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:

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

where

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

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.

Exercise 21 — Assume we have single predictor X and binary response Y and we would like to create a parametric model for $p(X) = Pr(Y = 1|X)$. (1p)

a) We might try using the linear regression model. Why it is not very good idea?

b) In the binary logistic regression algorithm we model the probability $p(X)$ with the logistic function

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

Prove that above equation is equivalent to the following log-odds (logit) representation:

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

What's the connection between the logistic $\sigma(x) = \frac{e^x}{1+e^x}$ and logit $l(p) = \ln\left(\frac{p}{1-p}\right)$ function?

Exercise 22 — (ISL) This problem has to do with odds. (1p)

a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

Exercise 23 — (ISL) Suppose we collect data for a group of students in a statistics class with variables X_1="hours studied", X_2="average grade", and Y = "receive 5.0". We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$. (1p)

a) Estimate the probability that a student who studies for $40h$ and has an average grade 3.5 gets 5.0 in the class.

b) How many hours would the student in part a) need to study to have a 50% chance of getting 5.0 in the class?

Exercise 24 — Explain how we may get the multinomial (multi-class) logistic model for K classes by running $K - 1$ independent binary logistic regression models. Hint: see here. (1p)

4 Maximum likelihood estimation, cross-entropy, tree models

Exercise 25 — Recall what is a maximum likelihood estimator and what you can say about its consistency and efficiency. (1p)

Exercise 26 — Assume that you have n observations $x_1, \ldots, x_n$ from the normally distributed random variable $X \sim N(\mu, \sigma^2)$ with unknown parameters μ and σ^2. Derive maximum likelihood estimator for the parameter μ. (1p)

Exercise 27 — Recall what is entropy, cross entropy and Kullback–Leibler divergence. Explain how these three measures are connected to each other. (1p)

Exercise 28 — During the lecture we have shown that for binary classification problems the log-likelihood function can be expressed in terms of cross-entropy. Show the similar result for multi-class classification problems. You may look at this hint. (1p)

Exercise 29 — For decision trees we usually use entropy or Gini index as the loss function. Explain how the entropy and Gini index can be interpreted. See e.g. ESL book, page 310. (1p)

Exercise 30 — (ESL, page 309) Suppose we use a decision tree for two-class problem with 400 observations in each class (denote this by (400, 400)) and suppose one split created nodes (300, 100) and (100, 300), while the other created nodes (200, 400) and (200, 0). Both splits produce a misclassification rate of 0.25, but the second split produces a pure node and is probably preferable. Show that both the Gini index and cross-entropy are lower for the second split. (1p)

Exercise 31 — Suppose we try to use a decision tree to data with a categorical predictor having q possible unordered values. What might be the problem? Hint: show there are $2^{q-1}-1$ possible partitions of the q values into two groups. (1p)

Exercise 32 — Provide a pseudo-code of Random Forest algorithm and its detailed explanation (for both classification and regression problem). (1p)

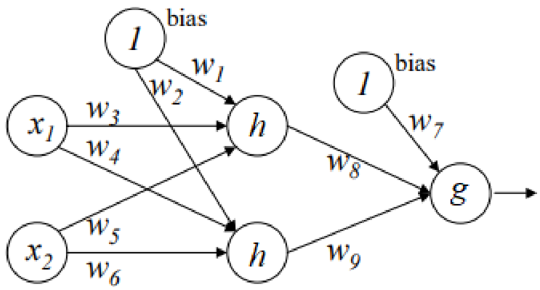
Exercise 33 — (ISL) Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $Pr(Class\ is\ Red|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7 \ and\ 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed earlier. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches? (1p)

5 Neural Networks

Exercise 34 — Suppose we solve a binary classification problem using a neural network with one hidden layer as shown in the figure below. The network models probability $P(y = 1|\mathbf{x})$, where $y \in \{0, 1\}$ and $\mathbf{x} = (x_1, x_2)$. At hidden units we use a linear activation function $h(z) = c \cdot z$ with constant c. At the output unit we use a sigmoid activation function $g(z) = \frac{1}{1+e^{-z}}$.



1. Express the output probability from the above neural network in terms of x_i, w_i and c.
2. Express classification decision boundary as an equation in terms of x_i, w_i and c.
   Is this decision boundary linear or non-linear in terms of input values x_1 and x_2?
3. Explain how you can use cross-entropy to evaluate the above neural network. (1p)

Exercise 35 — Draw a neural net with no hidden layer which is equivalent to the neural net given in the previous exercise. Express weights $v_1, v_2, \ldots$ of this new neural net in terms of weights w_i and c.

Can any multi-layered neural net with linear activation functions at hidden layers be represented as a neural net without any hidden layer? Justify your answer. (1p)

Exercise 36 — Assume we transform some vector $\vec{z} = (z_1, \ldots, z_n)$ using softmax function and we get vector $\vec{p} = (p_1(\vec{z}), \ldots, p_n(\vec{z}))$ where

$$p_j(\vec{z}) = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}}.$$

1. Describe the Jacobian matrix of the softmax function.
2. Show that $\frac{\partial p_j}{\partial z_k}$ is positive if $j = k$ and negative if $j \neq k$. What are the consequences? (1p)

Exercise 37 — Assume we transform a vector $\vec{z} = (z_1, \ldots, z_n)$ using softmax function and we get vector $\vec{p} = (p_1(\vec{z}), \ldots, p_n(\vec{z}))$. Assume also that we have some one-hot encoded vector $\vec{y}$ and we would like to minimize the cross-entropy loss function with respect to $\vec{z}$:

$$L(\vec{z}) = -\sum_{k=1}^{n} y_j \log p_j(\vec{z}) \ .$$

Find the formula for $\frac{\partial L}{\partial z_i}$. Recall how we can use this to iteratively minimize loss function. (1p)

Exercise 38 — Rectifier is an activation function defined as $f(x) = max(0, x)$. The unit in the neural network employing the rectifier is called a Rectified Linear Unit (ReLU).

1. What you can say about the derivative of $f(x)$?
2. A smooth approximation to the rectifier is softplus function defined as $s(x) = \ln(1 + e^x)$. Sketch a graph of $s(x)$ and find its derivative.
3. In Deep Learning book in Section 6.3.3 it is written that:

   „The use of the softplus is generally discouraged. The softplus demonstrates that the performance of hidden unit types can be very counterintuitive—one might expect it to have an advantage over the rectifier due to being differentiable everywhere or due to saturating less completely, but empirically it does not."

   What does it mean that function is saturated? What problem saturation can cause for the learning process? How you could deal with the saturation of rectifier? See e.g. Leaky ReLU. (1p)

J.L.