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.

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)

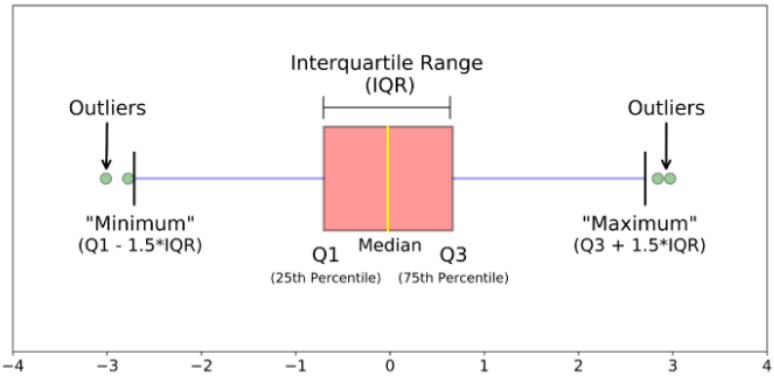
a) Estimate the probability that a student who studies for 40h and has an average grade 3.5 gets 5.0 in the class.

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. (2p)

Exercise 18

We can call a boxplot a standardized method of showing the distribution of specific data, using five numbers.



So in turn, they are:

1) "Minimum"

2) First quartile (Q1)

3) Median (Q2)

4) Third quartile (Q3)

5) "Maximum".

This is used to check what outliers are and their values. Boxplot also allows you to verify that the data is symmetrical, how tightly it is grouped, and how the data is skewed.

Let's define what each elements of boxplot are.

a) The "Minimum" should be considered the lowest data point excluding any outliers given (Q0 or 0th percentile).
b) First quartile, is the middle number between the smallest number (excluding "Minimum") and the median of the data set (Q1 or 25th percentile).
c) Median, it's the mid-point of the data, on the image above it's shown as the line that divide the box into two parts, it's then the middle value of the data set (Q2 or 50th percentile).
d) Third quartile, is the middle number between the median and the highest value (excluding "Maximum") of the data set (Q3 or 75th percentile).
e) The "Maximum" is the highest data point excluding any outliers given (Q4 or 100th percentile).
f) Whiskers, on the image above shown as those blue lines, the upper and the lower whiskers represent scores outside of the middle 50th percentile (i.e. the lower 25% of the scores and the upper 25% of the scores).
g) The Interquartile Range (IQR), is the boxplot showing the middle 50% of the scores, in the example above the range between the 25th and 75th percentile.
h) Outliers, on the image above shown as the green dots, they are an observations that are numerically distant from the rest of the data in the entire data set.

Exercise 19

As Naive Baies Classifier is a classification technique based on Bayes Theorem with an assumption of independence among used predictors (that's why it's "Naive"). Naive Bayes Classifier assumes that the presence of particular feature in a class is completely unrelated to the presence of any other feature. This is the easiest to understand on example:

We have a vegetable that may be considered an carrot, becouse it's orange, have oblong shape and it's length is around 12 cm. Even if those features depend on each other or upon existence of the other features, all of those properties independently contribute to the probability that this vegetable is in fact a carrot.

Naive Bayes model is known for it's simplicity, ease of construction and being particularly useful for very large data set.

Bayes theorem provides a way of calculating posterior probability $p(C_k|x)$ from $p(C_k)$, $p(x)$ and $p(x|C_k)$. Equation:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

$$p(C_k|X) = p(x_1|C_k) \times p(x_2|C_k) \times \ldots \times p(x_n|C_k) \times p(C_k)$$

$p(C_k|x)$ is the posterior probability of class (C_k, target) given predictor (x, attributes).

$p(C_k)$ is the prior probability of class.

$p(x|C_k)$ is the likelihood which is the probability of predictor given class.

$p(x)$ is the prior probability of predictor.

Looking further based on above equation and on a fact that Bayes theorem states the following relationship, that given class variable C and dependent feature vector x_1 through x_n:

$$p(C_k|x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n|C_k)p(C_k)}{p(x_1, \dots, x_n)}$$

We are using the "Naive" conditional independence assumption:

$$p(x_i|C_k, x_1, \dots, x_{i-1}, x_{i+1}, x_n) = p(x_i|C_k)$$

For all the i, this relationship will be simplified to:

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

We know that the $p(x_1, \dots, x_n)$ is a constant given input, then we can use following classification rule:

$$p(C_k|x_1, \dots, x_n) \propto p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

Where $\propto$ denotes proportionality, which means that we have here independence assumption, so the conditional distribution over class variable C is:

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

Where $p(X)$ is a scaling factor dependent only on the x values.

Exercise 20

a) We know that 0.1 represents a fraction of 10% the case if $x \in [0.05, 0.95]$, for the observations we will use the interval if $[x - 0.05,\ x + 0.05]$ and consequently respresents a length of 0.1. If the value of $x < 0.05$, for the obervations we will use the interval $[0,\ x + 0.05]$ that represents a fraction of $[100x + 5]\%$. And the last possiblity is that the value of $x > 0.95$, then the fraction will be $[105 - 100]\%$. The full calculation of the average fraction will be:

$$\int_{0.05}^{0.95} 10dx + \int_{0}^{0.05} (100x + 5)dx + \int_{0.95}^{1} (105 - 100x)dx = 9 + 0.375 + 0.375 = 9.75$$

9.75% is the fraction of available observations that will be used to make the prediction.

