CS195f Homework 2

Mark Johnson and Erik Sudderth

Homework due at 2pm, 1st October 2009

The first question asks you to analyse the following naive Bayes model that describes the weather in a mythical country.

$$
\begin{aligned}
\mathcal{Y} &= \{\mathsf{night}, \mathsf{day}\} \\
\mathcal{X}_1 &= \{\mathsf{cold}, \mathsf{hot}\} \\
\mathcal{X}_2 &= \{\mathsf{rain}, \mathsf{dry}\} \\
\mathrm{P}(X_1, X_2, Y) &= \mathrm{P}(Y)\mathrm{P}(X_1 \mid Y)\mathrm{P}(X_2 \mid Y) \\
\mathrm{P}(Y{=}\mathsf{day}) &= 0.5 \\
\mathrm{P}(X_1{=}\mathsf{hot} \mid Y{=}\mathsf{day}) &= 0.9 \\
\mathrm{P}(X_1{=}\mathsf{hot} \mid Y{=}\mathsf{night}) &= 0.2 \\
\mathrm{P}(X_2{=}\mathsf{dry} \mid Y{=}\mathsf{day}) &= 0.75 \\
\mathrm{P}(X_2{=}\mathsf{dry} \mid Y{=}\mathsf{night}) &= 0.4
\end{aligned}
$$

Question 1:

For each of the following formulae except the first, write an equation which defines it in terms of formulae that appear earlier in the list. (For example, you should give a formula for $\mathrm{P}(x_1, x_2)$ *in terms of* $\mathrm{P}(x_1, x_2, y)$*). Then given the model above, calculate and write out the value of the formula for possible each combination of values of the variables that appear in it.*

a) $\mathrm{P}(x_1, x_2, y)$.

b) $\mathrm{P}(x_1, x_2)$.

c) $\mathrm{P}(y \mid x_1, x_2)$.

d) $\mathrm{P}(x_1)$.

e) $\mathrm{P}(x_2)$.

f) $\mathrm{P}(x_1 \mid x_2)$.

g) $\mathrm{P}(x_2 \mid x_1)$.

h) $\mathrm{P}(x_1 \mid x_2, y)$.

i) $\mathrm{P}(x_2 \mid x_1, y)$.

Are X_1 and X_2 conditionally independent given Y? Are X_1 and X_2 marginally independent, integrating over Y? Provide a short proof for both answers.

Consider a binary categorization problem, and let $p(y_i \mid x_i)$ denote the posterior distribution of the latent class label $y_i \in \{0, 1\}$ given observation 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 λ 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 $p(y_i \mid x_i)$ and λ. 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 $[\theta_a, \theta_b]$, so that

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

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

Question 3:

Suppose that we are certain that $\theta_a = 0$, so that the only unknown parameter is $\theta_b \triangleq \theta$.

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

b) Suppose that we place the following prior distribution on θ:

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

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:* $(\alpha, \beta) = (0.1, 0.1)$; $(\alpha, \beta) = (2.0, 0.1)$; $(\alpha, \beta) = (1.0, 1.0)$.

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* θ*? 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* θ *minimizes the posterior expected quadratic loss? Simplify your answer as much as possible.*

f) *Suppose that we observe three observations* $x = (0.7, 1.3, 1.9)$. *Determine the posterior distribution of* θ *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?*

Question 4: (200-level credit)

a) *Consider a continuous parameter* $\theta \in \Theta$, *where* $\Theta \subset \mathbb{R}$ *is convex. Let* $p(\theta)$ *denote the prior distribution, and suppose that an observation* x *is observed with likelihood model* $p(x \mid \theta)$. *Derive the general form of the estimator* $\hat{\theta}(x)$ *which minimizes the posterior expected loss, where* $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$.

b) *Let* θ *be the probability that a possibly biased coin comes up heads. Let* $\theta \sim \text{Beta}(0.5, 0.5)$ *be our prior distribution (this reference prior could be motivated by objective Bayesian arguments). Suppose that out of 5 i.i.d. observations* $x_i \sim \text{Bernoulli}(\theta)$, *3 come up heads, and 2 tails. What is the posterior distribution* $p(\theta \mid x)$, *where* $x = (x_1, \ldots x_5)$*? Plot both the prior* $p(\theta)$ *and posterior* $p(\theta \mid x)$.

c) *Propose and implement a numerical method for approximating the estimator* $\hat{\theta}(x)$ *which minimizes the posterior expected loss* $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, *for the data in part (b). Compare your answer to the corresponding ML estimate.*

d) *Repeat parts (b-c) for the case where all five observations come up heads.*

e) *Generalize your answer in part (a) to the case where* $\theta \in \mathbb{R}^d$, *and* $L(\theta, \hat{\theta}) = \sum_{i=1}^{d} |\theta_i - \hat{\theta}_i|$.

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 *sensitivity*, or $TP/(TP+FN)$, versus *specificity*, or $TN/(TN+FP)$, 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 $\gamma = \text{P}[z_i = 1]$.
- Select decision $\hat{y}_1(x_i)$ if $z_i = 1$, or decision $\hat{y}_0(x_i)$ if $z_i = 0$.

Varying γ between 0 and 1 then creates a new family of decision rules.

Question 5: (200-level credit)

a) Consider a randomized decision rule as above. Derive formulas for the sensitivity and specificity of this decision rule as a function of γ, and the sensitivity and specificity of the base decision rules.

b) Consider the diagonal ROC line for which sensitivity = 1-specificity. 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 Λ 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 specificity but higher sensitivity.*

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?