Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси у информационом инжењерингу

Основи класификације

(материјали за предавања)

Садржај

1. Класификација
2. Регресија
3. Метод најближих суседа
4. Метод потпорних вектора
5. Стабла одлучивања
6. Извори и литература

Класификација

- задатак у којем треба утврдити којој од могућих класа припада посматрана појава
  - класа представља вредност обележја
  - утврђивање припадности класи се начелно изводи на основу вредности других одабраних обележја посматране појаве
  - резултат је одабрана класа
    - одабрана класа се може али и не мора поклапати са стварном класом којој припада посматрана појава

Класификација

- врсте класификације
  - бинарна
    - две могуће класе у разматрању
  - n-арна
    - n (више од две) могућих класа у разматрању

Класификација

Основне улоге обележја у класификацији

- циљно обележје
  - (зависно обележје, одговор)
  - обележје чију вредност за неку појаву треба одредити на основу вредности других обележја
  - циљно обележје је категоријско (квалитативно)
  - обично постоји једно циљно обележје
    - ознака Y
- предикторско обележје
  - (независно обележје, предиктор)
  - обележје чија се вредност за неку појаву користи у одређивању вредности циљног обележја
  - предикторско обележје може бити или категоријско (квалитативно) или нумеричко (квантитативно)
  - обично постоји више предикторских обележја
    - ознака $X = (X_1, X_2, \ldots, X_p)$

Класификација

Класификациони модел

(класификатор)

- модел који начелно успоставља везу између циљног обележја и предикторских обележја
  - може настати у поступку обучавања
- постојање варијација у врсти структуре и нивоу сложености
- сврха
  - предвиђање
  - разумевање

Класификација

Поступак обучавања

- формирање модела и подешавање параметара ради постизања што бољих перформанси у класификацији
- обучавање се изводи на основу расположивих података
  - за појаве заступљене у подацима познате су вредности циљног обележја

Класификација

Перформансе у класификацији

- перформансе могу варирати зависно од коришћених података
- метод који је напогоднији за један случај не мора бити најпогоднији за неки други случај
- тежити методи која обезбеђује ниску варијансу (енгл. *variance*) и ниску пристрасност (енгл. *bias*)
- теорема о непостојању бесплатног ручка

Класификација

Оцена перформанси

- грешка обучавања
  - проценат погрешно класификованих појава међу појавама које су коришћење током обучавања
- грешка тестирања
  - проценат погрешно класификованих појава међу појавама које нису коришћене током обучавања

Оцена перформанси

матрица конфузије – пример за бинарни случај

		Стварна класа	
		0 (не, није, нема, -)	1 (да, јесте, има, +)
Процењена класа	0 (не, није, нема, -)	број стварно негативних (енгл. *true negative*, *TN*)	број лажно негативних (енгл. *false negative*, *FN*)
	1 (да, јесте, има, +)	број лажно позитивних (енгл. *false positive*, *FP*)	број стварно позитивних (енгл. *true positive*, *TP*)

Оцена перформанси

- матрица конфузије – пример за бинарни случај
  - тачност
    - $(TP + TN) / (TP + TN + FP + FN)$
  - осетљивост
    - $TP / (TP + FN)$
  - специфичност
    - $TN / (TN + FP)$

Класификација

Процена грешке тестирања помоћу појава за обучавање

- валидациони приступ (енгл. *validation set approach*)
  - насумична подела расположивих појава у скуп за обучавање и валидациони скуп
    - модел бива подешаван на основу скупа за обучавање
    - подешени модел се примењује над валидационим скупом
      - рачуна се грешка над валидационим скупом

Класификација

Процена грешке тестирања помоћу појава за обучавање

- појединачна унакрсна валидација (енгл. *leave-one-out cross-validation*)
  - расположиве појаве се распоређују у валидациони скуп, који обухвата само једну појаву, и скуп за обучавање, који обухвата све остале појаве
    - модел бива подешаван на основу скупа за обучавање
    - подешени модел се примењује над валидационим скупом
      - рачуна се грешка над валидационим скупом
  - понавља се по истом принципу распоређивање појава у валидациони скуп и скуп за обучавање, при чему нека друга појава постаје једини елемент валидационог скупа
    - поново се по истом принципу изводе подешавање модела, примена модела и рачунање грешке
  - понављање се изводи док свака појава тачно једном не буде део валидационог скупа
  - по завршетку понављања рачуна се аритметичка средина за грешку над валидационим скупом

