# CSCI 1950-F Homework 2: ML & Bayesian Estimation

Brown University, Spring 2012

Homework due at 12:00pm on February 16, 2012

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 `train`. The class labels are stored in an $N \times 1$ vector `trainLabels`, where primary gamma signals have label 1 and background events have label 0. Similarly, test data is stored in `test` and `testLabels`.

**Question 1:**

*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 this naive Bayes model, defining parameters as appropriate.*

*b) Specify the equations for maximum likelihood (ML) estimation of the model parameters from the training data.*

*c) Implement this 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?*

Now consider a binary categorization problem, where we want to assign a label $y_i \in \{0, 1\}$ given observation $x_i$. Let $\rho_i = p(Y_i = 1 \mid x_i)$ and $1 - \rho_i = p(Y_i = 0 \mid x_i)$. Suppose that the classifier $\hat{y}(x_i)$ is allowed to make one of three decisions: choose class 0, choose class 1, or "reject" this data (refuse to make a decision). We can use a Bayesian decision theoretic approach to tradeoff the losses incurred by incorrect decisions and rejections.

**Question 2:**

*Suppose that the classifier incurs a loss of* 0 *whenever it chooses the correct class, a loss of* 1 *whenever it chooses the wrong class, and a loss of* $\lambda$ *whenever it selects the reject option. Express the optimal decision rule* $\hat{y}(x_i)$*, which minimizes the posterior expected loss, as a function of* $\rho_i$ *and* $\lambda \geq 0$*. Simplify your answer as much as possible.*

The next 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 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

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

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 $\theta$.

**Question 3:**

*a) Given N i.i.d. observations* $D = (x_1, \ldots, x_N)$*,* $X_i \sim \text{Unif}(0, \theta)$*, 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:* $(\alpha, \beta) = (0.1, 0.1)$*;* $(\alpha, \beta) = (2.0, 0.1)$*;* $(\alpha, \beta) = (1.0, 2.0)$*.*

*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 estimator for each hyperparameter choice? What estimator minimizes the quadratic loss for each hyperparameter choice?*

In this next question, we explore the geometry of the *receiver operating characteristic* (ROC) curves discussed in lecture. Let $\mathcal{Y} = \{0, 1\}$ denote the two possible classes in a binary categorization problem. For $N$ observations $x_i$ of instances with true class labels $y_i$, and any decision rule $\hat{y}(x_i)$, recall the following definitions:

**TP** Total number of *true positives*, which occur when $y_i = 1$, and $\hat{y}(x_i) = 1$.

**FP** Total number of *false positives*, which occur when $y_i = 0$, but $\hat{y}(x_i) = 1$.

**TN** Total number of *true negatives*, which occur when $y_i = 0$, and $\hat{y}(x_i) = 0$.

**FN** Total number of *false negatives*, which occur when $y_i = 1$, but $\hat{y}(x_i) = 0$.

The ROC curve is then a plot of the expected values of the sensitivity or *true positive rate*, $TPR = TP/(TP+FN)$, versus the false alarm or *false positive rate*, $FPR = FP/(FP+TN)$, achieved by some family of decision rules for this dataset.

Our analysis is based on the concept of a *randomized decision rule*. Given two base decision rules $\hat{y}_0(x_i)$, $\hat{y}_1(x_i)$, we classify each observation $x_i$ as follows:

- Sample $z_i \sim \text{Bernoulli}(\gamma)$, for some fixed $\gamma$.
- Select decision $\hat{y}_1(x_i)$ if $z_i = 1$, or decision $\hat{y}_0(x_i)$ if $z_i = 0$.

Varying $\gamma$ between 0 and 1 then creates a new family of decision rules.

**Question 4:**

*a) Consider a randomized decision rule as above. Derive formulas for the TPR and FPR of this decision rule as a function of $\gamma$, and the true positive and false positive rates of the base decision rules.*

*b) Consider the diagonal ROC line for which $TPR = FPR$. Prove that a classifier which achieves any performance on this line can always be constructed, regardless of the joint distribution $p(x_i, y_i)$.*

*c) A set $\Lambda$ is convex if, for any $\alpha \in [0, 1]$ and $\lambda_0, \lambda_1 \in \Lambda$, $(\alpha\lambda_0 + (1-\alpha)\lambda_1) \in \Lambda$. Consider a hypothetical family of decision rules for which the region under the ROC curve is* **not** *convex. Argue that these rules must be sub-optimal, i.e. that there exists a decision rule with equal FPR but higher TPR.*

*d) Suppose you test a learned classifier on a validation set, and discover that the region under the ROC is not convex. How could you construct a better classifier?*

*e) Suppose you test a learned classifier on a validation set, and discover that the $FPR$ is larger than the $TPR$. How could you construct a better classifier?*