

EEL709: Minor I

3rd February, 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: θ is the prior probability of a patient being healthy; p is the probability of a healthy prediction from FOB if the patient is healthy (also called the *specificity*); and q is the probability of a cancer prediction from FOB if the patient has cancer (also called the *sensitivity*).

- (a) Write down the joint likelihood of the data, as a function of the three model parameters θ , 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	K
FOB predicted cancer	1	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 specificity and sensitivity estimates obtained are $p' = 0.8$ and $q' = 0.96$. Obtain the expected loss for CTS as well; what is the critical value of K above 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 *Poisson distribution*, i.e.,

$$p(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \quad (k \in \{0, 1, 2, 3, \dots\})$$

- (a) It is known that the expected value of k under the Poisson distribution is λ ; i.e., in our case, λ corresponds to the expected time interval between successive arrivals. Given a

sequence of independent observations of time intervals, $\{k_1, k_2, \dots, k_N\}$, write down the likelihood function for this data under the Poisson distribution, and obtain a general expression for the maximum likelihood estimate for λ . [3]

(b) Now plug in the Metro data above to get the maximum likelihood estimate for the time interval at Hauz Khas station, based on the five observations. Do you think this is a good estimate for the actual expected waiting time between trains? Why or why not? [2]

(c) Suppose we now treat λ as a random variable and put a prior distribution on it. The *conjugate prior* for the Poisson distribution is given by the *Gamma distribution*, which is of the form:

$$p(\lambda) = \frac{1}{K} \lambda^{\alpha-1} e^{-\beta\lambda}. \quad (\lambda > 0)$$

Here α and β are parameters to be chosen, and K is a normalising constant which is independent of λ . Use this prior and the likelihood function from part (a) to obtain the posterior distribution of λ , as a function of the observed data, α , and β . Derive a general expression for the maximum a posteriori (MAP) estimate for λ . [3]

(d) The parameter β above is a *rate parameter*, which specifies the precision of the distribution; high β corresponds to less spread or variance. Suppose we fix $\beta = 10$. 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 α , in this case? (Look at the MAP expression derived in part (c).) Plug in your choice, along with $\beta = 10$ 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 (b). [3]

3. Suppose we have a data set $\{(\mathbf{x}_1, t_1); (\mathbf{x}_2, t_2); \dots; (\mathbf{x}_N, t_N)\}$, 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:

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

(Note that we have assumed different mean vectors for the two classes, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, but a common covariance matrix Σ .)

Show that the posterior distribution of the class label for a given data point \mathbf{x} can be written as a logistic function as follows:

$$p(\mathcal{C}_1|\mathbf{x}; \theta, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \frac{1}{1 + \exp(-(\mathbf{w}^T \mathbf{x} + w_0))},$$

where \mathbf{w} and w_0 are appropriate functions of θ , $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and Σ . [5]