HW #2 Due: 4/11/2019

In this homework, the dataset to be used is again the "Breast Cancer Wisconsin (Original) Data Set" (called cancer dataset). The full features are from attribute 2 to attribute 10, and the classification target is whether the subject is benign or malignant. As you know that there are some missing attributes, you may want to use the in-class average (from the training set) to replace the missing attributes.

1. In the lecture, we mention that $X^T X$ in PCA derivation is (a) orthogonally diagonalizable and (b) semi-definite. Prove the above arguments.
2. We mention the heart rate variability (HRV) in the lecture. Among the two Lorenz plots for R-R intervals, which one is more likely from a healthy subject? Why? Give references if necessary.
3. In the class, we mention the naive Bayes classifier, but only with discrete-type features. Consult any paper to learn how to extend this approach to continuous-type features. For pedagogic reasons, treat the attributes of the cancer dataset as if they were continuous values. (a) Repeat problem 3 in HW #1, but use the Naïve Bayes classifier instead of k-NN. Use all available samples in the dataset for the experiments. (b) We know that class imbalance is a serious problem for the Naïve Bayes classifier. What will you do if you want to avoid this problem during conducting experiments?
4. In this problem, you are asked to perform the wrapper-type feature selection using the *k*-NN with $k = 3$ for cancer dataset. To simplify the problem, you just need to select 3 attributes out of 9. To begin one experiment, randomly draw 50 % of the instances from each class for training, and 20% from each class for finding the best 3 attributes. Once the feature selection is complete, use the rest 30% for testing to obtain the accuracy. Remember to use only the three chosen attributes for *k*-NN ($k = 3$) to classify. During testing, your training set contains any samples not in test set. Repeat the experiments 10 times and report the average accuracy.
5. In this problem, you are asked to use the cancer dataset to perform PCA for the entire cancer dataset. (a) Plot PoV(k) for k from 1 to 9. Remove mean values before computing eigenvalues. (b) When do we need to repeat the experiment 10 times and taking average if PCA is involved for dimension reduction? Explain.