**Statistical Data Mining I**

**Concept Quiz IV**

**October 13, 2017**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model Interpretation  (1 poor –  3 good) | Model Assumptions  (briefly describe) | 2 Advantages | 2 Disadvantages |
| kNN | 1 | It makes no assumptions about underlying data . But the value of “k” has to be selected beforehand . | 1)Works well for small input size .  2)Accuracy is usually high . | 1)High Computational Cost  2)Depends on the value of “k” . |
| Regression on  An Indicator  Matrix | 3 | The assumption is that the indicator matrix will have 1s and 0s (Sparse) for predictors . | 1)Computationally Effective .  2)Easy to understand (High interpretability) | 1)Very primitive and simple compared to other methods  2)Has to deal with the phenomena of masking |
| LDA | 1 | Assumes the input is Gaussian . Assumes same covariance matrix for all classes . | 1)Does well when the true decision boundary is linear .  2)More stable than Linear Regression for small number of variables. | 1)Does not perform well for large data sets .  2)Assumption of Gaussian distributed inputs may not always be met . |
| QDA | 1 | Assumes the input is Gaussian . Assumes different covariance matrices for all classes . | 1) Does well when the decision boundary is moderately linear .  2)QDA does better than LDA for large data sets . | 1) LDA performs better than QDA for small data sets .  2)Assumption of Gaussian distributed inputs may not always be met . |
| RDA | 1 | Assumes the input is Gaussian. RDA shrinks different covariances of QDA towards the single one of LDA | 1)Introduces flexibility of model complexity based on selection of alpha .  2)Is a flexible combination of LDA and QDA . | 1)Less robust than Logistic Regression .  2)Assumption of Gaussian distributed inputs may not always be met . |
| Logistic Regression | 2 | Assumes the function to be squashed by a sigmoid to fit between 0 and 1 . Also the observations have to be independent . | 1)Performs better than LDA and QDA if the input is not truly Gaussian .  2)The model is more robust than LDA, QDA , RDA . | 1)Independent Observations Required .  2)Tendency to overfit |

1. Fill in the table

(2) You are handed a data set from a Mass Spec experiment that profiles blood serum from individuals that work in a factory. The outcome is a binary exposure indicator, which conveys if the individual has exhibited respitory illness linked to a poor environment. The data set consists of 200 individuals (100 with illness, and 100 without). 800 metabolites (continuous) are profiled in the Mass Spec machine, and they are continuous variables. There is some missing data. The objective is to assess the use of Mass Spec as a profiling device to predict respitory illness.

(a) What is more important, inference (model interpretation) or prediction (accuracy)?

Since the data set deals with medical information and the prediction would have consequences on the health of an individual / affect diagnosis / misdiagnose results , its better to focus on accuracy than interpretation . But interpretation cannot be too low as physicians/ anyone responding should be able to analyse data easily .

(b) Describe your analysis plan.

Divide the data set into train, validation and test .

Train a logistic regression model on the dataset as this is a classification problem .

Once the model is trained , and all the optimizations are done , we can predict the category of the new data based on the trained model .