Процена грешке тестирања помоћу појава за обучавање

- *k*-тострука унакрсна валидација (енгл. *k-fold cross-validation*)
  - расположиве појаве се насумично распоређују у *k* скупова исте или приближно исте величине
  - за сваки појединачни скуп изводи се посебан поступак
    - појединачни скуп служи као валидациони скуп
    - сви остали скупови заједно служе као скуп за обучавање
    - модел бива подешаван на основу скупа за обучавање
    - подешени модел се примењује над валидационим скупом
      - рачуна се грешка над валидационим скупом
  - по завршетку поступка за сваки појединачни скуп, рачуна се аритметичка средина за грешку над валидационим скупом

Класификација

Методи класификације

- разноврсни методи на располагању
- примери метода које се могу користити у класификацији
  - регресија
    - логистичка регресија
  - метод најближих суседа
  - метод потпорних вектора
  - стабла одлучивања
    - појединачна
    - вишеструка

Класификација

- статистичко учење (енгл. *statistical learning*)
  - „велики скуп алата за разумевање података” (по Џејмсу и сарадницима)
- машинско учење (енгл. *machine learning*)
  - „аутоматска детекција смислених образаца у подацима” (по Шалев-Шварцу и Бен-Давиду)
  - врсте машинског учења
    - надгледано учење (енгл. *supervised learning*)
    - потпомогнуто учење (енгл. *reinforcement learning*)
    - ненадгледано учење (енгл. *unsupervised learning*)

Скупови података коришћени у примерима

- скуп података **`abalon`**
  - подаци о абалонима с Тасманије (Аустралија)
  - 4177 записа
  - 9 обележја
    - пол, дужина, пречник, висина, различите тежине и број прстенова
  - датотека *abalone.data*
    - Abalone Data Set (од 1. 12. 1995)
      - https://archive.ics.uci.edu/ml/datasets/Abalone
        - (преузето 13. 4. 2021)
    - UCI Machine Learning Repository
      - Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Скупови података коришћени у примерима

скуп података **abalon** – припрема

```
1  # install.packages("tidyverse")
2
3  library(readr)
4  library(dplyr)
5  library(magrittr)
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Скупови података коришћени у примерима

скуп података **abalon** – припрема

```
abalon <- read_csv("abalone.data",
                   col_names=c("sex", "length", "diameter",
                               "height", "weight_whole",
                               "weight_shucked", "weight_viscera",
                               "weight_shell", "rings"),
                   col_types="fddddddddi")

abalon %<>%
  mutate(id=1:nrow(abalon), age=rings + 1.5) %>%
  select(id, everything())

abalon %<>%
  mutate(age_cat=factor(ifelse(age < median(age), "young", "old"),
                        levels=c("young", "old")))
```

УЛАЗ

Скупови података коришћени у примерима

скуп података **abalon** – припрема

```
 1 set.seed(5)
 2
 3 abalon.tst <- slice_sample(abalon, prop=0.2)
 4 abalon.trn <- setdiff(abalon, abalon.tst)
 5 nrow(abalon.tst)
 6 nrow(abalon.trn)
 7
 8
 9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Садржај

1. Класификација
2. **Регресија**
3. Метод најближих суседа
4. Метод потпорних вектора
5. Стабла одлучивања
6. Извори и литература

Линеарна регресија

општи облик

једнострука линеарна регресија

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

вишеструка линеарна регресија

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

елементи

- циљно обележје
- предикторска обележја
- параметри (коефицијенти)
- случајна грешка

Регресија

Линеарна регресија

- потребно спровести оцењивање параметара
  - може се искористити метод најмањих квадрата
- примарно погодна за нумеричка циљна обележја
- потребно проверити испуњеност више предуслова за примену

Линеарна регресија

пример

```
1  library(ggplot2)
2
3  linr <- lm(age ~ diameter + weight_whole, data=abalon)
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Линеарна регресија

пример

```
> summary(linr)

Call:
lm(formula = age ~ diameter + weight_whole, data = abalon)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2786 -1.6905 -0.7151  0.8972 15.9518

