Cvičení 7

Klasifikace

Jan Přikryl

ČVUT FD

9. dubna 2020

Příklad 1

Data z akciového trhu

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

Načteme a zobrazíme základní charakterisktiky

```
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)
```

Příklad 1

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?

Příklad 1

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.

Příklad 2

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?

Příklad 2

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 hodonotami `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)
```

Příklad 2

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 testovacísadu. Zkusme identifikovat model na datech z let 2001–2004 a ověřit jeho předpovědi na datech z roku 2005.

```
train = (smarket.Year<2005); % Logicky sloupcovy 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)
```

Příklad 2

Pokračování

Identifikujeme model a porovnáme jej na datech z roku 2005:

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

Jaká je chyba testovací sady?

Příklad 2

Pokračování

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

```
mdlt = fitglm(smarket_train,
        'Direction ~ Lag1+Lag2',
        'Distribution', 'binomial')
probs = predict(mdlt, 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?

Příklad 2

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

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

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

Příklad 3

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')
```

Příklad 3

Pokračování

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

```
% Vykreslime data a jejich tridu 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 = cmdl.Coeffs(1,2).Const;
L = cmdl.Coeffs(1,2).Linear;
% Vykreslime hranici
h2 = ezplot(f, [-6,6,-6,6]);
```

Příklad 3

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)
```

Samostatná práce

Kvadratická diskriminační analýza a KNN

Samostatně vyzkoušejte:

(a) kvadratická diskriminační analýzu,

(b) klasifikaci pomocí metody k nejbližších sousedů (KNN).