FIT2086 Studio 12
Revision Questions

Daniel F. Schmidt

October 21, 2017

Contents

1 Introduction

The Studio 12 questions are examples of the type of questions you will be asked on the exam. Please work on this questions during, and after the studio. Your demonstrator will go through some of the answers to the questions during the Studio with you.

2 Short Answer Questions

Please provide 2-3 sentence description of the following terms:

A: General comment. When answering short-answer questions of this form (in general), it is a good idea to use the following basic structure: your first sentence should describe *what* the object/item of interest is. The second and third sentences (or fourth, the $2-3$ is a guide and not a strict requirement!) should describe one or two properties of the object. This allows a marker to clearly see that you can (i) identify the object of interest, and (ii) you know something about the object of interest. All the answers below follow this basic structure.

1. Bias and variance of an estimator

   A: The bias of an estimator is the average amount by which the estimator under, or over-estimates a parameter. If the bias is zero, the estimator is unbiased.
   The variance of an estimator is the average squared-deviation of the estimator from its average value. It measures how much we would expect the estimate to vary if we drew a new sample from the population.

2. R^2 value

   A: The R^2 value is one minus the ratio of the residual sum of squares of a linear model over the total sum of squares. It measures how well a linear model fits data. The R^2 value varies from zero (model does not fit the data at all) to one (model fits the data perfectly).

3. An information criterion

   A: An information criterion is a measure of goodness-of-fit of a model that takes into account the model complexity. It is formed by adding a penalty which is based on the number of parameters in the model to the negative log-likelihood (which measures how well the model fits the data).

4. A p-value

   A: A p-value is used in hypothesis testing to measure evidence against the null hypothesis. A p-value is the probability of seeing a test-statistic as extreme, or more extreme, than the one we have observed, just by chance, if the null hypothesis was true.

5. Classification accuracy, sensitivity, specificity

   A: Classification accuracy is the percentage of times our model correctly classifies an individual/object.
   Sensitivity is the proportion of correct classifications of individuals as a "success" (or a "1")
   Specificity is the proportion of correct classifications of individuals as a "failure" (or a "0")

6. A decision tree

   A: A decision tree is supervised machine learning method. It works by sequentially splitting the data into disjoint sets based on the values of the attributes of each of the individuals in our data.

3 Maximum Likelihood Estimation

A random variable Y is said to follow a geometric distribution with probability p if

$$\mathbb{P}(Y = y \,|\, p) = (1-p)^y p$$

where $y \in \{0, 1, 2, \ldots\}$ is a non-negative integer. Imagine we observe a sample of n non-negative integers $\mathbf{y} = (y_1, \ldots, y_n)$ and want to model them using a geometric distribution. *(hint: remember that the data is independently and identically distributed).*

1. Write down the geometric distribution likelihood function for the data $\mathbf{y}$ (i.e., the joint probability of the data under a geometric distribution with probability parameter p).

   A: By independence

   $$\begin{aligned} p(y_1, \ldots, y_n \,|\, p) &= \prod_{i=1}^{n} p(y_i \,|\, p) \\ &= [(1-p)^{y_1} p] \cdot [(1-p)^{y_2} p] \cdots [(1-p)^{y_n} p] \\ &= (1-p)^{\sum_{i=1}^n y_i} p^n \end{aligned}$$

   using $e^a e^b = e^{a+b}$.

2. Write down the negative log-likelihood function of the data $\mathbf{y}$ under a geometric distribution with probability parameter p.

   A: Taking negative logarithm of the above expression and simplifying:

   $$\begin{aligned} -\log p(y_1, \ldots, y_n \,|\, p) &= -\log \left[(1-p)^{\sum_{i=1}^n y_i} p^n \right] \\ &= -\sum_{i=1}^{n} y_i \log(1-p) - n \log p \end{aligned}$$

   using $\log a^b = b \log a$.

