Úkol 4 - Diskriminační analýza

4ST512 Vícerozměrná statistika

Jiří Filip

Úvod

V této analýze se budeme zaobírat údaji o tělesných rozměrech mužů a žen (obvod břicha, předloktí a kolene) a do jaké míry lze tyto rozměry využít k rozhodnutí o pohlaví jedince. Využijeme metodu Lineární diskriminační analýzy (LDA). Data rovněž obsahují údaje o 10 jedincí neznámého pohlaví, jež se pokusíme zařadit buďto jako muže, nebo ženu.

id	o4	o8	o9	pohlavi
M001	75.4	26.1	31.7	muž
M002	80.2	26.1	36.2	muž
M003	92.4	29.2	38.4	muž
M004	67.4	26.6	32.8	muž
M005	87.3	28.4	35.5	muž
M006	99.0	29.0	41.2	muž

Průzkum dat z hlediska předpokladů LDA

Nejprve data prozkoumáme z hlediska předpokladů diskriminační analýzy. Data popisují muže a ženy a jejich tělesné rozměry. Naším cílem bude na základě těchto tělesných rozměrů rozhodnout o pohlaví deseti neznámých jedinců.

Průměry

Nejprve se podívejme, jak se jednotlivé třídy (muž a žena) liší, co se týče průměrů tělesných rozměrů. Vidíme, že všechny rozměry se nějakým způsobem liší - to nám dopomůže k lepší diskriminaci pohlaví.

pohlavi	o4	o8	o9
muž	87.62975	28.22975	37.18017
žena	83.88189	23.78583	35.29173

Směrodatné odchylky

Stejně tak se podívejme na směrodatné odchylky rozměrů. Ty nám (společně s informacemi o průměrech) dávají nějakou představu o tom, jak se jednotlivé distribuce rozměrů ve třídách (muž a žena) překrývají. Vidíme, že u prvního rozměru se odchylky liší výrazně, u druhých dvou už méně (až zanedbatelně). Toto by však mohlo být způsobeno jinou škálou, kdy desetinný rozdíl u druhých dvou rozměrů znamená více než u prvního.

pohlavi	o4	o8	o9
muž	8.40216	1.790391	2.290740
žena	10.00819	1.675031	2.586505

Shapirův-Wilkův test

Další tabulka představuje výsledky (p-hodnoty) Shapirova-Wilkova testu s nulovou hypotézou o normalitě dat. Lze vidět, že na hladině významnosti 5 % můžeme zamítnout hypotézu o normalitě dat u všech obvodů ženy a u prvního obvodu muže.

pohlavi	o4	o8	o9
muž	0.0071741	0.6357415	0.1369338
žena	0.0000493	0.0000016	0.0000004

Kovarianční matice

Porovnejme ještě konvarianční matice mužů a žen. Vidíme, že se dvě kovarianční matice od sebe značně liší.

Muž

o4	o8	o9
70.596290	6.572555	10.807398
6.572555	3.205501	2.167813
10.807398	2.167813	5.247489

