CSE471: Statistical Methods in AI

Monsoon 2016

Assignment # 2: Bayesian Decision Theory, Parameter Estimation and Component Analysis (PCA/LDA)

Due: Before 5:00pm on 13th October 2016 **Total Marks:** 80 (Mapped to 4% of course credits)

General Instructions:

- Assignment can be implemented in Matlab/Octave, Python, C/C++, R .
- Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious actions including an **F grade in the course** is possible.
- A single pdf file needs to be uploaded to the Courses Portal. The file should contain your answers as well as the code you have written and its output (Or as directed by the TA's).
- Include the assignment number, your name and roll number at the top-left of the first page of your submission.
- Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code as well viva based oral examination done by TA's.

Problem 1 (30 Marks)

Implement a Naïve Bayes Classifier for UCI Census-Income (KDD) Data Set using only the Discrete and Categorical attributes/features. If you think that some of the real attributes are useful, please convert them to discrete feature with appropriate binning. Use log-probabilities to avoid numerical errors. Randomly sample data into equal size training and test set (of the order of more than 1000 data points) for ten runs and report mean accuracy and standard deviation over all runs. Deal with any ties appropriately but mention this in observations part of the report. Explain how you choose to handle missing entries and why?

Problem 2 (10 Marks)

Derive Bayesian Parameter Estimation (BPE) for both Univariate and Multivariate Normal density functions.

Problem 3 (40 Marks)

Implement dimensionality reduction techniques, namely, Principal Component Analysis (PCA) and Fisher's Linear Discriminant Analysis for UCI Dorothea Data Set using a significantly large subset of 50K real features (neglect the probe feature). Further implement a Gaussian Naïve Bayes classifier in the resulting K-dimensional PCA space (where K=100 500 and 1000) as well as 1-dimensional LDA space and report the classification performance on validation data. Provide the pseudo code, assumptions made, accuracy results and your observations in a report format along with the code.