EE 511 – Assignment 3

Due: 2 Feb '19

Part I: written Assignment (6 points)

1. You have a classification problem where you need to predict what was the cause of a power outage, where the classes are: weather, bird/animal, vehicle accident (e.g. truck hits a telephone pole), equipment failure, or other. Your input vector includes:

   - four measurements for each of the following from the past 48 hours: temperature, humidity, barometric pressure, wind speed
   - wind direction (8 possible categories) associated with the wind measurements
   - rainfall in inches in the past two days
   - month, day of the week, and time of day of the outage

   You are given ten years worth of labeled data to learn your classifier.

   (a) Let's assume that no one has ever worked on this problem before. What would you use as a baseline?

   (b) If you designed a classifier based on a generative model, what distributions would you use? How many free parameters are in your model?

   (c) If you used logistic regression, what would the dimensionality of your feature vector be?

   (d) Explain how you might control the complexity of each model.

2. In a binary classification problem, assuming that the class conditional distributions are Gaussians is referred to as linear discriminant analysis (LDA) when the covariances are shared, $p(x|c_k) \sim N(\mu_k, \Sigma)$, since the discriminant functions can be simplified to:

$$f_k(x) = x^t \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^t \Sigma^{-1} \mu_k + \log \pi_k = x^t w + w_0$$

   where $P(C = c_k) = \pi_k$, and

$$\mu_k = \frac{1}{n_i} \sum_{i:y_i=k} x_i \qquad \Sigma = \frac{1}{N-2} \sum_{k=1,2} \sum_{i:y_i=k} (x_i - m_k)(x_i - m_k)^t$$

   (the sample mean and unbiased covariance estimate). This solution for w is the same as that obtained using Fisher's linear discriminant, which chooses w to maximize the separation of the distributions of the transformed features $\tilde{x} = w^t x$:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \frac{w^t S_B w}{w^t S_W w} \text{ where } S_B = (m_1 - m_2)(m_1 - m_2)^t \quad S_W = S_1 + S_2$$

where S_W is the shared covariance Σ without the normalization factor. This involves solvinging the generalized eigenvector problem: $S_B w = \lambda S_W w$. Suppose you wish to use LDA with a high dimensional transformation $h(x)$. You can use kernels to simplify the solution by letting

$$w = \sum_{i=1}^{n} \alpha_i h(x_i) \text{ and } \underset{\alpha}{\operatorname{argmax}} \frac{\alpha^t S_B^K \alpha}{\alpha^t S_W^H \alpha}$$

solving for α with kernel function $k(x_i, x_j) =< h(x_i), h(x_j) >$ and plugging back in to solve for w. We showed in class that $S_B^K = (K_1 - K_2)(K_1 - K_2)^t$ where $K_j(i) = \frac{1}{n_j} \sum_{l:y_l=j} k(x_i, x_l)$. Find a similar expression for S_W^K where you use a regularized within-class scatter $S_h + \gamma I$.

Part II: Computer Assignment (14 points)

1. Computer assignment

   For this problem we will be working with the 20 Newsgroups dataset, a collection of posts to twenty different newsgroups from 1997. The task is to predict which of the 20 newsgroups a post belongs to from the text. You get to choose if you want to use a linear Support Vector Machine (SVM) or logistic regression classifier. If you choose the SVM option then you will probably want to take advantage of one of the many available SVM solvers. If you do logistic regression, many libraries support this but it is also a great chance to introduce yourself to one of the deep learning libraries such as Tensorflow.

   (a) **Setup**

   Download and load the training and evaluation data.

   `http://ana.cachopo.org/datasets-for-single-label-text-categorization/20ng-train-all-terms`
   `txt`

   `http://ana.cachopo.org/datasets-for-single-label-text-categorization/20ng-test-all-terms`
   `txt`

   (b) **Vocabulary Selection**

   We will be using a unigram bag-of-words model to represent the text. This means that we consider anything separated by spaces to be a word and treat each document as an unordered collection of words. Each word becomes a feature in our model. Using the full vocabulary is definitely doable but since this is a homework assignment we will speed things up by choosing only a subset of the vocabulary.

   Our vocabulary selection method uses mutual information. Let X_i be a random variable that indicates if word i is present in a document or not and let Y be a random variable that indicates the newsgroup label for a document. For each i in the list of words in the training data, calculate the mutual information:

$$I(X_i; Y) = \sum_{x_i \in \{0,1\}} \sum_{y \in Y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}$$

Zero probabilities will cause problems. To counter this, use add-one (Laplace) smoothing when estimating $p(x_i, y)$. This means that you pretend like there is one extra document per class in the data and that each of those documents uses every word in the corpus.

Rank all the words in descending order of mutual information and select the top 5,000 words. In your report, include a table of the top ten words and their mutual information. As a sanity check you can confirm that all your mutual information values are greater than zero and less than $\log_2(20)$.

Optionally, you can also try using the most frequent 5,000 words.

(c) **Input Representation**

There are a few different ways to represent a document as a vector. You should assess three possibilities. (Note that for each of the three we will be ignoring all the words that are not in our selected vocabulary.) Method one is to use binary indicator variables. Method two is to use the raw counts where each entry in the feature vector holds the count of that word in the document. Each entry in the vector is set to $\max(1, t_d)$ where t_d is the number of occurences of term t in document d. Method three is to use the log-normalized counts where each entry becomes $\log(t_d + 1)$.

(d) **Classifier**

Choose either the logistic regression or linear support vector machine classifier and train a model on the data for each of the three input representations. For the SVM option, use the one-vs-rest classification strategy. For this assignment we are not specifying how to create the validation set. Choose some reasonable way of creating a held-out set from the training data or use k-fold cross validation. Just make sure to describe what you did in your write-up. Use the validation data to pick an input representation and to choose any hyperparameters for your classifier such as regularization penalties. Optionally, you can also assess the most frequent word vocabulary for the best feature set. If it is taking too long to train the classifier on your computer, feel free to try using a smaller vocabulary.

As a sanity check, calculate the accuracy of a model that always predicts the most frequent class. If your logistic regression or SVM model is not doing better than this baseline then you are doing something wrong.

(e) **Evaluation**

Evaluate your best model on the test data and report the accuracy. What are the five top misclassified label pairs in terms of absolute number of misclassifications? Generate the confusion matrix and find the five largest off-diagonal entries. Comment on whether these errors make sense given the topics.

(f) **Model Inspection**

In this part we will take a look at the weights of our model to see if we can gain any insight into how our model is working. Make a table of the ten features with the highest weights for each class. Do the same for the ten features with the most negative weight for each class. Compare the results to what you found using mutual information.