3. Derive the maximum likelihood estimator for p.

   A: To find the ML estimate of p, differentiate the negative-log-likelihood w.r.t. p, set it to zero and solve for p. The derivative is:

   $$\begin{aligned} \frac{d}{dp} \{-\log p(y_1, \ldots, y_n \,|\, p)\} &= -\sum_{i=1}^{n} y_i \frac{d}{dp} \{\log(1-p)\} - n \frac{d}{dp} \{\log p\} \\ &= \frac{\sum_{i=1}^n y_i}{1-p} - \frac{n}{p} \end{aligned}$$

where we use $d \log x/dx = 1/x$, and the chain rule. Then, set this to zero and solve for p:

$$\begin{aligned}
& \frac{\sum_{i=1}^n y_i}{1-p} - \frac{n}{p} = 0 \\
\Rightarrow \quad & \frac{p\sum_{i=1}^n y_i}{1-p} - n = 0 \\
\Rightarrow \quad & p\sum_{i=1}^n y_i - n(1-p) = 0 \\
\Rightarrow \quad & p\left(n + \sum_{i=1}^n y_i\right) - n = 0 \\
\Rightarrow \quad & p\left(n + \sum_{i=1}^n y_i\right) = n \\
\Rightarrow \quad & p = \frac{n}{(n + \sum_{i=1}^n y_i)}
\end{aligned}$$

4 Confidence Intervals and p-values

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.

   A: The sample mean of sample is

   $$\hat{\mu} = \frac{1}{7}(5 + 5.2 + 5.05 + 5.35 + 5.03 + 5.43 + 5.36) \approx 5.2$$

   where we rounded 5.2029 down to 5.2. The standard error is

   $$\text{se}_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{7}} \approx 0.227$$

   The formula for the 95% confidence interval for a mean with known variance is

   $$CI = (\hat{\mu} - 1.96\text{se}_{\hat{\mu}},\ \hat{\mu} + 1.96\text{se}_{\hat{\mu}})$$

   so we have

   $$CI = (5.2 - 1.96 \times 0.227,\ 5.2 + 1.96 \times 0.227) = (4.75,\ 5.64).$$

Summary of results: The estimated mean cholesterol level in our sample of size $n = 7$ of individuals being prescribed our drug of interest is $5.2mmol/L$. We are 95% confident that the population mean cholesterol level of people using this drug is between $4.75mmol/L$ and $5.64mmol/L$.

2. The mean cholesterol level in the general populace is known to be $4.8mmol/L$. 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. Calculate appropriate p-values to provide evidence against the null hypothesis (that the cholesterol levels are the same in the group taking the drug and in the general populace) against these two alternative hypotheses. What is your conclusion regarding these two questions?

   A: First, always state the hypotheses you are testing clearly. This helps to make sure you are doing the right thing.

   (i) To answer the first part of the question we are testing the null hypothosis $H_0 : \mu = 4.8$ vs $H_A : \mu \neq 4.8$. From the Lecture notes regarding testing the population mean of a normal population with known variance, we must first calculate the sample mean for our sample, which we have above ($\hat{\mu} = 5.2mmol/L$). Then we calculate the z-score

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{5.2 - 4.8}{0.6/\sqrt{7}} \approx 1.763$$

   where we can treat our population standard deviation as known and equal to $\sigma = 0.6$ (as per the assumptions made in the question), and our sample size is $n = 7$. Then, using this z-score we need to do a two-sided test, so our p-value is

$$p = 2\mathbb{P}(Z < -|z_{\hat{\mu}}|)$$

   To do this, look at the Standard Normal Distribution table in the Appendix. The first column is the absolute value of the z-score; the second column is $\mathbb{P}(Z < -|z|)$, which is what we need. Move down the rows to $|z| = 1.767$ (which is the closest entry to our z-score), and we see that $\mathbb{P}(Z < -1.767) \approx 0.0385$. Our p-value is twice this, as we are doing a two-sided test, so we have $p \approx 0.077$. This suggests there is some weak evidence against the null that the population mean cholesterol level of people using the drug is the same as the population mean cholesterol level in the general populace.

