Intelligent Data Analysis

Problem Set 1 **Due 3/14/2016**

1. We have three urns, I, II and III. Urn I contains three red balls. Urn II contains two red balls and one white ball. Urn III contains one red ball and two white balls. You may use Bayes' rule to solve the following problems.
   (a) An urn is drawn at random and one ball is chosen at random from it. Suppose it is a red ball. What is the probability that the two balls left in the drawn urn are also red?
   (b) An urn is drawn at random and two balls are chosen at random from it. Suppose they are two red balls. What is the probability that the ball left in the drawn urn is also red?

2. In a two-class, two-action problem (rejection is not allowed), if the loss function is $\lambda_{11} = \lambda_{22} = 0,\ \lambda_{12} = k,$ and $\lambda_{21} = 1 - k,$ show the optimal classification rule in terms of posterior probability so as to minimize the *risk*.

3. Suppose we are given a sample $\chi = \left\{x^t\right\}_{t=1}^{N},$ where each x^t is chosen randomly from the normal distribution

$$p\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Show that the maximum likelihood estimate (MLE) of μ is

$$\hat{\mu} = \frac{1}{N}\sum_{t=1}^{N} x^t,$$

and the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N}\left(x^t - \hat{\mu}\right)^2.$$

4. If a series of independent Bernoulli trials is terminated with the occurrence of the first success, the probability function for *X*, the length of the series, will be given by the geometric distribution $P(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, \ldots,$ where *p* is the unknown probability of success at any trial. If *N* such series are observed, the data will consist of *N* lengths, $\chi = \left\{k^1, k^2, \ldots, k^N\right\}.$ Find the MLE of *p*.

5. Prove the following statements.

   (a) Let X be a random variable and let a and b be constants. Define $Y = aX + b$. Then $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$.

   (b) Let $X_1, \ldots, X_N$ be independent random variables and let $Y = X_1 + \cdots + X_N$. Then $\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_N)$.

   (c) Let $\bar{X} = \frac{1}{N}(X_1 + \cdots + X_N)$, where X_i's are chosen randomly from a population whose mean is μ and variance is σ^2. Then $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{N}$.

6. Program Assignment This goal of this assignment is study univariate parametric classification using Fisher's Iris data. You can download the iris data set from e3. The iris data set contains measurements of 50 specimens from each of three different species of iris—setosa, versicolor, and virginica—on the following dimensions: sepal length, sepal width, petal length, petal width.

   (a) Plot the matrix of scatter plots by group (species), as shown in Chapter 3, pp 3.

   (b) For each species, show the histogram with normal probability fit and the normal probability plot. See Chapter 1, pp 16.

   (c) For each input (feature), i.e., *sepal length, sepal width, petal length, petal width*, use the training data set to devise three univariate classifiers:

   (1) Quadratic distriminant analysis (QDA)

   (2) Linear discriminant analysis (LDA)

   (3) Nearest mean classifier (NMC)

   So you will have a total of 12 classifiers. For each classifier, show the estimated parameter(s).

   Apply the 150 testing cases to each classifier, report the confusion matrix and then compute the accuracy. The following is an example of confusion matrix for 10 testing cases.

Predicted\Observed	Setosa	Versicolor	Virginica	Total
Setosa	1	1	0	2
Versicolor	0	3	1	4
Virginica	2	0	2	4
Total	3	4	3	10

   There are 1+3+2=6 correctly classified cases, so the accuracy is 6/10×100%

= 60%. Answer the following questions:

- Rank the four features based on the average test accuracy of the three univariate classifiers.
- For every feature, will QDA always outperform LDA and NMC? Will LDA outperform NMC? Write any comments or findings you get.