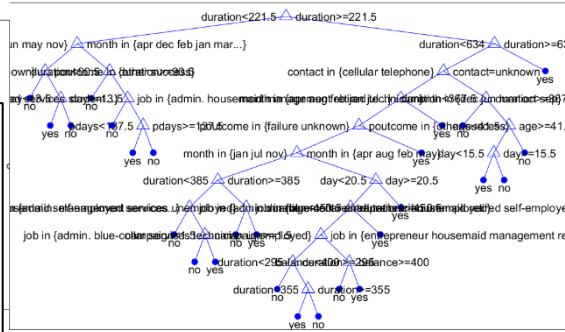
Matice cen za chybu

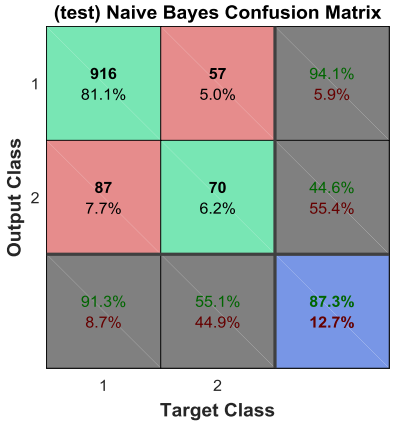
Skutečnost \ **Predikce**	no	yes
no	0	10
yes	80	0

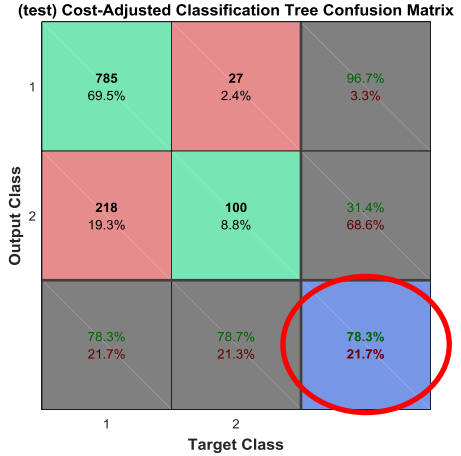
Pokud pošlu leták všem:
zisk = 90 * 127 – 10 * 1130 = **130**

Příklad









zisk			
4730	4310	3190	5820

Děkuji za pozornost

Některé snímky převzaty od:

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