

# 11MAI – Cvičení 6

Opakování úloh statistické analýzy dat

Jan Přikryl

ČVUT FD

pondělí 7. listopadu 2022

verze: 2022-11-07 12:24

# Problém 1

U každé části (a) až (d) uveďte, zda bychom obecně očekávali, že výsledky neparametrické statistické metody učení budou lepší nebo horší, než výsledky parametrické metody. Odůvodněte odpověď.

- (a) Velikost vzorku  $n$  je extrémně velká a počet prediktorů  $p$  je malý.
- (b) Počet prediktorů  $p$  je extrémně velký a počet pozorování  $n$  je malý.
- (c) Vztah mezi prediktory a odpovědí (závislou proměnnou) je velmi nelineární.
- (d) Rozptyl chybových členů, tj.  $\sigma^2 = \text{var}(x)$ , je extrémně vysoký.

## Problém 2

Vysvětlete, zda daný scénář reprezentuje klasifikační nebo regresní problém a uveďte, zda nás více zajímá inference (indukce) nebo predikce (dedukce). Uveďte  $n$  a  $p$ .

- (a) Shromáždili jsme soubor údajů o 500 nejlepších firmách v USA. Pro každou firmu jsme zaznamenali zisk, počet zaměstnanců, odvětví průmyslu a mzdu generálního ředitele. Chceme pochopit, jaké faktory ovlivňují výši platu generálního ředitele.

## Problém 2

Vysvětlete, zda daný scénář reprezentuje klasifikační nebo regresní problém a uveďte, zda nás více zajímá inference (indukce) nebo predikce (dedukce). Uveďte  $n$  a  $p$ .

- (a) Shromáždili jsme soubor údajů o 500 nejlepších firmách v USA. Pro každou firmu jsme zaznamenali zisk, počet zaměstnanců, odvětví průmyslu a mzdu generálního ředitele. Chceme pochopit, jaké faktory ovlivňují výši platu generálního ředitele.
- (b) Uvažujeme o uvedení nového produktu na trh a přejeme si vědět, zda bude úspěšný nebo ne. Shromáždíme údaje o průběhu uvedení na trh u 20 podobných produktů. Pro každý produkt zaznameneáme, zda byl či nebyl úspěšný, cenu účtovanou za produkt, marketingový rozpočet, konkurenční cenu a deset dalších proměnných.

## Problém 2

Vysvětlete, zda daný scénář reprezentuje klasifikační nebo regresní problém a uveďte, zda nás více zajímá inference (indukce) nebo predikce (dedukce). Uveďte  $n$  a  $p$ .

- (a) Shromáždili jsme soubor údajů o 500 nejlepších firmách v USA. Pro každou firmu jsme zaznamenali zisk, počet zaměstnanců, odvětví průmyslu a mzdu generálního ředitele. Chceme pochopit, jaké faktory ovlivňují výši platu generálního ředitele.
- (b) Uvažujeme o uvedení nového produktu na trh a přejeme si vědět, zda bude úspěšný nebo ne. Shromáždíme údaje o průběhu uvedení na trh u 20 podobných produktů. Pro každý produkt zaznamenáme, zda byl či nebyl úspěšný, cenu účtovanou za produkt, marketingový rozpočet, konkurenční cenu a deset dalších proměnných.
- (c) Zajímá nás předpověď procentuální změny kurzu amerického dolaru ve vztahu k týdenním změnám na světových akciových trzích. Z tohoto důvodu v každém týdnu roku 2019 sledujeme týdenní procentuální změnu kurzu dolaru a procentuální změny výkonnosti akcií na americkém, britském, německém a japonském trhu.

## Problém 3

Vraťme se nyní k rozkladu chyby na zkreslení a rozptyl.

(a) Do jednoho grafu načrtněte typickou závislost

- ▶ kvadrátu zkreslení,
- ▶ rozptylu,
- ▶ kvadrátu trénovací chyby,
- ▶ kvadrátu testovací chyby a
- ▶ kvadrátu Bayesovy chyby (tedy rozptylu neredukovatelné chyby)

v závislosti na flexibilitě modelu. Postupujte od méně flexibilních metod statistického učení k pružnějším přístupům, osa  $x$  by měla představovat množství flexibility v metodě a osa  $y$  by měla představovat hodnoty pro každou křivku. Každou z pěti výsledných křivek vhodně označte (barevně, typem čáry).

(b) Vysvětlete, proč každá z křivek má tvar, zobrazený v části (a).