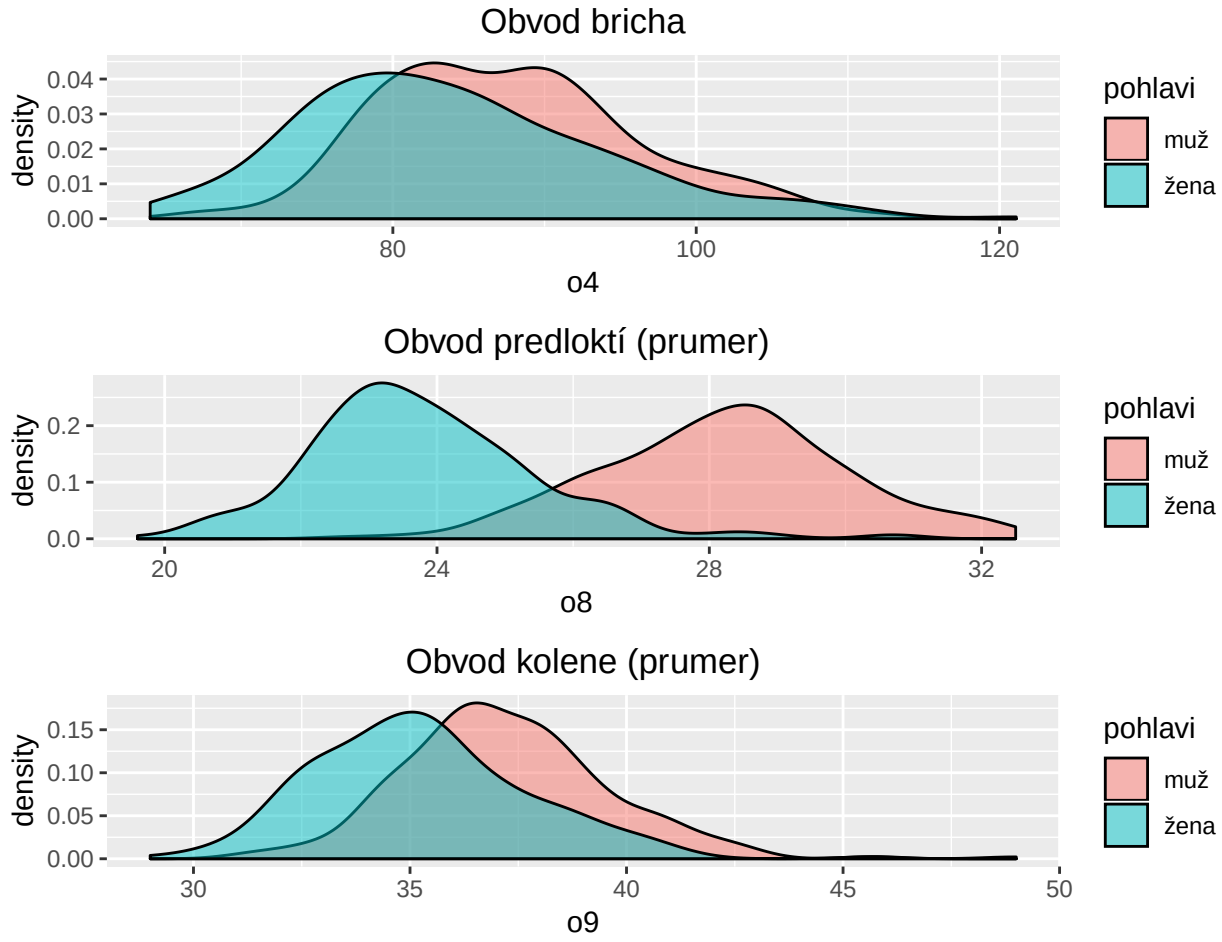
Žena

o4	o8	o9
100.16386	10.839149	15.992379
10.83915	2.805727	3.254744
15.99238	3.254744	6.690010