   (ii) To answer the second part of the question we are testing null hypothesis $H_0 : \mu \leq 4.8$ vs the alternative $H_A : \mu > 4.8$. From the Lecture notes, we see that for this one-sided test we need the same z-score as calculated above, but now our p-value is

$$p = 1 - \mathbb{P}(Z < z_{\hat{\mu}})$$

   We have $z_{\hat{\mu}} = 1.767$, so we can use the third column of the Table in the appendix to calculate $\mathbb{P}(Z < |z|)$. Again, find the row corresponding to $|z| = 1.767$ and this gives us $\mathbb{P}(Z < 1.767) \approx 0.961$. Our p-value is therefore $p \approx 1 - 0.961 = 0.039$. We see that there is moderate evidence against the null that the population mean cholesterol level of people using the drug is less than or equal to the population mean cholesterol level in the general populace.

   Overall, looking at both tests, we see there is some evidence to suggest that we can reject the null that the drug does not affect the mean cholesterol level of individuals compared individuals in the general populace, but it is not very conclusive. A larger study is probably required.

5 Regression

1. Please explain how we can use the principle of least squares to fit a linear model with predictor $\mathbf{x} = (x_1, \ldots, x_n)$ to the targets $\mathbf{y} = (y_1, \ldots, y_n)$

   A: The principle of least squares says we should find the values of the coefficient and intercept for this simple linear model that result in the model that minimises the sum-of-squared errors (residuals) between the model predictions and the data values $\mathbf{y}$.

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

   A: If our predictor is a categorical variable with K categories, we can handle this by creating $K - 1$ new dummy variables (predictors). The new variable number k will take on a "1" if an individual is in category $k + 1$ and a 0 otherwise. These are called indicator variables as they indicate which category an individual is in.

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:

   $$\texttt{BP} = 51 + 1.4\,\texttt{AGE} + 0.6\,\texttt{ALCOHOL}$$

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

   A: To get full marks for a question like this, you must explicitly state the size of the effect, the direction of the effect and the units of the variables involved. For example:

   i. For each year a person has lived, their expected blood pressure will increase by $1.4mmHg$.

   ii. For each additional standard drink a person consumes on average per week, their expected blood pressure will increase by $0.6mmHg$.

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

   A: Just plug the numbers into the equation: $51 + 1.4 \times 33 + 0.6 \times 2.5 = 98.7mmHg$. Don't forget units!

