Data Mining 2022/2023
List of exercises

I use the following abbreviations to indicate the source of an exercise:

1. ISL = An Introduction to Statistical Learning by G. James et al.
2. ESL = The Elements of Statistical Learning by T. Hastie et al.
3. ...

1 Introduction

Exercise 1 — Suppose that random variables X and Y are independent. Prove that (1p)

(a) $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$,

(b) $\mathbb{V}\mathrm{ar}(X+Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y)$.

Exercise 2 — Recall what is the bias and variance of an estimator. Give examples of biased and unbiased estimators. (1p)

Exercise 3 — (ESL, p. 223) Suppose that we have a training set of points $(x_1, y_1), \ldots, (x_n, y_n)$. Assume there is an underlying relation $y = f(x) + \epsilon$, where ϵ represents noise and is a random variable with zero mean and variance σ_ϵ^2. We use the training set to find $\hat{f}(x)$ that approximates $f(x)$. Show that we can decompose expected squared error at a new input x_0 as:

$$\mathrm{E}\Big[\big(y_0 - \hat{f}(x_0)\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(x_0)\big]^2 + \mathrm{Var}\big[\hat{f}(x_0)\big] + \sigma_\epsilon^2 \ .$$

What is the bias–variance tradeoff? (3p)

Exercise 4 — (ISL) For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer. (1p)

(a) The sample size n is extremely large, and the number of predictors p is small.

(b) The number of predictors p is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e. $\sigma^2 = \mathbb{V}\mathrm{ar}(\epsilon)$, is extremely high.

Exercise 5 — (ISL) Provide a sketch of typical squared bias, variance, training error, test error and irreducible error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. Explain the shape of each curve. (1p)

Exercise 6 — (ISL) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide the sample size n and the number of predictors p. (1p)

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

(c) We are interesting in predicting the percent change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the percent change in the dollar, the percent change in the US market, the percent change in the British market, and the percent change in the German market.

Exercise 7 — (ISL) You will now think of some real-life applications for statistical learning. (1p)

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) Describe three real-life applications in which cluster analysis might be useful.

Exercise 8 — (ISL) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred? (1p)

Exercise 9 — (ISL) Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages? (1p)

Exercise 10 — (ISL) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X1 = X2 = X3 = 0$ using K-nearest neighbors. (1p)

(a) Compute the Euclidean distance between each observation and test point $X1 = X2 = X3 = 0$.

(b) What is our prediction with K = 1? Why?

(c) What is our prediction with K = 3? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

2 Linear Regression

Exercise 11 — Assume we have n observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we consider a linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. We estimate parameters β_0 and β_1 by minimizing mean squared error:

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x))^2 \ .$$

Show that in such a case

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \ ,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are sample means. Argue that the obtained line always passes through the point $(\bar{x}, \bar{y})$. (2p)

Exercise 12 — Derive the bias, variance and standard error for estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. We assume that $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and all ε_i for $i \in \{1, \ldots, n\}$ are independent. (2p)

Exercise 13 — Recall how to prove that the sum of two independent normally distributed random variables is normally distributed. (2p)

Exercise 14 — Explain why there is approximately a 95% chance that the interval

$$\hat{\beta}_1 \pm 2\sqrt{\mathbb{V}\mathrm{ar}\left(\hat{\beta}_1\right)}$$

contains the true value of β_1. (2p)

Exercise 15 — Recall that for $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ we define R^2 as

$$R^2 = 1 - \frac{RSS}{TSS} \ .$$

What's the interpretation for R^2? Show that if we consider a model $Y = \beta_0 + \beta_1 X + \varepsilon$ we have

$$R^2 = Corr(X, Y)^2 \ ,$$

where $Corr(X, Y)$ is correlation coefficient. (3p)

Exercise 16 — Recall what's t-statistic and how we can use it in the context of linear regression. What's p-value? (3p)

Exercise 17 — Show that for a linear regression model with $k+1$ parameters we can obtain estimations of

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$$

as

$$\hat{\beta} = (XX^T)^{-1} X^T \vec{y} \ ,$$

where X is data matrix and $\vec{y}$ is vector of responses (see e.g. here). (3p)

3 Classification

Exercise 18 — Explain what are the elements of the boxplot. (1p)

Exercise 19 — Explain what is Naive Bayes classifier. Explain and prove the following formula:

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{p(\mathbf{x})} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

where

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid C_k) p(C_k) \quad \text{and} \quad \mathbf{x} = (x_1, \ldots, x_n) \ . \ (1p)$$

Exercise 20 — (ISL) When the number of features p is large, there tends to be a deterioration in the performance of k-nearest neighbors (KNN) and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse. (1p)

a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction?

b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make the prediction?

c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1$, 2 and 100, what is the length of each side of the hypercube? Comment on your answer.

⋮

J.L.