Homework III  
Statistics 640 / 444  Fall 2015  
Assigned: October 15  
Due: October 29  

Stat 444:

1. Spam Data. Please download the Spam Data from the course webpage to use for this problem.

   K-Fold CV - Choose any classifier with one tuning parameter for parts (a-e).

   (a) Write a function to perform K-fold cross-validation to select the tuning parameter for this classifier. You must code this up yourself and cannot use built-in functions (using a built in function for the base classifier is fine).

   (b) Select the optimal tuning parameter using the (i) minimum CV error rule and the (ii) one SE rule for K = 5 fold CV. Are the models selected different? Interpret these results and reflect on this.

   (c) Perform both K = 5 and K = 10 fold CV. Does this change the results? Is one of these preferable for this problem?

   (d) When reporting the CV error, try out different loss functions: (i) missclassification error, (ii) binomial deviance error, and (iii) hinge loss error. Which error function is best for CV and model selection? Why?

   (e) Reflect on your results. What have you learned about CV? Which approach to model selection to you think is best for this spam classification example? Why?

   Process of Statistical Learning - Decide which classifier is best for building a spam filter.

   (f) Use a model selection procedure to select tuning parameters for each of the following classifiers: Linear SVM, Gaussian Kernel SVM, and Polynomial Kernel SVM.

   (g) Report the accuracy (model assessment) of each classifier for this spam data set. Which one is best? Why? Interpret and reflect on your results.

   (h) Discuss why your model selection and assessment procedures are correct and justify any decisions you made.

     Note: For parts (f-h), you may use any built in functions. The question is purposefully vague as it is up to you to design and implement a correct model selection and model assessment scheme to decide which type of SVM classifier is best for building a spam filter.

2. Handwritten digits. Use PCA, NMF, and ICA to find patterns, reduce the dimension, and visualize the data. Please download the Digits Data from the ESL webpage.

   (a) Visualize results from the 3 methods. How would you visualize patterns among the samples? Among the features? Show these graphics, explain them, and interpret the results. What do these reveal? Did you find anything interesting?

   (b) How much variance is explained by each PC? What would be a good number of PC factors to retain for this data? Explain.
(c) How do results of ICA and NMF change when you take \( r = 10, 20, 50, 250 \) factors? Is there a way that you could decide how many factors to retain in a data-driven manner? Explain.

(d) Is there a quantitative and objective way that you can determine which is the best pattern recognition technique for this data set? How? Explain and implement your procedure.

Stat 640:


2. Non-Negative Matrix Factorization. Derive an algorithm to find a local optima of the non-negative matrix factorization problem commonly employed for count-valued data:

\[
\max_{W, H} \sum_{i=1}^{n} \sum_{j=1}^{p} \left( X_{ij} \log \left( \sum_{k=1}^{K} W_{ik} H_{kj} \right) - \sum_{k=1}^{K} W_{ik} H_{kj} \right)
\]

subject to \( W_{ik} \geq 0 \) & \( H_{kj} \geq 0 \) \( \forall i = 1, \ldots, n; j = 1 \ldots p; k = 1 \ldots K. \)

Hint: There are several possible algorithmic strategies that can be used. If you employ an MM (majorize-minimize) algorithm, you may find ESL problem 14.23 helpful.