COMS 4771: Machine Learning Due: Feb 9^{th} 2016

Homework 1 - Solutions

Instructor: *Satyen Kale*

Part 1

When using a Naive Bayes Classifier, we have 2 kinds of parameters which we estimate from the training set using maximum likelihood. The first, $\hat{\pi}$, is the parameter describing the class prior. In the `usps.mat` dataset, we have 10 possible labels, $y \in \{0, 1, \ldots 9\}$, and $\hat{\pi}_y$ gives the Maximum Likelihood Estimate of the probability of seeing the class y, i.e. $\Pr[Y = y]$. The other parameters describe the class conditional Bernoulli distributions, where we have 1 Bernoulli parameter, $\hat{\theta}_{j,y}$, for every binary feature j and class y, and is the Maximum Likelihood Estimate for the probability $\Pr[X_j = 1 \mid Y = y]$.

For the class prior parameters $\hat{\pi}_y$, we get the Maximum Likelihood Estimate as follows:

$$\hat{\pi}_y = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i = y\}$$

There is a parameter $\hat{\pi}_y$ for every value of k, ie., we have 10 such $\hat{\pi}_y$.

For the Bernoulli parameters describing the class conditional distributions, we get the following Maximum Likelihood Estimate:

$$\hat{\theta}_{j,y} = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_{i,j} = 1 \mid y_i = y\}}{\sum_{i=1}^{n} \mathbb{1}\{y_i = y\}}$$

$$\therefore \ \hat{\theta}_{j,y} = \frac{1}{n\hat{\pi}_y} \sum_{i=1}^{n} \mathbb{1}\{x_{i,j} = 1 \mid y_i = y\}$$

In the `usps.mat` dataset, we have 10 labels and 256 input dimensions, so we have 2560 class conditional parameters $\hat{\theta}_{j,y}$.

For any given input image x, the Bayes classifier predicts the label

$$\arg\max_y \Pr[Y = y] \Pr[X = x \mid Y = y] = \arg\max_y \Pr[Y = y] \prod_{j=1}^{d} \Pr[X_j = x_j \mid Y = y],$$

using the Naïve Bayes assumption that the features of independent conditioned on the class.

The Naïve Bayes classifier computes its prediction by plugging-in probabilities computed using Maximum Likelihood Estimation of the parameters. Since the features are assumed to be drawn from a Bernoulli distribution, the prediction thus becomes:

$$\arg\max_y \ \hat{\pi}_y \prod_{j=1}^{d} (\theta_{j,y} x_j + (1 - \theta_{j,y})(1 - x_j)),$$

since for a random variable Z drawn from a Bernoulli distribution with parameter θ, we have

$$\Pr[Z = z] \;=\; \theta z + (1-\theta)(1-z).$$

Part 2

Training error of Naïve Bayes Classifier: 0.1582 (1582 errors out of 10,000 examples)
Test error of Naïve Bayes Classifier: 0.167 (167 errors out of 1000 examples)

Part 3

In the k-Nearest Neighbors algorithm, to predict the class label of a given data point, we have to compute its distance from all the training data points, find k points that are closest to it, and assign it the label that is the plurality of the labels of those k points.

In the assignment, we have to compute the training loss using the first 1000 training examples and the test loss on the test data set. When $k = 1$ (and the training dataset does not have two different values of Y for the same X), the training loss must be 0, as we give any example the label of its closest neighbor. In the case of $k = 1$, every training example gets its own label, so there is no misclassification error. This property also provides us with a nice way to verify if the implementation of k-NN is correct.

Results of computing training and test error with k-NN on the `usps.mat` dataset:

	$k = 1$	$k = 3$	$k = 5$
Training error	0	0.0450	0.0580
Test error	0.0840	0.0850	0.0910