## Problém 4

Nyní se zamyslete nad některými reálnými aplikacemi pro statistické učení

- (a) Popište tři reálné aplikace, v nichž by mohla být užitečná klasifikace. Popište závislou proměnnou, stejně jako prediktory. Je cílem každé aplikace inference nebo predikce? Vysvětlete svoji odpověď.
- (b) Popište tři reálné aplikace, v nichž by mohla být užitečná regrese. Popište závislou proměnnou, stejně jako prediktory. Je cílem každé aplikace inference nebo predikce? Vysvětlete svoji odpověď.
- (c) Popište tři reálné aplikace, v nichž by mohlo být užitečné shlukování.

## Problém 5

Jaké jsou výhody a nevýhody velmi flexibilního (oproti méně flexibilnímu) přístupu k regresi nebo klasifikaci? Za jakých okolností může být preferován flexibilnější přístup než méně flexibilní přístup? Kdy může být preferován méně flexibilní přístup?



## Problém 6

Popište rozdíly mezi parametrickým a neparametrickým přístupem ke statistickému učení. Jaké jsou výhody parametrického přístupu k regresi nebo klasifikaci (na rozdíl od neparametrického přístupu)? Jaké jsou jeho nevýhody?

## Problém 7

Níže uvedená tabulka obsahuje soubor trénovacích dat obsahující šest pozorování, tři prediktory a jednu kvalitativní cílovou proměnnou.

#	$X_1$	$X_2$	$X_3$	$Y$
1	0	3	0	červená
2	2	0	0	červená
3	0	1	3	červená
4	0	1	2	zelená
5	-1	0	1	zelená
6	1	1	1	červená

## Problém 7 (pokračování)

Předpokládejme, že chceme použít tuto množinu dat k předpovědi  $Y$  pomocí metody  $k$  nejbližších sousedů ( $k$ -NN) v případě, kdy  $X_1 = X_2 = X_3 = 0$ .

- (a) Vypočítejte euklidovskou vzdálenost mezi každým pozorováním a bodem  $X_1 = X_2 = X_3 = 0$ .
- (b) Jaká bude Vaše předpověď s  $k = 1$ ? Proč?
- (c) Jaká bude Vaše předpověď s  $k = 3$ ? Proč?
- (d) Je-li Bayesova rozhodovací hranice v tomto problému vysoce nelineární, budete očekávat, že nejlepší hodnota pro  $k$  bude velká nebo malá? Proč?

## Problém 8

Popište nulové hypotézy, kterým odpovídají  $p$ -hodnoty uvedené v tabulce. Vysvětlete, jaké závěry můžete vyvodit na základě těchto  $p$ -hodnot. Vaše vysvětlení by mělo být formulováno z hlediska **sales**, **TV**, **radio**, a **newspaper**, spíše než z hlediska koeficientů lineárního modelu.

	$\beta_i$	$s(\beta_i)$	$t$ -statistika	$p$ -hodnota
–	2,939	0,3119	9,42	$< 0,0001$
<b>TV</b>	0,046	0,0014	32,81	$< 0,0001$
<b>radio</b>	0,189	0,0086	21,89	$< 0,0001$
<b>newspaper</b>	-0,001	0,0059	-0,18	0,8599

## Problém 9

Pečlivě vysvětlete rozdíly mezi KNN klasifikátory a KNN regresními metodami.

## Problém 10

Předpokládejme, že máme soubor dat s pěti prediktory,  $x_1 = \text{GPA}$ ,  $x_2 = \text{IQ}$ ,  $x_3 = \text{Gender}$  (1 pro ženu a 0 pro muže),  $x_4 = \text{Interakce mezi GPA a IQ}$  a  $x_5 = \text{Interakce mezi GPA a Gender}$ . Závislou proměnnou je počáteční plat po promoci v tisících dolarů. Předpokládejme dále, že k sestavení modelu použijeme metodu nejmenších čtverců a dostaneme  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0,07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0,01$ ,  $\beta_5 = -10$ .

(a) Které z následujících tvrzení je správné a proč?

- Pro danou hodnotu **IQ** a **GPA** vydělávají muži v průměru více, než ženy.
- Pro danou hodnotu **IQ** a **GPA** ženy vydělávají v průměru více, než muži.
- Pro danou hodnotu **IQ** a **GPA** vydělávají muži v průměru více, než ženy, pokud je **GPA** dostatečně vysoká.
- Pro danou hodnotu **IQ** a **GPA** vydělávají ženy v průměru více, než muži, za předpokladu, že je **GPA** dostatečně vysoká.

# Problém 10

pokračování

(b) Predikujte plat ženy s IQ 110 a GPA 4,0.

