# EEL709: Re-Minor I

April 13, 2013

Maximum Marks: 25

1. The AIIMS has been running a trial of the fecal occult blood (FOB) test for screening bowel cancer. They have data from 1000 patients, for each of whom the actual presence or absence of bowel cancer has been independently confirmed via endoscopy. The results of the trial are as follows:

| | **Confirmed healthy** | **Confirmed cancer** |
|---|---|---|
| **FOB predicted healthy** | 882 | 10 |
| **FOB predicted cancer** | 18 | 90 |

Suppose we set up a simple model for this as follows: $\theta$ is the prior probability of a patient having cancer; $p$ is the probability of a cancer prediction from FOB if the patient is healthy (also called the *False Positive Rate (FPR)*); and $q$ is the probability of a cancer prediction from FOB if the patient has cancer (also called the *True Positive Rate (TPR)*).

(a) Write down the joint likelihood of the data, as a function of the three model parameters $\theta$, $p$, and $q$. Obtain maximum likelihood estimates for each of these parameters. **[4]**

(b) Suppose the loss matrix for this task is defined as follows:

| | **Confirmed healthy** | **Confirmed cancer** |
|---|---|---|
| **FOB predicted healthy** | 0 | 1 |
| **FOB predicted cancer** | $K$ | 0 |

Using the parameter estimates computed above, obtain the expected loss for FOB as a function of $K$. **[2]**

(c) The AIIMS has also run a trial for another diagnostic test, a CT scan (CTS), on the same cohort of 1000 patients. For CTS, the FPR and TPR estimates obtained are $p' = 0.1$ and $q' = 0.96$. Obtain the expected loss for CTS as well; what is the critical value of $K$ below which CTS becomes preferable to FOB? **[3]**

2. One Sunday morning, the arrival times of six successive north-bound trains at the Hauz Khas Metro station were observed to be as follows: 8:04; 8:10; 8:42; 8:47; 8:51; 8:55. Let us denote by $k$ the time interval (in minutes) between the arrival of two successive trains; we will assume that this follows a *Gaussian distribution*, i.e.,

$$p(k|\lambda, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(k-\lambda)^2}{2\sigma^2}\right\}.$$

(a) The expected value of $k$ under the Gaussian distribution is $\lambda$; i.e., in our case, $\lambda$ corresponds to the expected time interval between successive arrivals. Given a sequence of

independent observations of time intervals, $\{k_1, k_2, ..., k_N\}$, write down the likelihood function for this data under the Gaussian distribution, and obtain general expressions for the maximum likelihood estimates for both $\lambda$ and $\sigma$, the standard deviation parameter. Now plug in the Metro data above to get the maximum likelihood estimates for the expected time interval at Hauz Khas station and its standard deviation, based on the five observations. Comment on the meaningfulness of these estimates. **[5]**

(b) Suppose we now treat $\lambda$ as a random variable and put a prior distribution on it. We may use another Gaussian for this:

$$p(\lambda|\alpha, \beta) = \frac{1}{\sqrt{2\pi}\beta} \exp\left\{-\frac{(\lambda - \alpha)^2}{2\beta^2}\right\}.$$

Here the $\alpha$ and $\beta$ are 'hyperparameters' to be chosen. Use this prior and the likelihood function from part (a) to obtain the posterior distribution of $\lambda$, as a function of the observed data, $\alpha$, $\beta$, and $\sigma$ (which we assume to be fixed). Derive a general expression for the maximum a posteriori (MAP) estimate for $\lambda$. **[3]**

(c) The parameter $\beta$ above specifies the spread or variance of the prior distribution. Suppose we set $\beta = \sigma/3$. Suppose also that you have heard somewhere that the time interval between successive north-bound trains at Hauz Khas station is about 5 minutes. What would be a reasonable choice for $\alpha$, in this case? (Look at the MAP expression derived in part (b).) Plug in your choice, along with $\beta = \sigma/3$ and the Metro data given above, to obtain the MAP estimate for the waiting time at Hauz Khas. Comment on the difference between this and the maximum likelihood estimate obtained in part (a). **[3]**

3. Suppose we have a data set $\{(\mathbf{x}_1, t_1); (\mathbf{x}_2, t_2); ...; (\mathbf{x}_N, t_N)\}$, where the $\mathbf{x}_i \in \mathbb{R}^n$ are $n$-dimensional feature vectors, and the $t_i \in \{\mathcal{C}_1, \mathcal{C}_2\}$ are categorical class labels. Further suppose that we adopt the following model for the class priors and the class-conditional likelihoods:

$$\begin{aligned} p(\mathcal{C}_1) &= \theta, \\ p(\mathcal{C}_2) &= 1 - \theta, \\ p(\mathbf{x}|\mathcal{C}_k) &= \frac{1}{(2\pi)^{n/2}} \frac{1}{|\mathbf{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^{\mathrm{T}} \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right\}. \quad (k \in \{1, 2\}) \end{aligned}$$

Obtain an expression for the posterior distribution of the class label for a given data point $\mathbf{x}$. What kind of separation boundary does this specify between the classes? Under what conditions does this boundary become linear? **[5]**