[CSC 5825 Fall 2017]

Due. Before the Class of Oct, 2, 2017 Homework 2

Full credit: 100 points with extra 20 points bonus

September 19, 2017

Question 1. (20 points) For a two-class problem, for the four cases of Gaussian densities in table 5.1 on textbook page 106, derive $\log \frac{p(C_1|x)}{p(C_2|x)}$ (5 points for each case).

Note that Table 5.1 on page 106 contains four different covariance matrixes and this question asks you to derive $\log \frac{p(C_1|x)}{p(C_2|x)}$ under the four different covariance matrixes using Bayes' rules staring from the equation 5.16 on page 100.

Question 2. (80 points in total) Programming question: Naive Bayes Classifier to Detect Credit Card Fraud. In this question you are asked to fit Naive Bayes Classifier to detect credit card fraud using de-identified credit card transactions labeled as fraudulent or genuine. The below description was adopted from kaggle website.

The datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Please treat V1 to V28 as continuous features and use Gaussian distribution for training as I demonstrated in the lecture, i.e., P(V1 mid Class) is Gaussian, to calculation the class-conditional probabilities in training. And treat 'Time' and 'Amount' as discrete features by discretizing them into 2 states for the former and 3 states for the latter, for example, Low/Medium/High and Near/Far. You should divide the data into a training set (80%) and a test set (20%) and calculate precision, recall and F-score. Data can be downloaded from: `https://www.kaggle.com/dalpozz/creditcardfraud/data`

Bonus: please discuss possible strategies to deal with the class imbalance. (20 points)

Submission Instructions

To earn the full credit, your must type your solutions with sufficient details either using LaTeX or Microsoft Word **embedded** with source code and submit it through the blackboard website. You will get points off for not following this requirement.

Late homework (w/o acceptable documents) will be accepted with penalty. 20% off penalty if late for 24 hours or less. 40% off penalty if late between 24 hours and 48 hours. 60% off penalty if late between 48 hours and 72 hours. 80% off penalty if late between 72 hours and 96 hours. No homework late more than 96 hours will be accepted.