IEOR E4525
Martin Haugh

Due: Monday 15 February 2016

Assignment 2

Answer Questions 1 and 2 and two of Questions 3, 4, 5 and 6

1. *(Naive Bayes and spam filtering)*

   (a) Use the spam data from Assignment 1 and naive Bayes to build a classifier that distinguishes spam from non-spam. You may do this in `R` or `Matlab`. (In `Matlab` for example, you could use the **NaiveBayes** functionality.) Your code should split the data into training and test sets and then estimate the generalization error of your classifier.

   (b) Randomly assign 80% of your data to the training set, 20% to the test set and now estimate the test error, E_{test}, of your classifier. Repeat this 10 times. How much variability do you see in E_{test}? What conclusions can you draw from this?

   (c) There are two types of error that a spam classifier can make. Should these errors be treated equally when constructing a classifier. Can we adapt our naive Bayes classifier to reflect this?

   Additional Comments:

   (a) A common approach to handling categorical data in naive Bayes is to use the multinomial distribution. (If you're using `Matlab`, for example, take a look at the help files for naive Bayes and you'll see a description of how to use it.)

   (b) When dealing with an independent categorical variable be careful not to assign a probability of zero to a given class simply because that class was not observed in the training data together with a particular value of that categorical variable. (If you use some built-in functions then these functions may or may not handle this issue automatically but you should investigate if this is the case and adapt your code if necessary. *Hint*: See *Laplace smoothing* which is implemented by the naive Bayes classifier in the *e1071* package in `R`.)

   (c) The "spampct" variable has a lot of missing data. There are a couple of issues to consider here:

      (a) Is there any relation between the missingness of "spampct" and whether or not the email is spam? If so, then there is information here that you can use by creating a new binary variable. If not, then you probably can safely ignore "spampct" when it's missing.

      (b) When the "spampct" variable has a numerical value, you can use it in your classifier (assuming you think it's informative.) If it doesn't have a numerical value you can ignore it completely as this case will be handled by (a). In

particular you can simply set the ratio $\hat{P}(X|\text{spam})/\hat{P}(X|\text{not spam})$ to 1 if you are missing the value of X on a particular training sample.

2. *(Reduced-Rank LDA)*
Let $\mathbf{B}$ and $\mathbf{W}$ be positive definite matrices and consider the following problem of maximizing the so-called *Raleigh quotient*:

$$\max_{\mathbf{a}} \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}}. \tag{1}$$

   (a) Use the method of Lagrange multipliers to solve this problem. In particular show that the optimal $\mathbf{a}^*$ is an eigen vector of a particular matrix. What is this matrix and what eigen vector does $\mathbf{a}^*$ correspond to?

   (b) By identifying $\mathbf{B}$ and $\mathbf{W}$ with the between-class and within-class covariance matrices, we can interpret the problem in (1) as the problem of finding the linear combination, $\mathbf{a}^\top \mathbf{x}$ so as to maximize the *between-class* variance relative to the *within-class* variance. Show that ${\mathbf{a}^*}^\top \mathbf{x}$ is the first discriminant variable.

   Hint: First note that $\mathbf{W} \equiv \mathbf{\Sigma}$ from the lecture slides and that $\mathbf{B}^* = \mathbf{D}^{-1/2}\mathbf{U}^\top \mathbf{B}\mathbf{U}\mathbf{D}^{-1/2}$ where $\mathbf{B}$ is as in (1) and the other matrices are as defined in the slides.

   Remark: The l^{th} discriminant direction, $\mathbf{a}_l^*$ say, can be obtained as the solution to (1) but with the additional constraint that $\mathbf{a}_l^*$ be $\mathbf{W}$-orthogonal to $\mathbf{a}_1^*, \ldots, \mathbf{a}_{l-1}^*$ where we say $\mathbf{u}$ and $\mathbf{v}$ are $\mathbf{W}$-orthogonal if $\mathbf{u}^\top \mathbf{W}\mathbf{v} = 0$.

3. *(Reproducing the reduced-rank plots for HTF's vowel data)*

   (a) Download `Hwk_Reduced_Rank_FDA.R` from *CourseWorks* and add additional code to this script so that you can reproduce any of the sub-figures in Figure 4.8 of HTF which is also on the lecture slides. (The `R Lab` from Chapter 3 of ISLR shows you how the `lda` function in the `MASS` package works.)

