FIT2086 Studio 12
Sample Exam Questions

Daniel F. Schmidt

October 15, 2018

Contents

1 Introduction

The Studio 12 questions are examples of the type of questions you will be asked on the exam, in number and length roughly commensurate with the real exam. Please work on this questions during, and after the studio. You may ask your demonstrator for some assistance on the questions your studio time.

2 Short Answer Questions

Please provide a short (2-3 sentences) description of the following terms:

1. Bias and variance of an estimator
2. R^2 value
3. A p-value
4. Classification accuracy, sensitivity, specificity
5. A decision tree
6. Penalized regression
7. A random variable

3 Maximum Likelihood Estimation

A random variable Y is said to follow an exponential distribution with a rate parameter β, if

$$\mathbb{P}(Y = y \mid \beta) = \beta \exp(-\beta y)$$

where $y > 0$ is a non-negative continuous number. Imagine we observe a sample of n non-negative real numbers $\mathbf{y} = (y_1, \ldots, y_n)$ and want to model them using an exponential distribution. *(hint: remember that the data is independently and identically distributed).*

1. Write down the exponential distribution likelihood function for the data $\mathbf{y}$ (i.e., the joint probability of the data under an exponential distribution with rate parameter β).
2. Write down the negative log-likelihood function of the data $\mathbf{y}$ under an exponential distribution with rate parameter β.
3. Derive the maximum likelihood estimator for β.

4 Confidence Intervals and p-values: I

A car company runs a fuel efficiency test on a new model of car. They perform 6 tests, and in each test they drive the car until the fuel tank is empty, then calculate the litres of fuel consumed per one-hundred kilometers of distance covered. The observed efficiencies (in litres per 100 kilometers, $L/100km$) were:

$$\mathbf{y} = (7.87, 8.10, 9.07, 8.83, 7.60, 8.91).$$

From previous efficiency experiments the car company has estimated the population standard deviation in fuel efficiency recordings (i.e., the experimental error) to be 0.3 ($L/100km$). We can assume that a normal distribution is appropriate for our data, and that the population standard deviation of fuel efficiency recordings for our experiment is the same as the population standard deviation of fuel efficiency recordings of previous experiments.

1. Using our sample, estimate the population mean fuel effiency for this brand of car. Calculate a 95% confidence interval for the population mean fuel efficiency and summarise your results appropriately.

2. The car company runs the same set of tests, on the same set of cars, but with a different brand of fuel. The new observed fuel efficiencies (again, in $L/100km$) were

   $$\mathbf{y}_B = (7.74, 7.74, 8.22, 7.88, 7.85, 8.27).$$

   The company wants to know if this fuel has made any difference to the fuel efficiency. Again, we can assume the population standard deviation for this new set of fuel efficiency measurements is known to be 0.3 $L/100km$. Using this information, please provide a p-value for testing the null hypothesis that the mean fuel efficiency for the two fuel types is the same. Please interpret this p-value.

5 Random Variables

Suppose Y_1 and Y_2 are two random variables distributed as per $Y_1 \sim \text{Poi}(2)$ and $Y_2 \sim \text{Poi}(4)$. Remember that $\text{Poi}(\lambda)$ denotes a Poisson distribution with rate parameter λ, which means the random variable follows the probability distribution:

$$\mathbb{P}(Y = y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}.$$

Recall that if $Y \sim \text{Poi}(\lambda)$, then $\mathbb{E}[Y] = \lambda$ and $\mathbb{V}[Y] = \lambda$. Let $S = Y_1 + Y_2$ denote the sum of these two variables; then:

1. What is the value of $\mathbb{E}[S]$?

2. What is the value of $\mathbb{V}[S]$?

3. What is the probability that $S = 0$?

4. What is the value of $\mathbb{E}[Y_1 Y_2]$?

5. What is the value of $\mathbb{E}\left[Y_1^2\right]$?

6 Confidence Intervals and p-values: II

Consider a drug targetting obesity being considered for introduction to the market by the Therapeutic Goods Administration (TGA). The drug has been demonstrated to substantially reduce BMI, but the TGA are concerned about possible side-effects. They have measured cholestrol levels (in millimols per L $mmol/L$) on a cohort of 7 individuals who have been administered our drug. The measurements were

$$\mathbf{y} = (5, 5.2, 5.05, 5.35, 5.03, 5.43, 5.36).$$

The population standard deviation for cholesterol levels is $0.6 mmol/L$. We can assume that a normal distribution is appropriate for our data, and that the population standard deviation of cholesterol levels for individuals in our sample is the same as the population standard deviation of cholesterol levels for the general population.

1. Using our sample, estimate the population mean cholesterol levels of people being administered the drug. Calculate a 95% confidence interval for the population mean cholesterol level. Summarise your results.