(c) Pravda nebo nepravda:

Vzhledem k tomu, že koeficient pro interakci GPA/IQ je velmi malý, existuje velmi málo důkazů o interakčním účinku. Odůvodněte odpověď.

## Problém 11

Shromáždíme sadu dat ( $n = 100$  pozorování) obsahujících jediný prediktor a kvantitativní odpověď. Poté na datech identifikujeme lineární regresní model a také ještě kubickou regresi, tj.  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$ .

- (a) Předpokládejme, že skutečný vztah mezi  $x$  a  $y$  je lineární, tj.  $y = \beta_0 + \beta_1 x + \epsilon$ . Zvažte trénovací zbytkový součet čtverců (RSS) pro lineární regresi a také trénovací RSS pro kubickou regresi. Budeme očekávat, že jedna hodnota bude nižší, než druhá, očekávali bychom, že budou stejné, nebo není dostatek informací k tomu, abychom to mohli říct? Odůvodněte odpověď.
- (b) Odpovězte na (a) pro případ, kdy použijete RSS spočtené na testovací množině a nikoliv trénovací RSS.



# Problém 12

## pokračování

- (c) Předpokládejme nyní, že skutečný vztah mezi  $x$  a  $y$  není lineární, že ale nevíme, jak daleko je od lineárního. Uvažujte trénovací RSS pro lineární regresi a také trénovací RSS pro kubickou regresi. Budeme očekávat, že jedna hodnota bude nižší, než druhá, očekávali bychom, že budou stejné, nebo není dostatek informací k tomu, abychom to mohli říct? Odůvodněte odpověď.
- (d) Odpovězte (c) pro případ, kdy použijete RSS spočtené na testovací množině a nikoliv trénovací RSS.

## Automobily – jednoduchá regrese

Vyzkoušejte si jednoduchou lineární regresi na datové sadě `islr_auto.csv`.

- (a) Použijte funkci `mdl=fitlm()` pro stanovení jednoduché lineární regresní závislosti s `mpg` jako odezvou a `horsepower` jako prediktorem. Pro vypsání výsledků použijte obsah `mdl`. Komentujte výstup. Například:
- Existuje nějaký vztah mezi prediktorem a odpovědí?
  - Jak silný je vztah mezi prediktorem a odpovědí?
  - Je vztah mezi prediktorem a odpovědí pozitivní nebo negativní?
  - Jaké je předpokládané `mpg` spojené s `horsepower` 98? Jaké jsou 95 % intervaly spolehlivosti koeficientů a predikce?

# Automobily – jednoduchá regrese

## pokračování

Pokračujeme s grafy:

- (b) Pomocí funkce `plot(md1)` vykreslete odpověď a prediktor a zobrazte regresní přímku nejmenších čtverců.
- (c) Pomocí funkce `coefCI(md1)` vypište intervalové odhady jednotlivých parametrů regrese.
- (d) Použijte funkci `plot()` pro vytvoření diagnostických grafů nejmenších čtverců. Komentujte jakýkoli problém s proložením dat přímkou, který zaznamenáte.

## Automobily – vícenásobná regrese

Nyní se přesuneme k vícenásobné lineární regresi na téže datové sadě.

- (a) Pomocí `gplotmatrix` vytvořte matici korelačních diagramů, zahrnující všechny proměnné v datové sadě.
- (b) Vypočtěte matici korelací mezi proměnnými pomocí funkce `corr()`. Budete muset vyloučit proměnnou `name`, která je kvalitativní?
- (c) Použijte `mdl2=fitlm()` k určení vícenásobné lineární regrese s `mpg` jako odezvou a všemi ostatními proměnnými s výjimkou `name` jako prediktory. Pomocí `mdl2` vytiskněte výsledky. Komentujte výstup. Například:
  - i. Existuje vztah mezi prediktory a odpovědí?
  - ii. Které prediktory mají statisticky významný vztah k odpovědi?
  - iii. Co naznačuje koeficient pro proměnnou `year`?

# Automobily – vícenásobná regrese

## pokračování

Pokračujeme s grafy:

- (d) Vytvořte diagnostické grafy lineární regrese. Komentujte jakýkoli problém, který vidíte s proložením. Naznačují grafy reziduí nějaké neobvykle velké odchylky? Vykazuje leverage graf nějaké pozorování s neobvykle vysokým pákovým efektem?
- (e) Použijte symboly "\*" a ":" pro vytvoření lineárních regresních modelů s interakčními efekty. Jeví se nějaké interakce jako statisticky významné?
- (f) Vyzkoušejte několik různých transformací proměnných, jako je například  $\xi = \log(\mathbf{x})$ ,  $\xi = \sqrt{\mathbf{x}}$ ,  $\xi = \mathbf{x}^2$ . Komentujte svá zjištění.

