Exercises L7

2022-06-13

Exercise 1

Note: it is often convenient to use the "data" parameter of the lm function to build models, as this makes it easier to use "predict". For example here we train a model with the first 50 observations (flowers, individuals) of the "iris" dataset (training set) and use it to predict pepal length in the remaining 100 observations (testing set).

```
lm_length_width <- lm(Petal.Length ~ Petal.Width, data = iris[1:50, ])
pred <- predict(lm_length_width, newdata = iris[51:150,])
```

1. Explain why this is *not* a good choice of training and testing for this dataset

The training set has not been chosen wisely since it contains flowers of the species "Setosa" while the testing set contains flowers of the other two species. A better training set would have been obtained by randomly sampling 50 flowers from the whole dataset.

2. Show that indeed the R^2 on the "testing" set is actually negative, implying that the prediction error is greater than the variance of the testing set (compare with the formula for R^2 in the previous lecture)

```
compute_r2 <- function(y, y_pred){
  1 - sum((y - y_pred)^2)/sum((y - mean(y))^2)
}
```

```
r2 <- compute_r2(iris[51:150, "Petal.Length"], pred)
```

Exercise 2

```
library(MASS)
```

The diabetes data of the Pima dataset (from the library "MASS") are randomly split into a training set ("Pima.tr", with 200 subjects) and a test set ("Pima.te", with 332 subjects).

```
dim(Pima.tr)
```

```
## [1] 200   8
```

```
dim(Pima.te)
```

```
## [1] 332   8
```

1. Use the training set (Pima.tr) to generate:

- a model in which age predicts glucose level
- a model in which age and BMI predict glucose level

```
age_glu <- lm(glu ~ age, data = Pima.tr)
age_bmi_glu <- lm(glu ~ age + bmi, data = Pima.tr)

find_r2 <- function(y, y_pred){
  1 - sum((y - y_pred)^2)/sum((y - mean(y))^2)
}
```

Based on the training data only, does BMI add useful information with respect to age only?

2. Does the full model (age+BMI) predict glucose level better (in terms of R^2) than the age-only model also in the test set?

```
pred_age <- predict(age_glu, newdata = Pima.te)
r2_age <- find_r2(Pima.te$glu, pred_age)
r2_age
```

```
## [1] 0.02610344
```

```
pred_age_bmi <- predict(age_bmi_glu, newdata = Pima.te)
r2_age_bmi <- find_r2(Pima.te$glu, pred_age_bmi)
r2_age_bmi
```

```
## [1] 0.08743608
```

Yes, the addition of BMI increases the R^2 in the testing set from 0.026 to 0.087.

3. Now add also blood pressure (“bp”) to the model developed on the training set. Does blood pressure provide useful additional information on glucose levels? How does the predictive power of this model on the test set compare with the model without blood pressure?

```
age_bmi_bp_glu <- lm(glu ~ age + bmi + bp, data = Pima.tr)

pred_age_bmi_bp <- predict(age_bmi_bp_glu, newdata = Pima.te)
r2_age_bmi_bp <- compute_r2(Pima.te$glu, pred_age_bmi_bp)
r2_age_bmi_bp
```

```
## [1] 0.08045386
```

The addition of bp actually decreases the testing set R^2 to 0.080.