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.2254     0.2858  14.783   <2e-16 ***
diameter      16.8749     1.0858  15.541   <2e-16 ***
weight_whole   0.3925     0.2197   1.786   0.0741 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.638 on 4174 degrees of freedom
Multiple R-squared:  0.3307,  Adjusted R-squared:  0.3304
F-statistic:  1031 on 2 and 4174 DF,  p-value: < 2.2e-16
```

КОНЗОЛА

Линеарна регресија

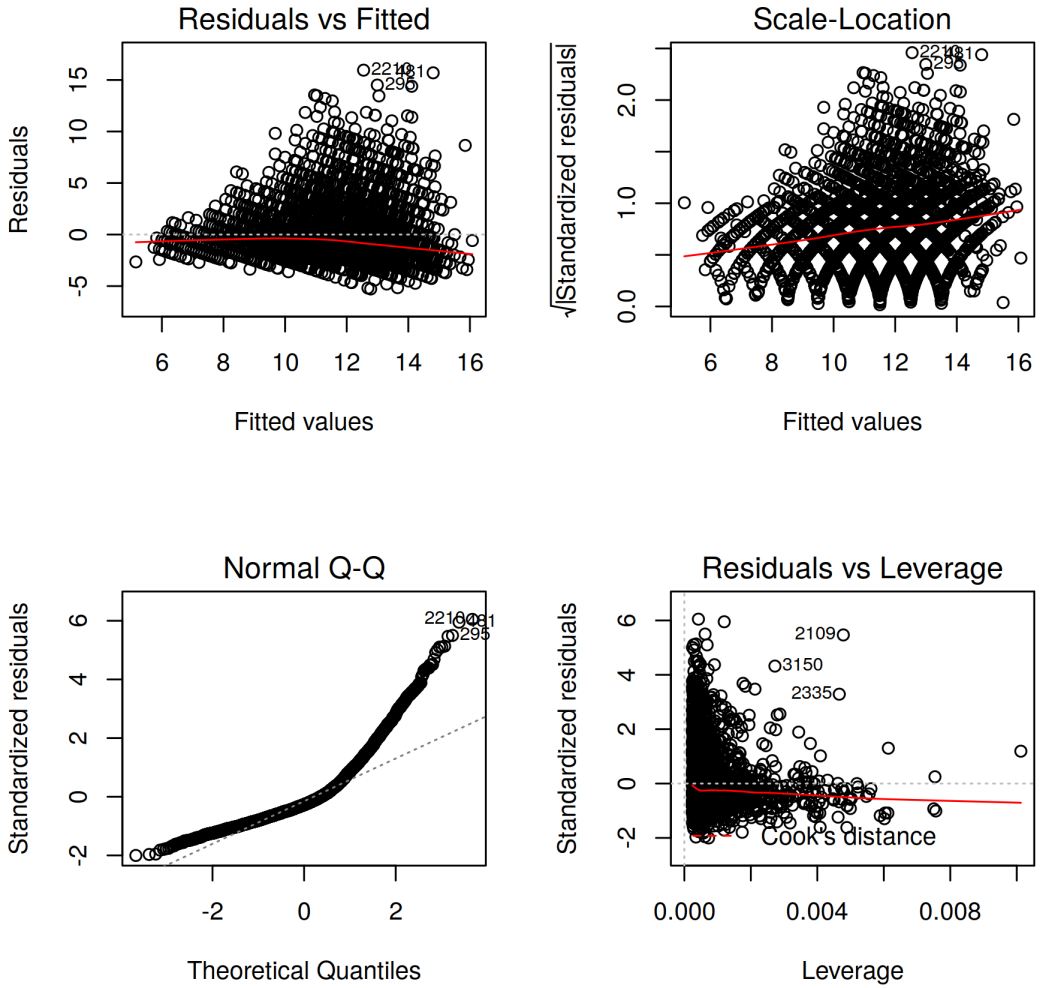
пример

