CSC2515 Fall 2015 Assignment 1

January 27, 2015

Instructions

There are several questions on this assignment; only the last question involves coding. Please turn in your write-up at the beginning of class on Feb 9th (or submit it electronically into CDF), but do not attach your code. For your implementation, submit it electronically.

Late assignments will have 25% subtracted from the total out of which they are graded for each day or part of a day that they are late. They will be more than one day late if they are not emailed within 24 hours of the due date/time.

1 Naive Bayes and Logistic Regression

1. (5 points)

   Write down and briefly describe the objective functions that Naive Bayes (NB) and Logistic Regression (LR) optimize.

2. (10 points)

   Given a two class problem, $Y \in \{0, 1\}$, derive the posterior class probabilities, $P(Y = i|X)$ given Poisson distributed class conditional densities of the form:

   $$P(X|Y = i) = \prod_j \frac{\lambda_i^{x_j}}{x_j!} e^{-\lambda_i}$$

   and unspecified priors $P(Y = i)$. [Hint: The Bishop book contains part of the derivation for multi-variate Gaussian class-conditional densities.]

3. (20 points)

   Consider the two class problem where class label $y \in \{T, F\}$ and each training example X has 2 binary attributes $X_1, X_2 \in \{T, F\}$. Let the class prior be $P(Y =$

$T) = 0.5$ and also let $P(X_1 = T|Y = T) = 0.8$ and $P(X_1 = F|Y = F) = 0.7$, $P(X_2 = T|Y = T) = 0.5$ and $P(X_2 = F|Y = F) = 0.9$. For this problem, you should assume that the *true* distribution of X_1, X_2, and Y satisfies the Naive Bayes assumption of conditional independence with the above parameters.

(a) Assume X_1 and X_2 are truly independent given Y. Write down the Naive Bayes decision rule given $X_1 = x_1$ and $X_2 = x_2$.

(b) Which attribute provides stronger evidence about the class label? Hint: a formal way to do this is to use mutual information, which for two random variables X and Y is defined as: $I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right)$.

(c) Now suppose that we create a new attribute X_3, which is an exact copy of X_2 (i.e., for every training example, attributes X_2 and X_3 have the same value, $X_2 = X_3$). Are X_2 and X_3 conditionally independent given Y?

(d) The error rate is defined as the summed probability that the decision rule misclassifies an observation generated by either class (see Figure 1.24 in Bishop for a relevant illustration). Would you expect the error rate of naive Bayes to increase when we add the attribute X_3 as above? Why or why not? For this question it is sufficient to just use intuition.

(e) Would logistic regression be affected by the addition of X_3? Explain why or why not.

2 Generalization and Model Complexity

This question asks you to show your general understanding of underfitting and overfitting as they relate to model complexity and training set size. Consider a continuous domain and a smooth joint distribution over inputs and outputs, so that no test or training case is ever duplicated exactly.

1. (5 points)

   For a fixed training set size, sketch a graph of the typical behavior of training error rate versus model complexity in a learning system. Add to this graph a curve showing the typical behavior of the corresponding test error rate (for an infinite test set drawn independently from the same joint distribution as the training set) versus model complexity, on the same axes. Mark a vertical line showing where you think the most complex model your data supports is; choose your horizontal range so that this line is neither on the extreme left nor on the extreme right. Mark a horizontal line showing the error of the optimal Bayes classifier. Indicate on your vertical axes where zero error is and draw your graphs with increasing error upwards and increasing complexity rightwards.

2. (5 points)

   For a fixed model complexity, sketch a graph of the typical behavior of training error rate versus training set size in a learning system. Add to this graph a curve showing the typical behavior of test error rate (again on an iid infinite test set) versus training set size, on the same axes. Mark a horizontal line showing the error of the optimal Bayes classifier. Indicate on your vertical axes where zero error is and draw your graphs with increasing error upwards and increasing training set size rightwards.

3 Experimenting with Logistic Regression and Naive Bayes

For this part you will compare Naive Bayes and Logistic Regression (LR) on a data set you can download from the course website.

The data set that we have chosen is about spam; you are asked to train a binary classifier to classify each instance as being spam or not (ham). The data set was compiled by Prof. Sam Roweis, using his personal emails, a few years ago. Each instance represents an individual email which has been summarized using 185 binary features. Features represent the occurrence of particular words which are indicative of spam.

The zip file on the course's website contains a Matlab version of the data set. The data set has been split into 1000 training instances and 4000 validation instances. In addition to the labelled instances, the features' names, that is the words they represent, are in the `feature_names` variable. Your job is to complete the functions in the files found in the my_code directory.

Here are details of the implementations.

Logistic Regression: We have provided you with a template for a Logistic Regression classifier. You will need to fill in the remaining components of the LR classifier, including an ℓ_2 regularizer. You will also need to implement a checkgrad function to make sure that your gradients are correct.

For LR, you will need to choose some parameters. Reasonable parameters for training are: weight initialization with 0.01*randn; a learning rate between 0.001 and 0.1; and an ℓ_2 penalty coefficient of between 0.01 and 1.

Naive Bayes: Implement the decision rule of a Naive Bayes classifier. When choosing how to model the class conditional densities keep in mind that the instance features are binary.

Below you will find details of what to turn in.

1. (20 points)

   Provide a brief description of both of your implementations (pseudocode). For LR, the description must include the equations for the negative log-likelihood and it's gradient with respect to the weights. For NB you should include your choice of distribution for the class conditional densities as well as for the prior class distribution. Make sure that you properly, yet succinctly, justify your choices. This should not be a printout of your code, but a high-level outline. [A paragraph or two]

2. (15 points)

   Train the weights in logistic regression with no regularization for 2000 iterations. Plot training and validation curves. Explain how you can stop training based on validation error. Compare the performance of the system if you stop based on the training error versus the validation error. Hint: For the optimization of the loss function we suggest you use steepest descent but let us know if you have used conjugate gradient descent.

   [One or two plots and a few sentences.]

3. (15 points)

   Add the regularization, and again plot the training and validation curves. Compare your results to those you obtained without the regularizer. Explain the effects of the regularization. [One or two plots and a few sentences.]

4. (5 points) For LR with regularization report the feature names of the 10 features with the highest learned weights and 10 features with lowest learned weights. [The names, and associated weights.]

5. (10 points) What are the training and validation errors that naive Bayes achieves? Do you think the naive Bayes assumption holds for this data?

6. (5 points) For NB report the feature names of the 10 features with the highest learned weights and 10 features with lowest learned weights.