# Boston

## Kriminalita v Bostonu

Tento problém zahrnuje datovou sadu Boston, uloženou v `islr_boston.csv`. Pokusíme se na ní předpovědět míru kriminality na obyvatele za použití dalších proměnných v tomto souboru údajů. Jinak řečeno: míra kriminality na obyvatele je odpovědí a ostatní proměnné jsou předpovědi.

- (a) Pro každý prediktor použijte jednoduchý model lineární regrese, který předpovídá odpověď. Popište své výsledky. V kterém z modelů existuje statisticky významná souvislost mezi prediktorem a odpovědí? Vytvořte nějaké obrázky, které vaše tvrzení podpoří.
- (b) Použijte model s vícenásobnou regresí pro předpověď pomocí všech prediktorů. Popište své výsledky. Pro které prediktory lze odmítnout nulovou hypotézu  $H_0 : \beta_j = 0$ ?

- (c) Jak se vaše výsledky z (a) srovnávají s výsledky z podle (b)? Vytvořte graf zobrazující jednorozměrné regresní koeficienty z (a) na ose  $x$  a více regresních koeficientů z (b) na ose  $y$ . To znamená, že každý prediktor je zobrazen jako jediný bod v grafu. Jeho součinitel v jednoduchém modelu lineární regrese je zobrazen na ose  $x$  a jeho odhad koeficientu v modelu vícenásobné lineární regrese je zobrazen na ose  $y$ .
- (d) Existuje důkaz polynomiální závislosti mezi některým z prediktorů a odpovědí? Chcete-li odpovědět na tuto otázku, použijte pro každý prediktor  $x_j$  model

$$y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3 + \epsilon.$$

a vyhodnoťte jej.

# Data z akciového trhu

Lze odhadnout vývoj akciového indexu?

Nejprve prozkoumáme data z akciových trhů, konkrétně denní vývoj indexu S&P v letech 2001–2005.

Načteme a zobrazíme základní charakteristiky

```
smarket = readtable('islr_smarket.csv');  
summary(smarket)
```

Proměnná **Direction** je kategorická (buď **Up** nebo **Down**) a je potřeba to Matlabu sdělit:

```
smarket.Direction = categorical(smarket.Direction)
```



# Data z akciového trhu

## Pokračování

Pokud bychom chtěli zkoumat korelace mezi jednotlivými numerickými proměnnými, nabízí se funkce `corrcoef()` ...

```
corrcoef(smarket)
```

...jenže jako vstup je potřeba matice:

```
smarket_matrix = table2array(smarket(:,2:end-1))  
smarket_cc = corrcoef(smarket_matrix);
```

**Vysvětlete**, proč indexujeme `smarket(:,2:end-1)`.

Najdete v korelační matici hodnoty naznačující, že nějaké veličiny jsou korelované? Pokud ano, kterým proměnným odpovídají?

# Data z akciového trhu

## Pokračování

Pokud bychom chtěli zkoumat korelace mezi jednotlivými numerickými proměnnými, nabízí se funkce `corrcoef()` ...

```
corrcoef(smarket)
```

...jenže jako vstup je potřeba matice:

```
smarket_matrix = table2array(smarket(:,2:end-1))  
smarket_cc = corrcoef(smarket_matrix);
```

**Vysvětlete**, proč indexujeme `smarket(:,2:end-1)`.

Vykreslíme

```
plot(smarket.Volume)
```

**Na základě grafu vysvětlete**, proč je mezi `Year` a `Volume` pozitivní korelace.

# Data z akciového trhu

## Logistická regrese

Natrénujeme generalizovaný lineární model závislosti **Direction** na **Lag1** až **Lag5**.  
Logistickou závislost specifikujeme volbou '**Distribution**', '**binomial**':

```
mdl = fitglm(smarket,  
             'Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume',  
             'Distribution', 'binomial')
```

Který regresní koeficient má nejmenší  $p$ -hodnotu? Naznačuje tato hodnota silnou vazbu na výstup modelu?