```
> select(filter(abalon, id==1), age, diameter, weight_whole)
# A tibble: 1 x 3
    age diameter weight_whole
  <dbl>    <dbl>        <dbl>
1  16.5    0.365        0.514
> predict(linr, newdata = abalon[1,])
       1
10.58652
>
```

КОНЗОЛА

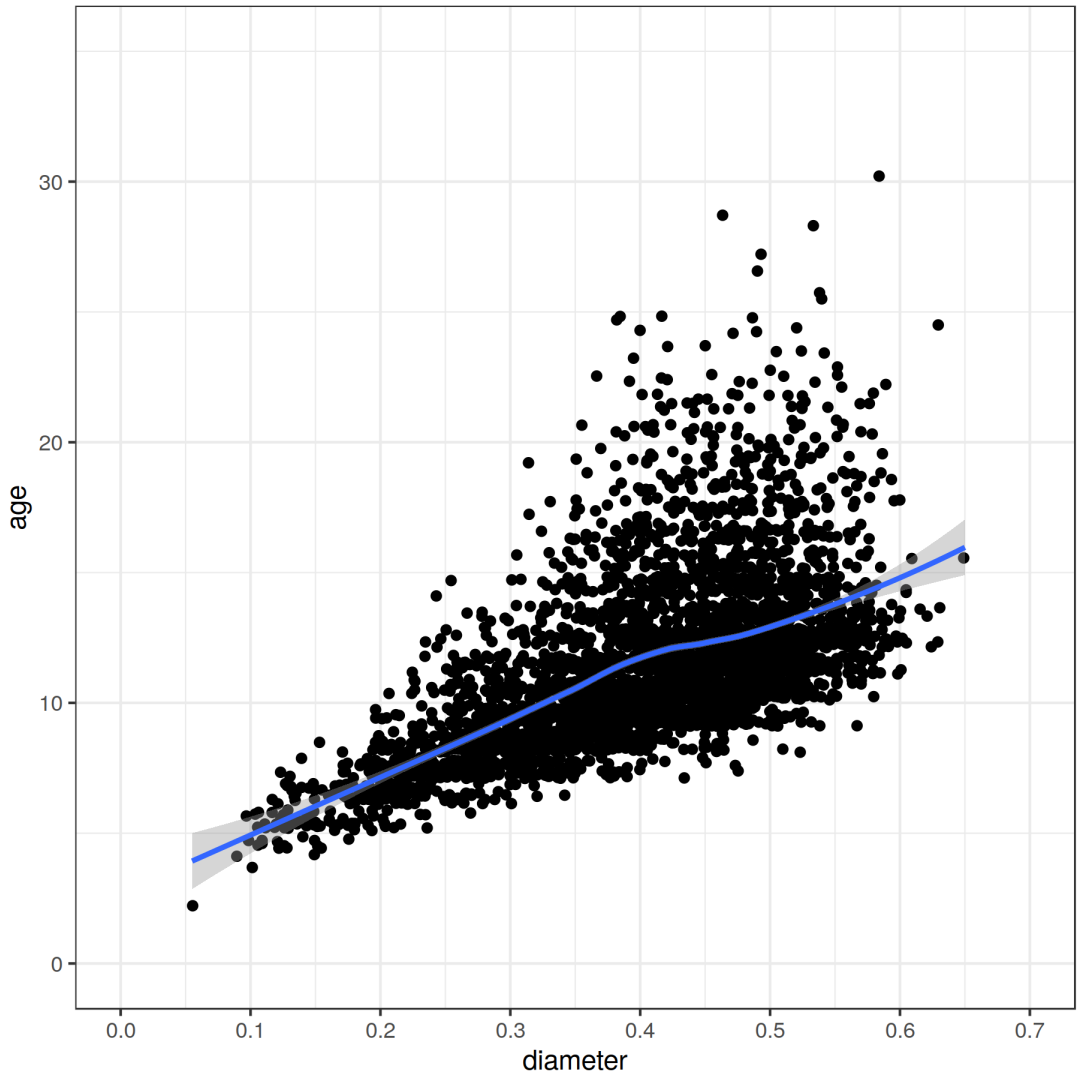
Линеарна регресија

пример



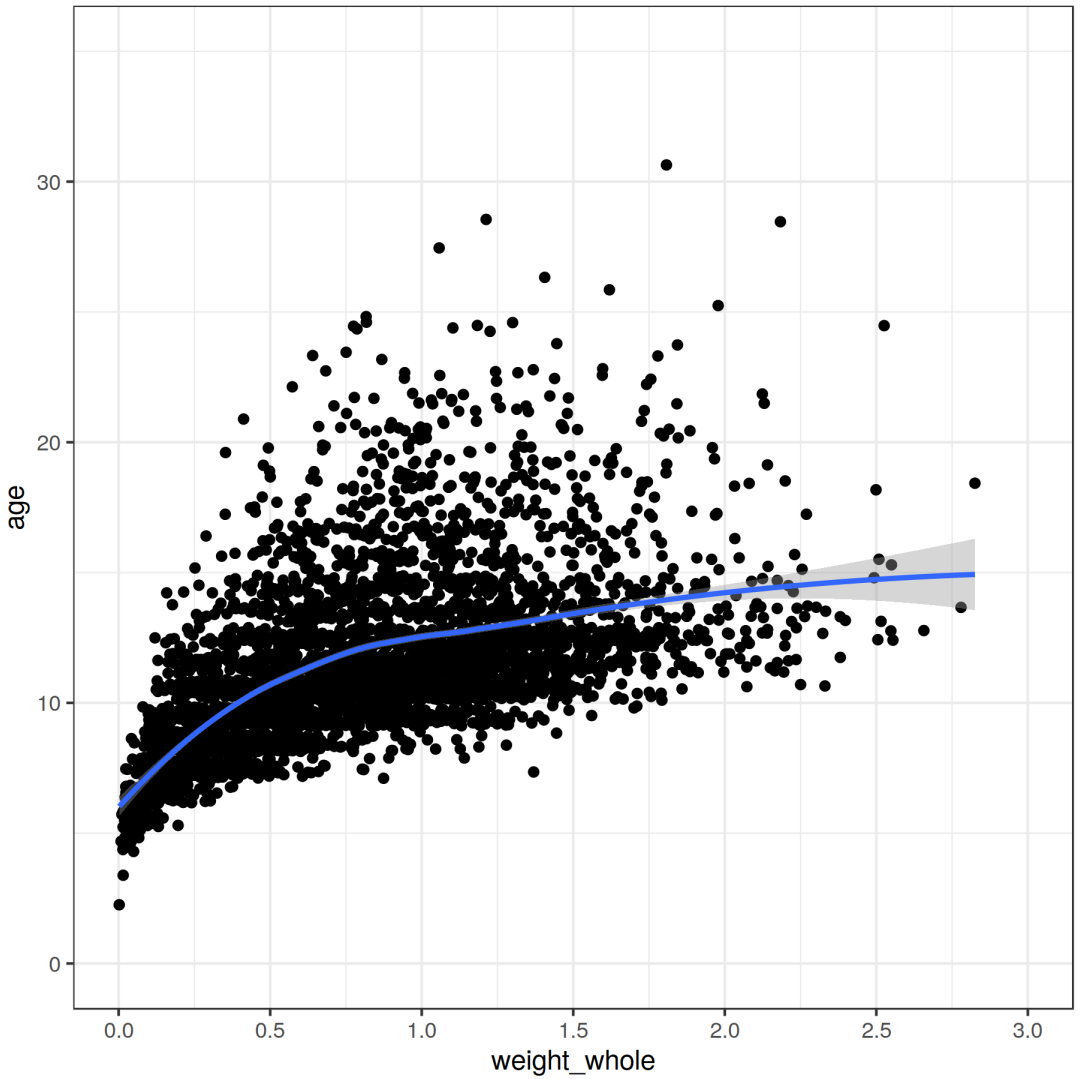
Линеарна регресија

пример



Линеарна регресија

пример



Логистичка регресија

општи облик

једнострука логистичка регресија

$$p(X) = Pr(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

вишеструка логистичка регресија

$$p(X) = Pr(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}}$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Регресија

Логистичка регресија

- потребно спровести оцењивање параметара
  - може се искористити метод максималне веродостојности
- погодна за категоријска циљна обележја
- потребно проверити испуњеност више предуслова за примену

Логистичка регресија

пример

