Stat 406 Spring 2010: homework 4

1 Logistic regression vs LDA/QDA

(Source: Jaakkola)
Suppose we train the following binary classifiers via maximum likelihood.

1. GaussI: A generative classifier, where the class conditional densities are Gaussian, with both covariance matrices set to $\mathbf{I}$ (identity matrix), i.e., $p(\mathbf{x}|y = c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{I})$. We assume $p(y)$ is uniform.

2. GaussX: as for GaussI, but the covariance matrices are unconstrained, i.e., $p(\mathbf{x}|y = c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.

3. LinLog: A logistic regression model with linear features.

4. QuadLog: A logistic regression model, using linear and quadratic features (i.e., polynomial basis function expansion of degree 2).

After training we compute the performance of each model M on the training set as follows:

$$L(M) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}, M) \tag{1}$$

(Note that this is the *conditional* log-likelihood $p(y|\mathbf{x}, \hat{\boldsymbol{\theta}})$ and not the joint log-likelihood $p(y, \mathbf{x}|\hat{\boldsymbol{\theta}})$.) We now want to compare the performance of each model. We will write $L(M) \le L(M')$ if model M *must* have lower (or equal) log likelihood (on the training set) than M', for any training set (in other words, M is worse than M', at least as far as training set logprob is concerned). For each of the following model pairs, state whether $L(M) \le L(M')$, $L(M) \ge L(M')$, or whether no such statement can be made (i.e., M might sometimes be better than M' and sometimes worse); also, for each question, briefly (1-2 sentences) explain why.

1. GaussI, LinLog.

2. GaussX, QuadLog.

3. LinLog, QuadLog.

4. GaussI, QuadLog.

5. Now suppose we measure performance in terms of the average misclassification rate on the training set:

$$R(M) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \ne \hat{y}(\mathbf{x}_i)) \tag{2}$$

   Is it true in general that $L(M) > L(M')$ implies that $R(M) < R(M')$? Explain why or why not.

2 Class conditional densities for binary data

Consider a generative classifier for C classes with class conditional density $p(\mathbf{x}|y)$ and uniform class prior $p(y)$. Suppose all the d features are binary, $x_j \in \{0, 1\}$. If we assume all the features are conditionally independent (the naive Bayes assumption), we can write

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{d} \text{Ber}(x_j|\theta_{jc}) \tag{3}$$

This requires dC parameters.

1. Now consider a different model, which we will call the "full" model, in which all the features are fully dependent (i.e., we make no factorization assumptions). How might we represent $p(\mathbf{x}|y = c)$ in this case? How many parameters are needed to represent $p(\mathbf{x}|y = c)$?
2. Assume the number of features d is fixed. Let there be n training cases. If the sample size n is very small, which model (naive Bayes or full) is likely to give lower test set error, and why?
3. If the sample size n is very large, which model (naive Bayes or full) is likely to give lower test set error, and why?
4. What is the computational complexity of fitting the full and naive Bayes models as a function of n and d? Use big-Oh notation. (Fitting the model here means computing the MLE or MAP parameter estimates. You may assume you can convert a d-bit vector to an array index in $O(d)$ time.)
5. What is the computational complexity of applying the full and naive Bayes models at test time to a single test case?

3 Spam classification using naive Bayes

We will re-examine the dataset from Question 3 in homework 3.

1. Use `naiveBayesBerFit` and `naiveBayesBerPredict` on the binarized spam data. What is the training and test error? (You can try different settings of the pseudocount α if you like (this corresponds to the Beta(α, α) prior each θ_{jc}), although the default of $\alpha = 1$ is probably fine.) Turn in your error rates.
2. Modify the code so it can handle real-valued features. Use a Gaussian density for each feature; fit it with maximum likelihood. What are the training and test error rates on the standardized data and the log transformed data? Turn in your 4 error rates and code.