Assignment (CSE6328 F12)

Due: in class on Oct 10, 2012

You have to work individually. Hand in a hardcopy of your answers before the deadline. No late submission will be accepted. No handwritting is accepted. Direct your queries to Hui Jiang (hj@cse.yorku.ca).

1. Assume we have a random vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ which follows a bivariate Gaussian distribution: $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$, where $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ is the mean vector and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma 2 \\ \rho\sigma_1\sigma 2 & \sigma_2^2 \end{pmatrix}$ is the covariance matrix. Derive the formula to compute mutual information between x_1 and x_2, i.e., $I(x_1, x_2)$.

   Hints: Refer to the related sections in the reading assignment [W2]. Note there is a mistake in [W2]: equation (99) should be

   $$\Gamma(n+1) = n \cdot \Gamma(n).$$

2. Assume we have two Gaussian distributions: $\mathcal{N}(x|\mu_1, \sigma_1^2)$ and $\mathcal{N}(x|\mu_2, \sigma_2^2)$, where μ_1 and μ_2 are their means, and σ_1^2 and σ_2^2 are their variances. Derive the formula to computer the K-L divergence between these two Gaussian distribution.

3. In many pattern classification problems, one has the option either to assign the pattern to one of N classes, or to *reject* it as being unrecognizable. If the cost for rejection is not too high, rejection may be a desirable action. If we observe feature $\mathbf{x}$ of a pattern (assume its true class id is ω_i), let's define the loss function for all actions α_j as:

   $$\lambda(\alpha_j|\omega_i) = \begin{cases} 0 & : \quad j = i \ \text{(correct classification)} \\ \lambda_s & : \quad j \neq i \ \text{ and } \ 1 \leq j \leq N \text{(wrong classification)} \\ \lambda_r & : \quad \text{rejection} \end{cases}$$

   where λ_s is the loss incurred for making any a wrong classification decision, and λ_r is the loss incurred for choosing the rejection action. Show the minimum risk is obtained by the following decision rule: we decide ω_i if $p(\omega_i|\mathbf{x}) \geq p(\omega_j|\mathbf{x})$ for all j and if $p(\omega_i|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$, and reject otherwise. What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?

   (Hint: consider the average loss for each action.)

4. Suppose we have three classes in two dimensions with the following underlying distributions:

   - class ω_1: $p(\mathbf{x}|\omega_1) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
   - class ω_2: $p(\mathbf{x}|\omega_2) \sim \mathcal{N}\left(\binom{1}{1}, \mathbf{I}\right)$
   - class ω_3: $p(\mathbf{x}|\omega_3) \sim \frac{1}{2}\mathcal{N}\left(\binom{0.5}{0.5}, \mathbf{I}\right) + \frac{1}{2}\mathcal{N}\left(\binom{-0.5}{0.5}, \mathbf{I}\right)$

   where $\mathcal{N}(\mu, \Sigma)$ denotes 2-d Gaussian distribution with mean vector μ and covariance matrix Σ, and $\mathbf{I}$ is identity matrix. Assume class prior probabilities $P(\omega_i) = 1/3, i = 1, 2, 3$.

   (a) By explicit calculation of posterior probabilities, classify the feature $\mathbf{x} = \binom{0.3}{0.3}$ based on the MAP decision rule.

   (b) Suppose that for a particular pattern the first feature is missing. Classify $\mathbf{x} = \binom{*}{0.3}$ for minimum probability of error.

5. you have collected a set of data samples $x_1, x_2, \cdots, x_n$. If we assume the data follow an exponential distribution as

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & : \quad x \leq 0 \\ 0 & : \quad \text{otherwise.} \end{cases}$$

   Derive the maximum-likelihood estimate for the parameter θ.

6. Assume we have c different classes, $\omega_1, \omega_2, \cdots, \omega_c$. Each class ω_i $(i = 1, 2, \cdots, c)$ is modeled by a univariate Gaussian distribution with mean μ_i and variance σ, i.e., $p(x \mid \omega_i) = \mathcal{N}(x \mid \mu_i, \sigma^2)$, where σ is a common variance for all c classes. Suppose we have collected n data samples from these c classes, i.e., $\{x_1, x_2, \cdots, x_n\}$, and let $\{l_1, l_2, \cdots, l_n\}$ be their labels so that $l_k = i$ means the data sample x_k comes from the i-th class, ω_i.

   Based on the given data set, derive the maximum-likelihood estimates for all model parameters, i.e., all means μ_i $(i = 1, 2, \cdots, c)$ and the common variance σ.