```
logr <- glm(age_cat ~ diameter + weight_whole, data=abalon.trn,
            family="binomial")

ver.logr <- predict(logr, newdata=abalon.tst, type="response")

klas.logr <- factor(ifelse(ver.logr > 0.5, "old", "young"),
levels=c("young","old"))
```

УЛАЗ

Логистичка регресија

пример

```
> summary(logr)

Call:
glm(formula = age_cat ~ diameter + weight_whole, family = "binomial",
    data = abalon.trn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8053  -0.5798   0.3073   0.6247   2.2990

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.3649     0.4516 -11.879  < 2e-16 ***
diameter      12.6844     1.7516   7.241 4.44e-13 ***
weight_whole   1.5411     0.3876   3.976 7.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

КОНЗОЛА

Логистичка регресија

пример

```
> head(ver.logr)
        1         2         3         4         5         6
0.5711173 0.9864490 0.9856813 0.9726682 0.1762120 0.9339484
> head(klas.logr)
    1     2     3     4     5     6
  old   old   old   old young   old
Levels: young old
>
```

КОНЗОЛА

Логистичка регресија

пример

```
1  library(caret)
2
3  cm.logr.tst <- confusionMatrix(data=klas.logr,
4  reference=abalon.tst$age_cat)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Логистичка регресија

пример