2. The mean cholesterol level in the general populace is known to be $4.8 mmol/L$. The TGA wants to know two things: (i) is the population mean cholesterol level in people being given the drug different from the general population, and (ii) is it higher than in the general population. Specify appropriate null and alternative hypotheses for these two questions, and calculate appropriate p-values to provide evidence against each null hypothesis. What is your conclusion regarding these two questions?

7 Regression

1. Please explain the intuition behind the principle of least squares that is used to fit a linear model with predictor $\mathbf{x} = (x_1, \ldots, x_n)$ to the targets $\mathbf{y} = (y_1, \ldots, y_n)$, and write down the least-squares objective function.

2. If one of our predictors in a regression, or logistic regression model, is categorical, how can we handle it?

3. Imagine we model a persons blood pressure in $mmHg$ (`BP`) using a linear regression. Two predictors are fitted as part of the model: (i) the persons age in years (`AGE`), and the amount of alcohol they consume on average per week `ALCOHOL` (in standard drinks). The model we arrived at is:

$$\mathbb{E}\left[\texttt{BP}\right] = 51 + 1.4\,\texttt{AGE} + 0.6\,\texttt{ALCOHOL}$$

   (a) From this model, how does a person's blood pressure change as their age and alchohol consumption vary?

   (b) If a person is 33 years old, and drinks on average 2.5 standard drinks per week, what is their expected blood pressure?

	No Breast Cancer ($C = 0$)	Breast Cancer ($C = 1$)
Non-dense Breasts ($D = 0$)	0.15	0.10
Dense Breasts ($D = 1$)	0.20	0.55

Table 1: Population joint probabilities of having dense breasts (D), and breast cancer by age 60 (C).

8 Classification

Breast cancer is one of the leading causes of death of women in Western populations. It is believed that mammographic density, which is defined as the amount of non-fat tissue in a woman's breast, is strongly associated with the risk of developing breast cancer. We define a woman's breasts to be "dense" if they contain over $70cm^3$ of non-fat tissue. Table 1 shows the joint probabilities of having dense breasts and contracting breast cancer by age 60.

1. What is the probability of contracting breast cancer by age 60 given that a woman does not have dense breasts?

2. What is the probability of contracting breast cancer by age 60 given that a woman does have dense breasts?

3. Do you think that having dense breasts is a good predictor of contracting breast cancer by age 60, and why/why not?

9 Machine Learning

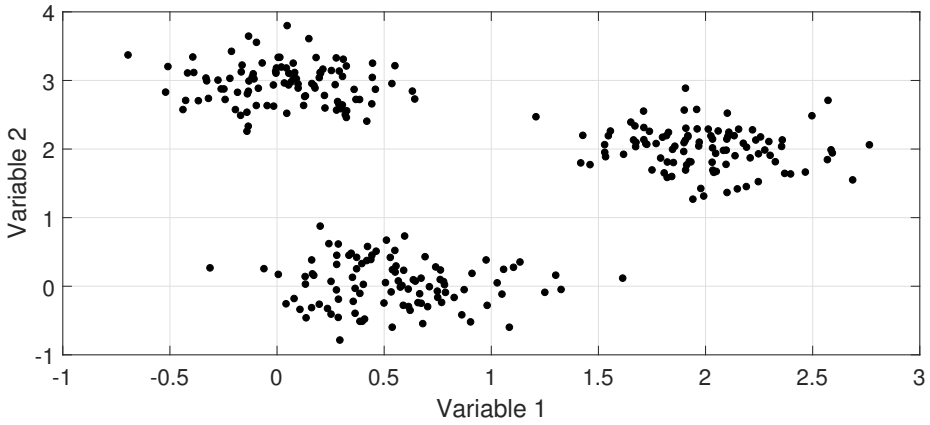

Figure 1: Figure 4: Scatter-plot of two variables.

1. The k-means algorithm is a popular method for clustering. Please explain how this algorithm works.

2. What is one limitation of the k-means algorithm?

3. Figure 1 shows a scatter plot of data points. If we were using the k-means clustering algorithm to cluster this data, what value of k do you think would be appropriate? (1 mark)

```
> cv$best.tree
node), split, n, deviance, yval
      * denotes terminal node

 1) root 442 2621000 152.10
   2) S5 < 4.60015 218  706500 110.00
     4) BMI < 26.95 171  366600  96.31 *
     5) BMI > 26.95 47  191500 159.70 *
   3) S5 > 4.60015 224 1150000 193.20
     6) BMI < 27.75 116  475100 162.70 *
     7) BMI > 27.75 108  451900 225.90
      14) BMI < 32.75 77  305400 208.60
        28) BP < 99.5 33  171800 178.20 *
        29) BP > 99.5 44   80330 231.30 *
      15) BMI > 32.75 31   66120 268.90 *
>
```

Figure 2: R output describing a decision tree learned using cross-validation for the diabetes progression dataset.