6 Machine Learning and Classification

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

 1) root 668 868.90 N ( 0.645210 0.354790 )
   2) PLAS < 127.5 418 408.20 N ( 0.808612 0.191388 )
     4) BMI < 26.95 119  11.55 N ( 0.991597 0.008403 ) *
     5) BMI > 26.95 299 345.30 N ( 0.735786 0.264214 )
      10) AGE < 29.5 165 125.80 N ( 0.872727 0.127273 )
        20) SKIN < 28.2 62   0.00 N ( 1.000000 0.000000 ) *
        21) SKIN > 28.2 103 104.20 N ( 0.796117 0.203883 ) *
      11) AGE > 29.5 134 183.30 N ( 0.567164 0.432836 )
        22) INS < 89.5 33  24.38 N ( 0.878788 0.121212 ) *
        23) INS > 89.5 101 139.50 Y ( 0.465347 0.534653 ) *
   3) PLAS > 127.5 250 330.00 Y ( 0.372000 0.628000 )
     6) PLAS < 154.5 142 196.70 N ( 0.514085 0.485915 )
      12) BMI < 29.95 42  46.11 N ( 0.761905 0.238095 ) *
      13) BMI > 29.95 100 135.40 Y ( 0.410000 0.590000 ) *
     7) PLAS > 154.5 108 103.50 Y ( 0.185185 0.814815 ) *
>
```

Figure 1: R output describing a decision tree learned using cross-validiation for the Pima Indians Diabetes dataset.

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

   A: The k-means algorithm tries to find the k centroids of k clusters of data. It works as follows:

   (a) Randomly initialise the k centroids

   (b) Find which data points are closest to which of the k centroids (form clusters)

   (c) Re-estimate the location of the cluster centroids using the data closest to that centroid

   (d) Repeat from Step (b) until convergence.

   In general you are advised to look at the marks associated with a question – this will give you a roughish guide as to how much you should probably need to write as an answer.

2. We have collected data on $n = 768$ Pima ethnic indian people, with and without disease. Figure 1 shows the R output after using the `tree` package to learn a decision tree to predict diabetes status (no or yes) using the predictors in the Pima Indians dataset. The predictors used are as follows: `PLAS` is the plasma glucose level, `BMI` is body-mass index (kg/m^2), `AGE` is age in years, `SKIN` is the triceps skin fold thickness (mm) and `INS` is 2-hour serum insulin (milli-units/ml).

   A: The figure shows a textual representation of the tree. This was described in the solutions to Studio 9 (but for a regression tree – this is a classification tree). To re-iterate: for each row, the first number is an (arbitrary) node number. The second quantity is the condition required to reach this node from the previous parent node. The third quantity is the number of samples that satsify all the conditions required to arrive at the node. The fourth number is a goodness-of-fit value that can be ignored. For a classification tree, the 5th quantity is the most likely class for people in the node (in this case, either "N" or "Y" for no diabetes or yes diabetes). The numbers in the parenthesis are the probabilities of having diabetes – the first number is the probability

of not having diabetes, while the second is the probability of having diabetes. Note that when a node predicts a "N", the first number is larger than the second, and when it predicts a "Y" the second number is higher than the first.

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

A: The leaf nodes are terminal nodes, and are starred – so in this case, there are 8 leaf nodes.

(b) If PLAS = 123, BMI = 29.8, AGE = 46, SKIN = 26.2, INS = 85, what is the odds of a person having diabetes?

A: To find the odds, we first must find the probability of having diabetes. To do this, we simply need to traverse the tree for the predictors we have. First, we note that we have PLAS < 127.5, so we move to Node #2. Then, our BMI > 26.95 so we move to Node #5. Then, we have AGE > 29.5, so we move to Node #11. The value of SKIN is not used in this branch of the tree, so we see that INS < 89.5 for this individual, so we move to Node #22, which is a terminal (leaf node). The probability of having diabetes is the second number in the parenthesis, so we see that $P(\texttt{DIABETES} = \texttt{Y}) \approx 0.121$. The odds is then

$$\text{odds}(\texttt{DIABETES} = \texttt{Y}) = \frac{P(\texttt{DIABETES} = \texttt{Y})}{1 - P(\texttt{DIABETES} = \texttt{Y})} = \frac{0.121}{1 - 0.121} \approx 0.137$$

So an individual with that combination of predictor values is 0.137 times more likely to have diabetes than to not have diabetes.

(c) What combination of predictors leads to the smallest probability of having diabetes?

A: To answer this question, we must find the leaf node that has the smallest probability of diabetes, and then work back down the tree to figure out what combination of predictor values we need to arrive at this node. Node #20 has a probability of having diabetes of 0, so this is the leaf node with the lowest probability of having diabetes in the tree. To arrive at this node, we need to go from the root to Node #2 (PLAS < 127.5), then to Node #5 (BMI > 26.95), then to Node #10 (AGE < 29.5) and finally to Node #20 (SKIN < 28.2). So to summarise, we need:

- PLAS < 127.5;
- BMI $> 26.95 kg/m^2$;
- AGE < 29.5 years;
- SKIN $< 28.2 mm$.

7 Appendix I: Standard Normal Distribution Table

$\lvert z\rvert$	$\mathbb{P}(Z < -\lvert z\rvert)$	$\mathbb{P}(Z < \lvert z\rvert)$	$\lvert z\rvert$	$\mathbb{P}(Z < -\lvert z\rvert)$	$\mathbb{P}(Z < \lvert z\rvert)$
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 1: Cumulative Distribution Function for the Standard Normal Distribution $Z \sim N(0, 1)$