```
> cm.logr.tst
Confusion Matrix and Statistics

          Reference
Prediction young old
     young   177  53
     old      99 506

               Accuracy : 0.818
                 95% CI : (0.7901, 0.8436)
    No Information Rate : 0.6695
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5706

 Mcnemar's Test P-Value : 0.0002623

            Sensitivity : 0.6413
            Specificity : 0.9052
```

КОНЗОЛА

Логистичка регресија

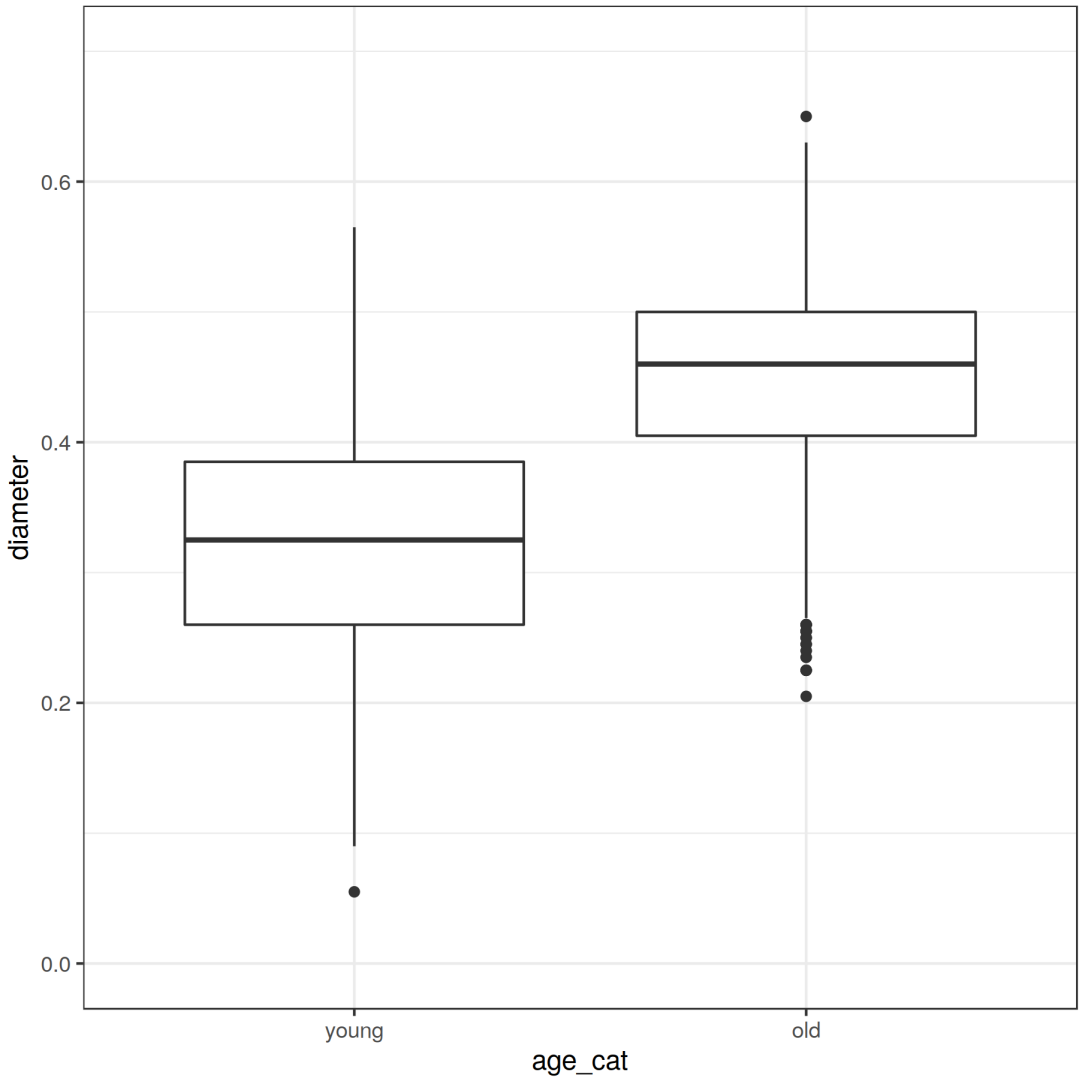
пример

```
> cm.logr.tst$overall["Accuracy"]
[1] 0.8179641
>
```

КОНЗОЛА

Логистичка регресија

пример



Логистичка регресија

пример