Machine Learning I

Mid-Semester Examination

Ramakrishna Mission Vivekananda University
M.Sc. in Data Science and M.Sc. in Computer Science

Date : 18 March 2017 Maximum Marks : 60 + 20 (bonus) Duration : 3 Hours

This is an open-resources examination. You may use any online or offline academic material as reference during the test, and you are free to borrow codes from any non-interactive source on the Internet. However, posting any query on discussion forums, emails, blogs, and interactive online portals during the test will be considered unethical, and will draw penalty. Discussion or any form of digital communication with your fellow students during the test will be considered unethical, and will draw penalty. You must refer to each and every online and offline source you use during the test at the end of your answer script.

Problem 1 (Compulsory) [40 points]

Consider the training dataset `wineTrain.csv` (attached), which has 12 variables, as follows.

```
"FixedAcidity"        "VolatileAcidity"      "CitricAcid"            "ResidualSugar"
"Chlorides"           "FreeSulphurDioxide"   "TotalSulphurDioxide"   "Density"
"pH"                  "Sulphates"            "Alcohol"               "Quality"
```

The target is to fit a linear regression model to predict the `"Quality"` of the wine based on the other 11 variables. In this connection, please answer the following questions, in given sequence.

A. What is the *null model* on this data? What is the value of Total Sum of Squares (TSS) on the dataset, assuming the null model? What is the value of R^2 in this case? [5]

B. What is the *full model* on this data? What is the value of Residual Sum of Squares (RSS) and R^2 on the dataset, assuming the full model? Are these the best possible values? [5]

C. Is the full model best to predict the response variable `"Quality"`? If you think so, justify your answer. If not, then in your opinion, what is the *best model* to predict the response variable `"Quality"` in case of the given dataset? Justify your choice for *best model*. [15]

D. How did you ensure that your *best model* predicts reasonably well for any *test data* of the same type, and not just the *train data* given to you? Briefly justify your approach. [10]

E. Were you happy with the distribution of the train data when you chose your *best model*, or were there any undesired outlier? How did you deal with such outliers, if any? [5]

Answer any one of the following problems. The other one will be considered as bonus.

Problem 2 (Optional) [20 points]

Consider the training dataset `wineTrain.csv` (attached), which has 12 variables, as follows.

```
"FixedAcidity"       "VolatileAcidity"      "CitricAcid"           "ResidualSugar"
"Chlorides"          "FreeSulphurDioxide"   "TotalSulphurDioxide"  "Density"
"pH"                 "Sulphates"            "Alcohol"              "Quality"
```

The target is to find out an optimal number of *components* to represent the data without major loss of information. In this connection, please answer the following questions, in given sequence.

Treat the dataset as a 1000×12 matrix D and compute its *Singular Value Decomposition (SVD).*

A. What are the *singular values* of D? What is the rank of D? What is its *Frobenius Norm*? Is there any relation between the Frobenius Norm and the singular values of D? [5]

B. Suppose you are given only the first two columns of u, the first two singular values in d, and the first two columns of v. Can you reconstruct the complete original matrix D, or an approximation of D? Describe the reconstruction process. What will be the rank of the reconstructed matrix, and how *accurate* will be your reconstruction? [10]

C. What is the minimum number of columns of u, singular values in d, and columns of v, required to reconstruct the original matrix D with 95% accuracy? What is the rank of the approximate reconstruction $\hat{D}$, and what is its Frobenius Norm? [5]

OR

Problem 3 (Optional) [20 points]

Consider the training dataset `pimaTrain.csv` (attached), which has 8 variables, as follows.

```
"NumPreg"       "PlasmaGlucose" "DiastolicBP"   "TricepSkin"
"BodyMassIndex" "Pedigree"      "Age"           "Diabetic"
```

The target is to fit a logistic regression model to predict the `"Diabetic"` variable based on the other 7 variables. In this connection, please answer the following questions, in given sequence.

A. Construct the *best model*, in your opinion, to predict the categorical response variable `"Diabetic"` in case of the given dataset? Justify your choice for *best model.* [15]

B. Suppose we choose a threshold t to classify $\Pr(\texttt{Diabetic} \mid X) > t$ as `"Diabetic" = Yes`. How would you choose the optimal threshold t such that the aforesaid classification achieves maximum *accuracy* for your *best model*? Justify your choice. [5]

Good Luck! ☺