Homework 3: Gaussian Classification

Brown University CSCI1420 & ENGN2520: Machine Learning

Homework due at 11:59pm on October 8, 2015

Question 1:

We begin by considering examples, produced by a sophisticated simulator, of data which might be collected by a gamma telescope observing high energy particles. The raw data, “showers” of particles on a planar detector, have been converted into 10 continuous features as outlined here: http://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope. Our goal is the binary classification of the “primary” gamma signals of scientific interest from background, hadronic shower events.

We have converted this data to Matlab format. The $D = 10$ continuous features for each of the $N = 15,216$ training examples are stored in a $N \times D$ matrix $\text{train}$. The class labels are stored in an $N \times 1$ vector $\text{trainLabels}$, where primary gamma signals have label 1 and background events have label 0. Similarly, test data is stored in $\text{test}$ and $\text{testLabels}$. We will model the gamma telescope data with a naive Bayes model, in which a Gaussian distribution is used to model each feature of each class. Each of these distributions has a potentially distinct mean and variance.

a) Give an equation for the joint log-likelihood of the training data under this naive Bayes model. Let $x_{nd} \in \mathbb{R}$ be the value of feature $d$ for training example $n$, $t_n \in \{0, 1\}$ the class label for example $n$, $\pi$ the probability of gamma signals (class 1), $\mu_{kd}$ the mean of feature $d$ for examples of class $k$, and $\sigma^2_{kd}$ the variance of feature $d$ for examples of class $k$.

b) Specify the equations for maximum likelihood (ML) estimation of the model parameters from the training data. A detailed derivation is not necessary, but include some explanation for why your equations are correct.

c) Implement this ML parameter estimation algorithm. Compute and plot an ROC curve on the test data to evaluate your classifier.

d) Suppose that the frequencies of the classes are as in the training data, and that all errors are equally costly. Determine the optimal Bayesian classification rule. What are the true positive rate and false positive rate of this rule on the test data?

e) Suppose that the frequencies of the classes are as in the training data, and that it is 50 times more costly to classify signals as background (missed detections) as to classify background as signals (false alarms). Determine the optimal Bayesian classification rule. What are the true positive rate and false positive rate of this rule on the test data?
Question 2:

In this question, we consider the handwritten digit data from Homework 1, and compare a naive Bayes classifier to the nearest neighbor classifier. Because the digits are represented by continuous intensities, we use a Gaussian distribution to model each feature (pixel).

a) Reuse your naive Bayes parameter estimation code from question 1 to determine maximum likelihood (ML) estimates of the Gaussian mean $\mu_{kd}$ and variance $\sigma_{kd}^2$ parameters for each feature/pixel, $d$, of each class, $k$. For some “background” pixels near image boundaries, the resulting variance estimates may be extremely small (or even 0). To avoid numerical problems, threshold all estimated variances to a minimum value of $10^{-6}$. Use the same 200 training examples per class ($N = 600$ total examples) considered in Homework 1, and plot the mean and variance of each class as images.

b) Using the naive Bayes models from part (a), estimate categorization accuracy on the held-out test data. How does this compare to the best nearest-neighbor classifier? How does the computational cost of the naive Bayes and nearest-neighbor classifiers compare?

c) Suppose that we trained both the nearest neighbor and naive Bayes classifiers on a much larger database, with 6,000 examples per class. Which classifier do you think would see the larger increase in accuracy from using more training data? Why?

We now consider a modification of the standard Gaussian naive Bayes model. As before, we provide a different mean parameter $\mu_{kd}$ for each of the $D$ features/pixels of each class. However, we constrain all features of each class to have the same variance $\sigma_k^2$, so that

$$p(x_n \mid t_n = k, \mu_k, \sigma_k^2) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left\{ -\frac{(x_{nd} - \mu_{kd})^2}{2\sigma_k^2} \right\}$$

For each class of handwritten digit, there are $D + 1 = 785$ unknown parameters: 784 scalar mean parameters, and 1 scalar variance parameter.

d) Derive formulas for the maximum likelihood (ML) estimates of the mean and variance parameters of the model above. Simplify your answers as much as possible.

e) Using formulas from part (d), estimate constrained naive Bayes models for each of the three digit classes, using the same 200 training examples per class ($N = 600$ total examples). What is the categorization accuracy of the resulting models on the held-out test data? How does this compare to the unconstrained naive Bayes classifier from part (b)?

f) Consider an extremely simple classifier, which measures the Euclidean distance of the test digit to each of the three mean vectors estimated in part (e), and assigns the class label of the closest. This is like applying a nearest-neighbor classifier to a “synthetic” dataset with only one example per class. What is the performance of this classifier? Is it equivalent to the constrained naive Bayes classifier from part (e)? Why or why not?
Question 3:

This question asks you to devise ML and Bayesian MAP estimators for a simple model of an uncalibrated sensor. Let the sensor output, $X$, be a random variable that ranges over the positive real numbers. We assume that, when tested over a range of environments, its outputs are uniformly distributed on some unknown interval $[0, \theta]$, so that

$$
p(x \mid \theta) = \begin{cases} 
\frac{1}{\theta} & \text{if } 0 \leq x \leq \theta, \\
0 & \text{otherwise}, 
\end{cases}
= \frac{1}{\theta} \mathbb{I}_{0,\theta}(x).
$$

Here, $\mathbb{I}_{0,\theta}(x)$ denotes an indicator function which equals 1 when $0 \leq x \leq \theta$, and 0 otherwise. We denote this distribution by $X \sim \text{Unif}(0, \theta)$. To characterize the sensor’s sensitivity, we would like to infer $\theta$.

(a) Given $N$ i.i.d. observations $x = (x_1, \ldots, x_N)$, $X_i \sim \text{Unif}(0, \theta)$, what is the likelihood function $p(x \mid \theta)$? What is the maximum likelihood (ML) estimator for $\theta$? Give an informal proof that your estimator is in fact the ML estimator.

(b) Suppose that we place the following prior distribution on $\theta$:

$$
p(\theta) = \alpha \beta^{\alpha} \theta^{-\alpha-1} \mathbb{I}_{\beta,\infty}(\theta).
$$

This is known as a Pareto distribution. We denote it by $\theta \sim \text{Pareto}(\alpha, \beta)$. Plot the three prior probability densities corresponding to the following three hyperparameter choices: $(\alpha, \beta) = (0.1, 0.1); (\alpha, \beta) = (2.0, 0.1); (\alpha, \beta) = (1.0, 2.0)$. Briefly describe the influence these parameters have on the properties of the Pareto distribution.

(c) If $\theta \sim \text{Pareto}(\alpha, \beta)$ and we observe $N$ uniformly distributed observations $X_i \sim \text{Unif}(0, \theta)$, derive the posterior distribution $p(\theta \mid x)$. Is this a member of any standard family?

(d) For the posterior derived in part (c), what is the corresponding MAP estimator of $\theta$? How does this compare to the ML estimator?

(e) Recall that the quadratic loss is defined as $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. For the posterior derived in part (c), what estimator of $\theta$ minimizes the posterior expected quadratic loss? Simplify your answer as much as possible.

(f) Suppose that we observe three observations $x = (0.7, 1.3, 1.7)$. Determine the posterior distribution of $\theta$ for each of the priors in part (b), and plot the corresponding posterior densities. What is the MAP estimate for each hyperparameter choice? What estimate minimizes the quadratic loss for each hyperparameter choice?