To potvrzuje i Boxův M test, u něhož na hladině významnosti 5 % zamítáme hypotézu o homogenitě kovariančních matic.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data.train %>% select(o4, o8, o9)
## Chi-Sq (approx.) = 42.344, df = 6, p-value = 0.0000001572
```

Pro přehled si ještě zobrazme jednotlivá rozdělení tělesných rozměrů pro muže a pro ženy. Vidíme, že zdaleka nejméně se překrývají dvě rozdělení u obvodu předloktí. Očekávali bychom tak, že obvod předloktí bude nejlépe diskriminovat.

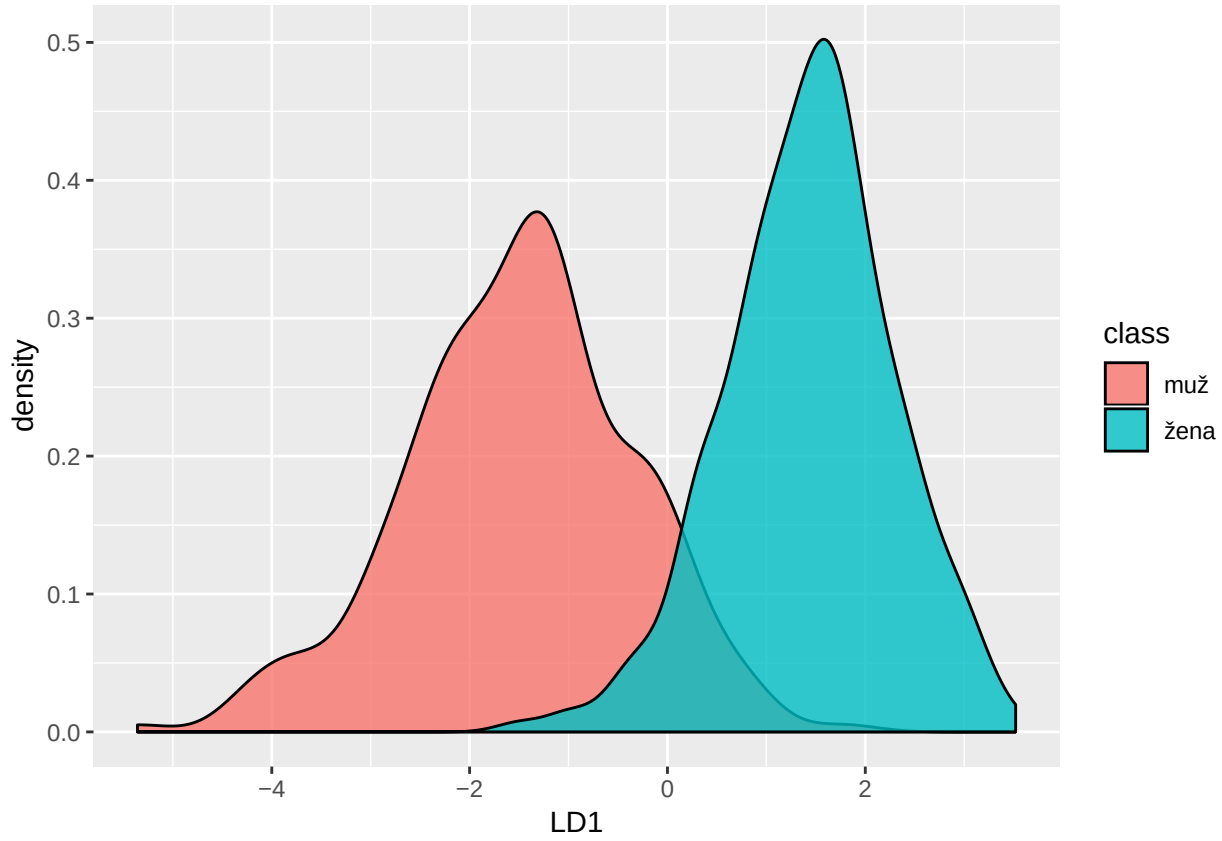


LDA

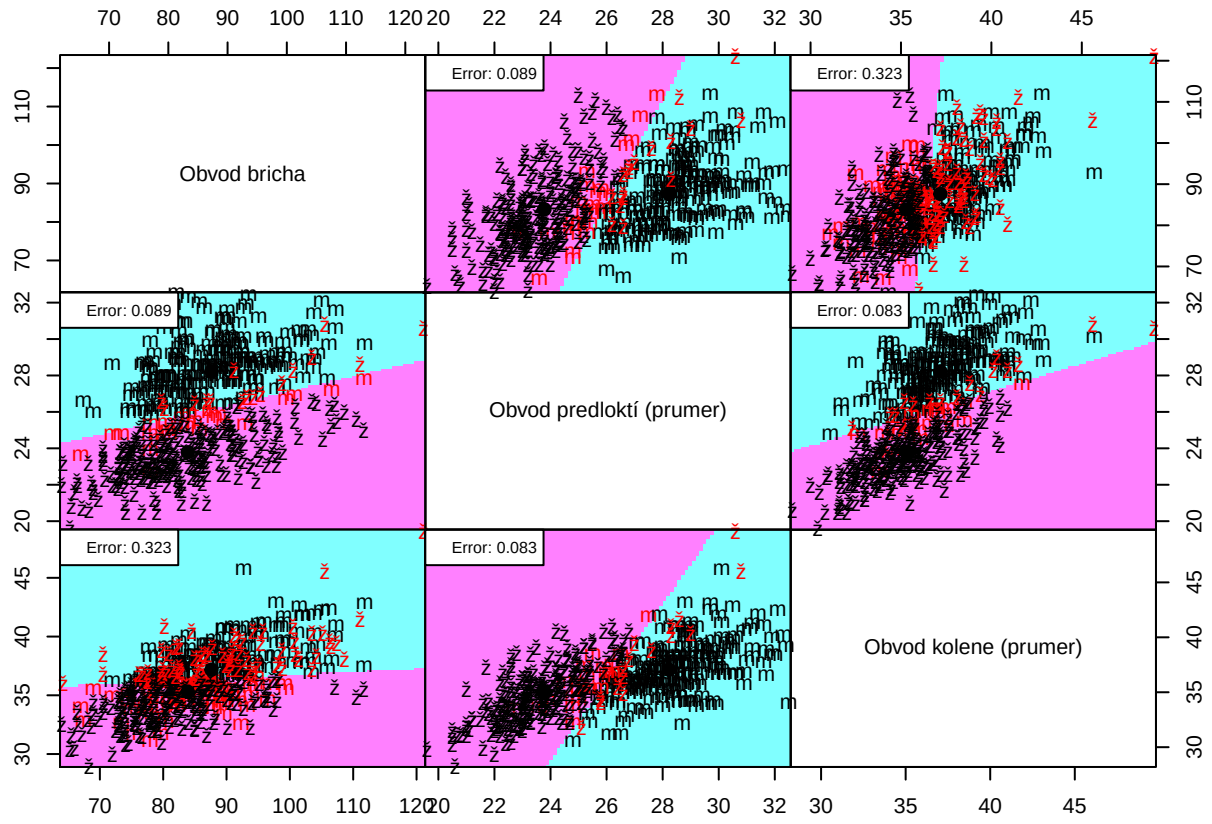
Nyní spusťme LDA na trénovacích datech a podívejme se na jednotlivé koeficienty u diskriminační funkce. Největší absolutní hodnota koeficientu je právě u obvodu předloktí, a sice $ 0.754 $. Nejhůře potom diskriminuje obvod břicha s absolutní hodnotou koeficientu $ 0.039 $.

$$LD_1 = (0.039 * o4) + (-0.754 * o8) + (0.148 * o9)$$

Zde vidíme jednotlivé hodnoty transformované pomocí lineární diskriminační funkce. Vidíme zde rozdělení pro muže a ženy a lze vidět, že po transformaci se překrývají jen v menší části uprostřed.