   (b) Now perform a reduced-rank LDA in each of the subspaces H_L for $L = 1, \ldots, 10$. As part of your solution you should reproduce Figure 4.10 of HTF (which is also on the lecture slides). Note that you can use the `LDA` function again here.

   A description of the vowel data and associated classification problem can be found at
   http://www-stat.stanford.edu/ tibs/ElemStatLearn/datasets/vowel.info

4. *(Logistic regression is a linear classifier and convergence of IRWLS)*

   (a) Show that binary classification using logistic regression yields a linear classifier.

(b) Consider a binary classification problem where the data is linearly separable. Show that the likelihood function for logistic regression can then be made arbitrarily close to 1. *Hint*: consider a vector $\mathbf{w}$ that separates the data.

(c) What does this observation from part (b) say about convergence of the MLE algorithm? Can you propose a method for resolving this issue?

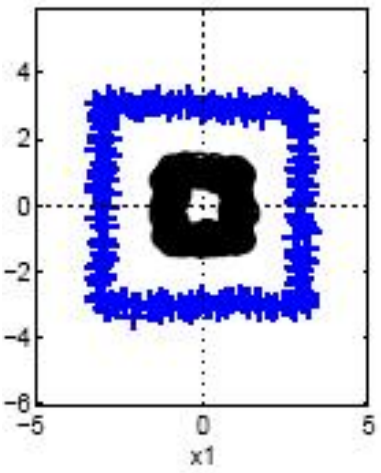
5. *(Comparing logistic regression with Gaussian naive Bayes)*
Consider a naive Bayes classifier for a binary classification problem where all the class-conditional distributions are assumed to be Gaussian with the variance of each feature, X_j, being *equal* across the two classes. That is we assume $(X_j \,|\, G = k) \sim \mathrm{N}(\mu_{jk}, \sigma_j^2)$ for $k = 0, 1$.

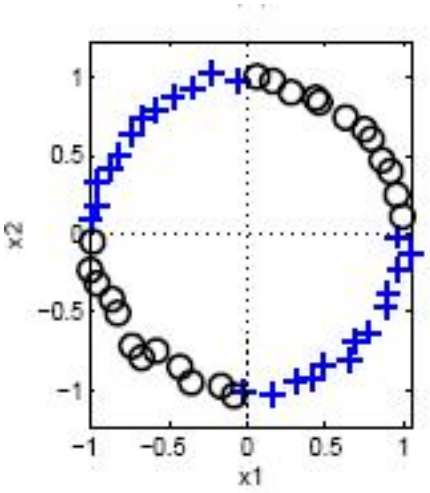
(a) Show that the decision boundary is a linear function of $\mathbf{X} = (X_1, \ldots, X_m)$ and hence that it has the same parametric form as the decision boundary given by logistic regression.

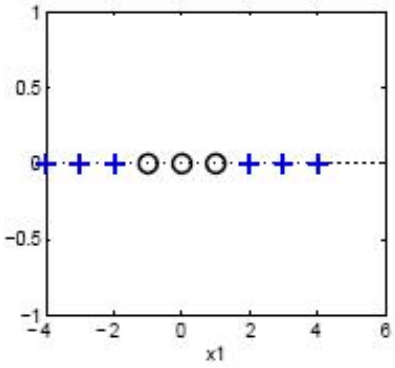
(b) Does the result of part (a) imply that in this case, Gaussian naive Bayes and logistic regression will find the same decision boundary? Justify your answer.

(c) If indeed the class conditional distributions are Gaussian with $(X_j \,|\, G = k) \sim \mathrm{N}(\mu_{jk}, \sigma_j^2)$ and the assumptions of naive Bayes are true, which classifier do you think will be "better": the naive Bayes classifier of part (a) or logistic regression? Justify your answer.

6. In each of the the following three figures there are two classes denoted by the colors blue and black. In each case, the classes are not linearly separable. For each case suggest basis functions which would render the classes linearly separable.




1
0.5
0
−0.5
−1
x2
−1
−0.5
0
0.5
1
x1


1
0.5
0
−0.5
−1
−4
−2
0
2
4
6
x1