4. We have collected data on $n = 442$ diabetic people. Figure 2 shows the R output after using the `tree` package to learn a decision tree to predict their degree of diabetes progression (a non-negative integer) using three predictors in the dataset. The predictors used were as follows: `BMI` is body-mass index (kg/m^2), `BP` is blood pressure in millimeters of Mercury and `S5` is serum measurement (in millimeters).

   (a) How many "leaf" nodes does the tree have?

   (b) If `BMI` $= 23$, `BP` $= 29.1$, `S5` $= 5.5$ what is the degree of diabetes progression predicted by this tree?

   (c) What combination of predictors leads to the greatest degree of diabetes progression?

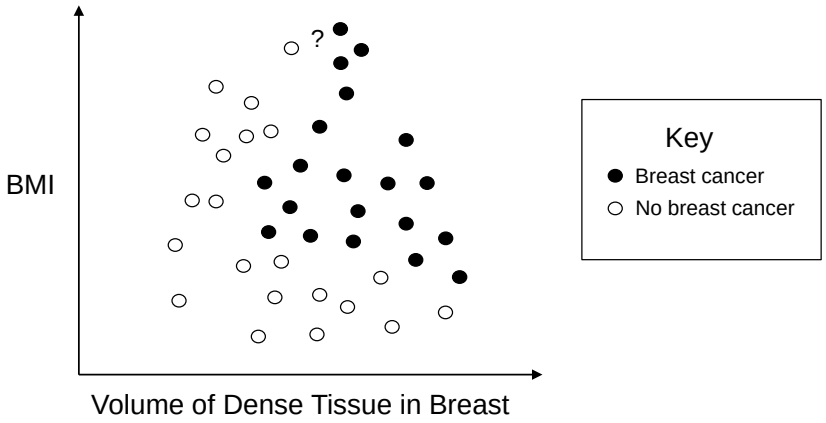

Figure 3: Scatter-plot of body mass index (BMI) against percent dense-tissue in a woman's breast for women with and without breast cancer.

5. Discuss one advantage that a decision tree has in comparison to a linear regression model.

6. The k-nearest neighbours method is a commonly used machine learning algorithm.

   (a) Figure 3 shows a scatter plot of a training sample of women, both with and without breast cancer. The x and y axis are the predictors volume of dense tissue in the woman's breast and body mass index, respectively. The question mark shows a new individual we have obtained data on, but for whom we do not know disease status. Would a k-nearest neighbour algorithm, using standard Euclidean distance and $k = 4$ nearest neighbours, predict them to have breast cancer or not? Please justify your answer.

   (b) Looking at the configuration of the data points in Figure 3, do you think that a logistic regression model would be appropriate for seperating women with and without breast cancer on the basis of the volume of dense tissue in a woman's breast and her body mass index? If so, why do you think so, and if not, why do you think it is not appropriate?

10 Appendix I: Standard Normal Distribution Table

$\|z\|$	$\mathbb{P}(Z < -\|z\|)$	$\mathbb{P}(Z < \|z\|)$	$\|z\|$	$\mathbb{P}(Z < -\|z\|)$	$\mathbb{P}(Z < \|z\|)$
0.000	0.500000	0.500000	2.047	0.020353	0.979647
0.093	0.462943	0.537057	2.140	0.016196	0.983804
0.186	0.426204	0.573796	2.233	0.012789	0.987211
0.279	0.390096	0.609904	2.326	0.010020	0.989980
0.372	0.354912	0.645088	2.419	0.007790	0.992210
0.465	0.320924	0.679076	2.512	0.006009	0.993991
0.558	0.288375	0.711625	2.605	0.004598	0.995402
0.651	0.257471	0.742529	2.698	0.003491	0.996509
0.744	0.228382	0.771618	2.791	0.002630	0.997370
0.837	0.201237	0.798763	2.884	0.001965	0.998035
0.930	0.176125	0.823875	2.977	0.001457	0.998543
1.023	0.153093	0.846907	3.070	0.001071	0.998929
1.116	0.132151	0.867849	3.163	0.000781	0.999219
1.209	0.113273	0.886727	3.256	0.000565	0.999435
1.302	0.096403	0.903597	3.349	0.000406	0.999594
1.395	0.081455	0.918545	3.442	0.000289	0.999711
1.488	0.068326	0.931674	3.535	0.000204	0.999796
1.581	0.056894	0.943106	3.628	0.000143	0.999857
1.674	0.047024	0.952976	3.721	0.000099	0.999901
1.767	0.038577	0.961423	3.814	0.000068	0.999932
1.860	0.031410	0.968590	3.907	0.000047	0.999953
1.953	0.025381	0.974619	> 4.000	< 0.000032	> 0.999968

Table 2: Cumulative Distribution Function for the Standard Normal Distribution $Z \sim N(0, 1)$