Zde můžeme vidět, jak lze pohlaví rozdělit na původním prostoru proměnných. Červeně jsou označeny instance, jež jsou nesprávně klasifikovány. Ze zběžného pohledu můžeme vidět, že většina instancí je správně klasifikována a pouze malá část jich je zabarvena červeně. To nám napovídá, že klasifikace má vysokou přesnost. Pojďme se nyní na tuto přesnost podívat v konkrétních číslech.



Ověření na trénovacích datech

Nyní predikujme hodnoty pohlaví za základě dat, s nimiž jsme spustili LDA a podívejme se na matici záměn. Vidíme, že na trénovací množině je odhad přesnosti LDA přibližně 93%. LDA chybně predikovalo 12 žen jako muže a 24 mužů jako ženy. Kromě toho se zde vyskytují údaje jako sensitivita a specificita. Sensitivita nám říká, že přibližně $ 90 % $ mužů bylo správně klasifikováno jako muži. Specificita nám pak říká, že přiblížně $ 95 % $ žen bylo správně klasifikováno jako ženy.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction muž žena
##       muž  218   12
##       žena  24  242
##
##                Accuracy : 0.9274
##                  95% CI : (0.9009, 0.9487)
##     No Information Rate : 0.5121
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8546
##
##  Mcnemar's Test P-Value : 0.06675
##
```