b) For this one we need to assume that the $X_1$and X_2 are independent, then the fraction of available observations used to make prediction will be $9.75\% * \ 9.75\% = 0{,}0975 * \ 0{,}0975 = \ 0{,}00950625 * 100\% = 0{,}950625\%$

c) According to the arguments with the previous tasks, we may conclude that the fraction of available observations used to make prediction will be $9.75\%^{100} \simeq 0$

d) According to previous tasks, we know that the fraction of available observations used to make predicions is $(9.75\%)^p$, with the p as the number of features. For $p \ \to \infty$ we will have then:

$$\lim_{p \to \infty} (9.75\%)^p = 0$$

e) For $p = 1$, we will have $l = 0.1$, for $p = 2$ the $l = 0.1^{0.5}$, and for $p = 100$ we will have $l = 0.1^{0.01}$

Exercise 21

a) Linear regression model is based on assumption that outcome Y is continuous, with some errors (after removing systematic variation in mean due to covariates $X_1, \dots, X_n$). If the outcome Y is binary, overall inferences will be invalid.

With binary data the variance is a function of the mean, and in particular is not constant as the mean changes. This violates one of the standard linear regression assumptions that the variance of the residual errors is constant.

For a binary outcome the mean is the probability of a 1, or success. If we use linear regression to model a binary outcome it is entirely possible to have a fitted regression which gives predicted values for some individuals which are outside of the $(0{,}1)$ range or probabilities.

b) We have then logistic function:

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

Let's suppose that we want to predict probability, and the dependent variable is binary. Then the calculations for odds will be:

$$odds = \frac{p}{1 - p}$$

In logistic regression, the dependent variable is logit, which is in fact the natural logarithm of the odds:

$$log(odds) = logit(p) = ln\left(\frac{p}{1 - p}\right)$$

Then the logit is a log of odds and the odds are a function of p. In logistic regression, we find that:

$$logit(p) = \beta_0 + \beta_1 X$$

We assume that the logit is linearly related to X.

If we take the log out from both sides of the equation $log(odds) = logit(p) = ln\left(\frac{p}{1-p}\right)$, we can convert odds to a simple probability:

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

Then:

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

What is the conection between $\sigma(x) = \frac{e^x}{1+e^x}$ and $logit(p) = ln\left(\frac{p}{1-p}\right)$?

$\sigma(x) = \frac{e^x}{1+e^x}$ is standard logistic function, and from the mathematical point of view, the $logit(p)$ is the inverse of it:

$$logit(p) = \sigma^{-1}(p) = ln\left(\frac{p}{1-p}\right) \text{ for } p \in (0,1)$$

Exercise 22

a) We need to calculate what fraction of people on average, with an odds of 0.37 of defaulting on their credit card paymen,t will in fact default.
Then:

$$\frac{p(X)}{1-p(X)} = 0.37$$
$$p(X) = \frac{0.37}{1+0.37} = 0.27$$

On average then a fraction of 27% people will default on their credit card.

b) In the next task we suppose that individual has 16% chance of defaulting on her credit card payment. And we need to calculate the odds that she will default.
Then:

$$\frac{p(X)}{1-p(X)} = \frac{0.16}{1-0.16} = 0.19$$

So the odds that she will default is 19%.

Exercise 23

a) We will use the equation for predicted probability for given values:

$$\hat{p}(X) = \frac{e^{(\beta+\alpha_1 X_1+\cdots+\alpha_n X_n)}}{1+e^{(\beta+\alpha_1 X_1+\cdots+\alpha_n X_n)}}$$

$$\hat{p}(X) = \frac{e^{(-6+0.05*X_1+X_2)}}{(1+e^{(-6+0.05*X_1+X_2)})} = 0.3775$$

0.3775 is the probability that the student who studied for 40 hours and has average grade of 3.5 will get 5.0 in the class.

b)

$$\hat{p}(X) = \frac{e^{(-6+0.05*X_1+3.5)}}{(1+e^{(-6+0.05*X_1+3.5)})} = 0.5$$

From that we can assume that $e^{(-6+0.05*X_1+3.5)} = 1$, and by taking logarithm by both sides:

$$X_1 = \frac{2.5}{0.05} = 50$$

The student need to study 50 hours then.

Exercise 24

So we get the multi-class logistic model, and we run it for K possible outcomes by running $K-1$ independent binary logistic regression models. We choose the K, the last outcome as pivot, then separately we got the $K-1$ outcomes regressed against the received pivot outcome.

Then it will be like this:

$$ln\ \frac{p(Y_i=1)}{p(Y_i=K)} = a_1 * X_i$$

$$a_2, a_3 \ldots$$

$$ln\ \frac{p(Y_i=K-1)}{p(Y_i=K)} = a_{K-1} * X_i$$

We want to solve for the probabilities for each given model, then we exponentiate (exp) both sides:

$$p(Y_i=1) = p(Y_i=K) * e^{a_1*X_i}$$

$$p(Y_i=K-1) = p(Y_i=K) * e^{a_{K-1}*X_i}$$

We know that sum of all the value of all K of the probabilities must be 1:

$$\sum_{a=1}^{K} p(Y=a|X_i) = \sum_{a=1}^{K-1} \frac{e^{a_i*X_i}}{1+\sum_{j=1}^{K-1} e^{a_j*X_i}} + \frac{1}{1+\sum_{j=1}^{K} e^{a_j*X_i}} = 1$$

From this we can find other probabilities:

$$p(Y_i = 1) = \frac{e^{a_1 * X_i}}{1 + \sum_{j=1}^{K-1} e^{a_j * X_i}}$$

$$p(Y_i = K - 1) = \frac{e^{a_{K-1} * X_i}}{1 + \sum_{j=1}^{K-1} e^{a_j * X_i}}$$