Helwan University
Faculty of Computers and Information

MidTerm Exam
Fall 2007

Department of Computer Science
Machine Learning
Fall 2007

The grade is out of 10, while there are 18 points on the problems.

3 marks

1. In Machine Learning (ML) applications, the set of features constitute a random vector X that has a mean vector μ_X and a covariance matrix Σ_X. Prove that the random variable $Y = X - c$, where c is any constant vector, has a mean $\mu_Y = \mu_X - c$ and a covariance matrix $\Sigma_Y = \Sigma_X$.

7 marks

2. In some medical applications in ML, it has been noticed that some features constitute a random vector X with a mean vector μ_X and a covariance matrix Σ_X, where

$$\mu_X = (1, 2)', \text{and}$$
$$\Sigma_X = \begin{pmatrix} 1 & .2 \\ .2 & .5 \end{pmatrix}$$

   (a) What is the dimensionality of X.

   (b) What is the correlation coefficient between the two features of X. What is the mean and variance of every component?

   (c) Assume that very large number of i.i.d. observations, from X, are collected in an experiment. How does their data plot most likely look like?

   (d) Assume that we had to transform each observation x_i by the transformation A, where

$$A = \begin{pmatrix} 50/46 & -20/46 \\ -20/46 & 100/46 \end{pmatrix};$$

   each transformed observation is called y_i that comes from the random vector $Y = AX$. How does the transformed data set look like?

   (e) Can we calculate the distribution of Y? If yes, how and what is it?; if no, why?

   (f) Can we say anything about the "statistical dependence" between the two components of X? Give reason!

   (g) Can we say anything about the "statistical dependence" between the two components of Y? Give reason!

5 marks

3. The PH of water samples from a specific lake is a random variable Y with probability density function given by

$$f_Y(y) = \begin{cases} c(7-y)^2, & 5 \le y \le 7 \\ 0, & \text{elsewhere.} \end{cases}$$

   (a) Find c.

   (b) Sketch the PDF.

   (c) Find $\mathrm{E}\, Y$ and $\mathrm{Var}\, Y$.

   (d) Find an interval that starts at 5 and is shorter than $(5, 7)$ in which at least $3/4$ of the PH measurements must lie.

   (e) Would you expect to see a PH measurement below 5.5 very often? Why?

3 marks

4. A random variable X is distributed as $U(0, \theta)$, where θ is a constant. Sketch the PDF, then find the mean and variance of X.