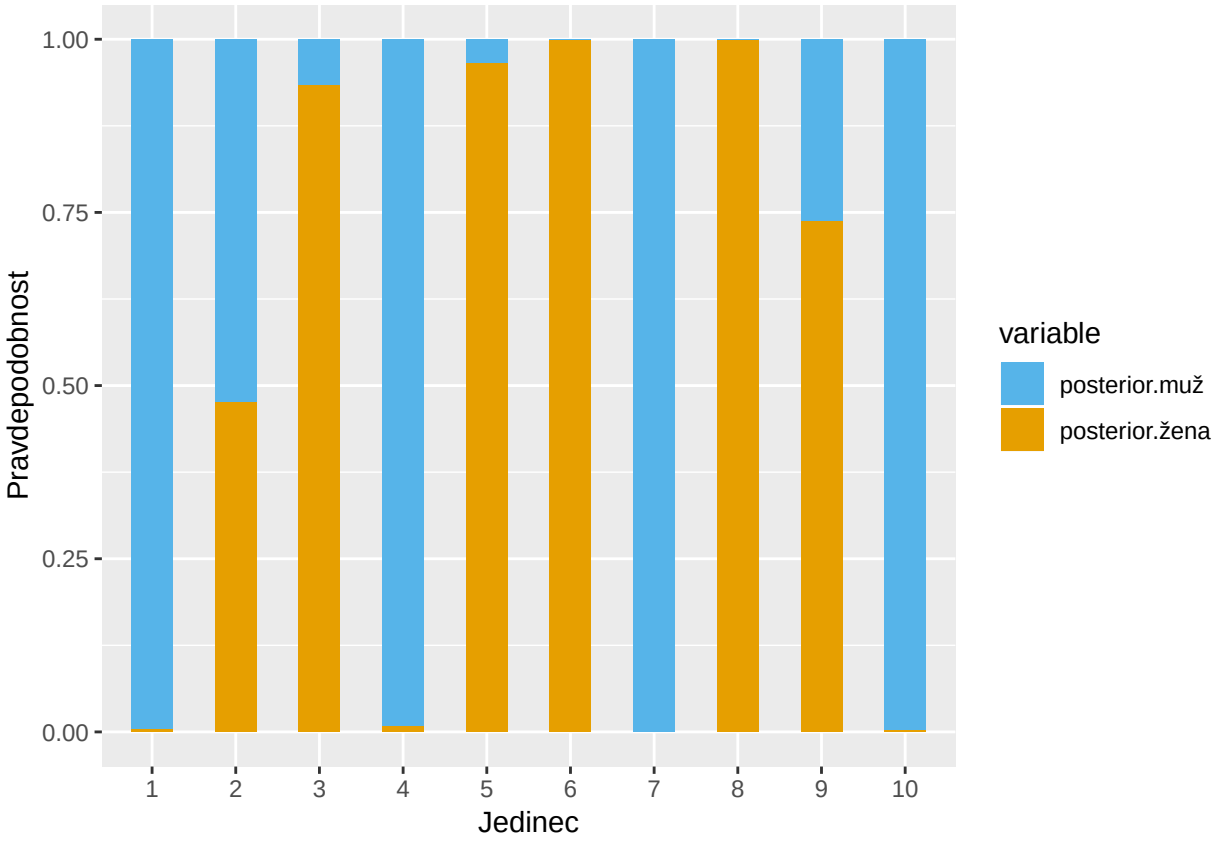
```
##             Sensitivity : 0.9008
##             Specificity : 0.9528
##          Pos Pred Value : 0.9478
##          Neg Pred Value : 0.9098
##              Prevalence : 0.4879
##          Detection Rate : 0.4395
##    Detection Prevalence : 0.4637
##       Balanced Accuracy : 0.9268
##
##        'Positive' Class : muž
##
```

Přesnost byla však vypočtena na trénovacích datech (na nichž jsme LDA spustili). Pro spolehlivější odhad přesnosti použijeme desetinásobnou křížovou validaci. Průměrná přesnost napříč deseti validacemi je přibližně $ 93 % $. Tedy stejně, jako kdybybychom použili k ověření jen trénovací data. Získali jsme tedy relativně přesný odhad. Máme jistotu přibližně $ 93 % $, že jedince správně klasifikujeme.

```
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction  muž žena
##       muž  43.3  2.4
##       žena  5.4 48.8
##
##  Accuracy (average) : 0.9214
```

Jistota klasifikace nových instancí

Následující graf zobrazuje, s jakou pravděpodobností náleží jedinec do dané třídy. Můžeme z něj vyčíst, zda si můžeme být klasifikací jistí (v případě, kdy jasně převládá velikost pravděpodobnosti jednoho z pohlaví), nebo ne (v případě, kdy jsou velikosti téměř vyrovnané).



Zde už vidíme zařazení jednotlivých jedinců do tříd (pohlaví) podle té pravděpodobnější z nich.

Jedinec	Pohlaví
1	muž
2	muž
3	žena
4	muž
5	žena
6	žena
7	muž
8	žena
9	žena
10	muž