FIT2086 Studio 12
Revision Questions

Daniel F. Schmidt

October 16, 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:

1. Bias and variance of an estimator
2. R^2 value
3. An information criterion
4. A p-value
5. Classification accuracy, sensitivity, specificity
6. A decision tree

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

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.

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. 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?

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 $\mathbf{y} = (y_1, \ldots, y_n)$?

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:

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

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

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

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.

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 (yes or no) 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) How many "leaf" nodes does the tree have?

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

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

7 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 1: Cumulative Distribution Function for the Standard Normal Distribution $Z \sim N(0, 1)$