# Data z akciového trhu

## Pokračování

Podívejme se na predikci modelu na původních pozorováních:

```
probs = predict mdl
```

Jak dobře model predikuje vývoj trhu zjistíme porovnáním s trénovacími hodnotami v **Direction**. Musíme ale **probs** převést na kategorickou proměnnou s hodnotami **Up** a **Down**:

```
predictions = repmat(categorical({'Down'}), mdl.NumObservations, 1);  
predictions(probs>0.5) = 'Up'; %
```

Pokračujeme maticí záměn:

```
confusionmat(predictions, smarket.Direction)  
(507+145)/1250  
mean(predictions == smarket.Direction)
```

# Data z akciového trhu

## Pokračování

Je náš model lepší, než náhodné rozhodování? Jaká je jeho trénovací chyba?  
Lepší odhad chyby, kterou model bude v reálu vykazovat, lze získat rozdělením na trénovací a testovací sadu. Zkusme identifikovat model na datech z let 2001–2004 a ověřit jeho předpovědi na datech z roku 2005.

```
train = (smarket.Year<2005); % Boolovský sloupcový vektor true/false  
smarket_train = smarket(train,:);  
smarket_test = smarket(~train,:);
```

Co znamená `smarket(train,:)`, `smarket(~train,:)`?

Jak velká je trénovací a testovací množina?

```
size(smarket_train)  
size(smarket_test)
```

# Data z akciového trhu

## Pokračování

Identifikujeme model a porovnáme jej na datech z roku 2005:

```
mdl1 = fitglm(smarket_train,  
'Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume',  
'Distribution', 'binomial')  
probs = predict mdl1, smarket_test);  
% Prevod na Up/Down  
predictions = repmat(categorical({'Down'}),mdl1.NumObservations,1);  
predictions(probs>0.5) = 'Up';  
% Matice zamen a procento spravnych predpovedi  
confusionmat(predictions, smarket_test.Direction)  
mean(predictions == smarket_test.Direction)
```

Jaká je chyba testovací sady?

# Data z akciového trhu

## Pokračování

Identifikujeme jednodušší model pouze se členy **Lag1** a **Lag2**, které v originální logistické regresi měly nejsilnější vztah k výstupu:

```
mdl = fitglm(smarket_train,  
  'Direction~Lag1+Lag2',  
  'Distribution', 'binomial')  
probs = predict(mdl, smarket_test);  
predictions = repmat(categorical({'Down'}),252,1);  
predictions(probs>0.5) = 'Up';  
confusionmat(predictions, smarket_test.Direction)  
mean(predictions == smarket_test.Direction)
```

Jaký je odhad testovací chyby nyní? Jaká je pravděpodobnost předpovědi růstu trhu?  
Poklesu trhu?

# Data z akciového trhu

## Pokračování

Na závěr si ukážeme, jak spočítat predikce u nových hodnot **Lag1** a **Lag2** daných následující tabulkou:

Lag1	Lag2
1,2	1,1
1,5	-0,8

```
% Vytvoríme novou Matlabi tabulku  
pt = table([1.2;1.5], [1.1;-0.8], 'VariableNames', {'Lag1', 'Lag2'});  
% Vyhodnotíme model na datech uložených v 'pt'  
predict(mdlt2, pt)'
```

Místo tabulky můžete v tomto případě použít i **pt** reprezentované maticí. Jak to uděláte?



# Data z akciového trhu

## Diskriminační analýza

Nyní zkusíme to samé pomocí lineární diskriminační analýzy. V Matlabu je na to obecná metoda `fitcdiscr()`, implementující i vyšší polynomiální reprezentace hranice.

Vstupem metody je zvlášť matice prediktorů a zvlášť odpověď modelu:

```
x = [ smarket_train.Lag1, smarket_train.Lag2 ];  
y = smarket_train.Direction;  
cmdl = fitcdiscr(x,y)
```

Vidíme, že `cmdl` neobsahuje údaje o názvech proměnných, doplníme:

```
cmdl = fitcdiscr(x,y,  
'PredictorNames',{'Lag1','Lag2'},  
'ResponseName','Direction')
```

# Data z akciového trhu

## Pokračování

Zkusíme si vykreslit hranici a hodnoty v jednotlivých třídách. Podívejte se nejprve, k čemu slouží funkce `gscatter()` a `ezplot()`.

```
% Vykreslíme data a jejich třídu Up/Down
gscatter(smarket.Lag1, smarket.Lag2, smarket.Direction);
hold on
% Definice funkce pro ezplot()
f = @(x1,x2) K + L(1)*x1 + L(2)*x2;
K = cmd1.Coeffs(1,2).Const;
L = cmd1.Coeffs(1,2).Linear;
% Vykreslíme hranici
h2 = ezplot(f, [-6,6,-6,6]);
```

# Data z akciového trhu

## Pokračování

Matice záměn a celková testovací chyba modelu je totožná s logit modelem:

```
xtest = [smarket_test.Lag1, smarket_test.Lag2];  
predictions = predict(cmdl, xtest);  
confusionmat(predictions, smarket_test.Direction)  
mean(predictions == smarket_test.Direction)
```