Module Code
COM00028H

BEng, BSc, MEng and MMath Degree Examinations 2019–20

DEPARTMENT OF COMPUTER SCIENCE

Fundamentals of Machine Learning

Open Individual Assessment

Issued: 20 November 2019 (12 noon)
Submission due: 4 December 2019 (12 noon)
Feedback and marks due: 9 January 2020

All students should submit their answers through the electronic submission system: http://www.cs.york.ac.uk/student/assessment/submit/ by 4 December 2019 (12 noon). An assessment that has been submitted after this deadline will be marked initially as if it had been handed in on time, but the Board of Examiners will normally apply a lateness penalty.

Your attention is drawn to the section about Academic Misconduct in your Departmental Handbook: https://www.cs.york.ac.uk/student/handbook/.

Any queries on this assessment should be addressed by email to James Cussens at james.cussens@york.ac.uk. Answers that apply to all students will be posted on the VLE.

Your exam number should be on the front cover of your assessment. You should not be otherwise identified anywhere on your submission.

Rubric

- All data and any other additional files are available at the FUML VLE site in the **Assessment** section.
- Your submission should be a single zip file containing a PDF and a number of Python programs. **If your submission is not a single zip file you will be penalised**. The PDF is your *report*. The questions below specify what should be in your report and which Python programs should be submitted.
- All Python programs should be in Python3.
- Your Python code can assume that a working installation of scikit-learn is available.
- Your Python code must run correctly when run under Linux on the Ubuntu distribution which is used in our department's software labs (i.e. the distribution you used during practicals).
- Unless otherwise indicated there are no word limits on your answers.

1 Bayesian and non-Bayesian estimation of probabilities (40 marks)

Download the file `discrete.csv` from the Assessment section of the FUML VLE site. You will see that this file contains data for 4 binary random variables $X1$, $X2$, $X3$ and Y. The first line of the file is a header line stating the names of the variables and the following lines are 500 datapoints. All values are separated by commas.

Your task in this question is to write a Python program that takes this file as input and produces as output estimates of the following probability distributions: $P(Y)$, $P(X1|Y)$, $P(X2|Y)$, $P(X3|Y)$. Two sorts of estimates are required: maximum likelihood estimates and a Bayesian estimate. Your code does not need to deal with the possibility of undefined MLEs, all MLEs for this data are defined. The Bayesian estimate should be the mean of the relevant posterior distribution where the prior is the uniform distribution in all cases. The specific output format required is shown in Fig 1 (where, of course, you need to ensure the $\star$ is replaced with the correct value). Note that only the estimate for the value of 0 is required.

1. You are only allowed to use modules in the Python standard library to do this question.
2. Estimates should be given to at least 3 significant figures (more is OK if you like).
3. You are not asked to provide any explanation of your answers. Marks for this question are determined entirely by the correctness of your output.

```
MLE:
P(Y=0) = ⋆
P(X1=0|Y=0) = ⋆
P(X1=0|Y=1) = ⋆
P(X2=0|Y=0) = ⋆
P(X2=0|Y=1) = ⋆
P(X3=0|Y=0) = ⋆
P(X3=0|Y=1) = ⋆

Bayesian
P(Y=0) = ⋆
P(X1=0|Y=0) = ⋆
P(X1=0|Y=1) = ⋆
P(X2=0|Y=0) = ⋆
P(X2=0|Y=1) = ⋆
P(X3=0|Y=0) = ⋆
P(X3=0|Y=1) = ⋆
```

Figure 1: Required output format for Question 1

4. The efficiency or ‘elegance’ of your code has no bearing on your mark for this question.

(Although you don’t need to know this to do this question, the probabilities you are estimating here are those required for a *naive Bayes* model which is a popular machine learning for classification.)

What to submit: Put the output of your program in your report. Also submit your program and call it `q1.py`

2 Which regression? (40 marks)

Download the file `continuous.csv` from the Assessment section of the FUML VLE site. You will see that this file contains data for 11 continuous random variables $X1 \ldots X10, Y$. The first line of the file is a header line stating the names of the variables and the following lines are the 8 datapoints. All values are separated by commas.

Your task here is to create a regression model for predicting values of Y based on the values of $X1 \ldots X10$. You need to choose which type of regression to do, and then use scikit-learn to estimate the relevant parameters from these 8 datapoints. I will be testing your regression model on a test set and computing the R^2 value of the predictions. Your goal is for this R^2 value to be as close to 1 as possible. You do not have access to the test set.

You are required to submit a Python program called q2.py such that when it is called (from the command line) like this:

```
python q2.py continuous.csv secrettestset.csv
```

then it (1) estimates the parameters of your regression model from the data in continuous.csv and (2) outputs the predicted Y value for each data point in secrettest.csv and also the R^2 value for the fitted regression model when evaluated on secrettest.csv.

Although you should be trying to maximise the R^2 value for the unseen test set your marks for this question are not determined by this value (since you might get a good/bad R^2 value by luck!).

Mark distribution for this question:

- Appropriateness of your chosen regression model (15 marks)
- Justification given for your chosen regression model (15 marks)
- Correctness of your Python program (10 marks)

What to submit: Submit your program and call it q2.py. In your report state which regression approach you used and provide a justification for your choice. Your justification must be less than 100 words. You can also use figures/diagrams in your justification (these do not contribute to the word count).

3 Dimensionality reduction (20 marks)

Download the file pca_ex.csv from the Assessment section of the FUML VLE site. You will see that this file contains data for 3 continuous random variables $X1, X2, X3$. There are 100 datapoints in the data. Also download the file classes.txt. This contains 100 datapoints for a binary variable (let's call it Y) corresponding to the values in pca_ex.csv.

Do these 3 scatterplots using the data in pca_ex.csv: $X1$ vs $X2$, $X1$ vs $X3$ and $X2$ vs $X3$. In each case use different colours to differentiate datapoints corresponding to the different values of Y.

Now do a 4th scatterplot where the two axes are the first two principal components of the data in pca_ex.csv. Again differentiate datapoints for the different classes using colour.

Suppose you had to build a classifier (predicting future values of Y) from this data which was only allowed to use two variables to make predictions. Based just on these plots, what would be a reasonable choice of classifier? Does it matter which pair of variables is chosen from these 4 possible options: (1) $X1, X2$, (2) $X1, X3$, (3) $X2, X3$ and (4) the two first principal components? If it does matter explain which would be good and which bad choices.

Mark distribution for this question:

- Correctly producing the first 3 scatterplots (5 marks).
- Correctly producing the principal components scatterplot (5 marks).
- Justification of choice of classifier and choice of two variables to use (10 marks).

What to submit: You should not submit any code for this question. Include the following in your report: the 4 scatterplots and your answer to the questions posed in the preceding paragraph.

End of examination paper