COMP4702/COMP7703 - Machine Learning
Homework 3 - Multivariate Parametric Models and Density Estimation

Marcus Gallagher

Core Questions

1. Consider the nearest mean classifier (see Alpyadin p.104). Given a dataset representing a K-class classification problem:
$$\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_{t=1}^N$$
How many distance values would need to be calculated to classify $\mathcal{X}$ using this method?

2. Given the following sample covariance matrix:
$$C = \begin{pmatrix} 3.8600 & -0.0200 & 0.0100 \\ -0.0200 & 0.0100 & -0.0200 \\ 0.0100 & -0.0200 & 0.0700 \end{pmatrix}$$
Calculate the sample correlation matrix, R.

3. In the course Datasets folder, you will find two datasets that are a subset of the (Wisconsin) breast cancer dataset from the UCI Machine Learning Repository. These datasets have 9 inputs/features and class labels in the final column. Apply quadratic discriminant analysis (see Alpaydin, Section 5.5) to the breast cancer (training) dataset, and provide the following (as a percentage value to two decimal places):

   (a) Training error

   (b) Validation error

   Note: you do not need to work directly with the discriminant function ($g_i(\mathbf{x})$) to do this. You proceed by estimating class densities and using Bayes rule.

Extension Questions

4. Read Section 5.8 of Alpaydin and then exercise 7 in 5.10. Using the approach described, fit a 2-D quadratic regression model to the dataset `reg2d.csv`, which has the inputs as the first two columns and the target function values in the third column. What are the coefficient values for your linear model?

5. Imagine you have a 7-class classification problem, where the dataset contains 9 input features. You decide to build a classifier using a "mixture of mixtures", i.e. using a Gaussian mixture model for each likelihood ($p(\mathbf{x}|\theta)$). 3 mixture components are used with diagonal covariance matrices for each mixture model. Calculate the total number of model parameters in the classifier (do not consider priors).