COMP4702/COMP7703/DATA7703 - 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?

   Once the mean point of each class has been calculated (no distance calculations), we need to compute the distance between each mean and each data point: $K \times N = KN$.

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.

```
%HW3 Q1 Solution

S = [3.86 -0.02 0.01; -0.02 0.01 -0.02; 0.01 -0.02 0.07];
%Diagonal elements of R are equal to 1 by definition
R(1,1) = 1;
R(2,2) = 1;
R(3,3) = 1;
%Calc r_ij using formula
R(1,2) = S(1,2)/(sqrt(S(1,1))*sqrt(S(2,2)));
R(1,3) = S(1,3)/(sqrt(S(1,1))*sqrt(S(3,3)));
R(2,3) = S(2,3)/(sqrt(S(2,2))*sqrt(S(3,3)));
%R matrix is symmetric
R(2,1) = R(1,2);
R(3,1) = R(1,3);
R(3,2) = R(2,3);
%Done!
R

>> hw3q1
```

```
R =

    1.0000   -0.1018    0.0192
   -0.1018    1.0000   -0.7559
    0.0192   -0.7559    1.0000
```

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 to the breast cancer (training) dataset, and provide the following (as a percentage value to two decimal places):

   (a) Training error - 3.17%

   (b) Test error - 6.85%

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`. What are the coefficient values for your linear model?

```
% HW3 Q4
% 2D quadratic regression by transformation to linear model

%Create some data
%x1 = rand(100,1);
%x2 = rand(100,1);
%True function is actually quadratic plus noise
%r = (x1.^2 + x2.^2) + 0.1*randn(100,1);
%OR, use the one I generated for this question
load('reg2d.mat');
x1 = reg2d(:,1);
x2 = reg2d(:,2);
r = reg2d(:,3);

%Now to do regression as per Alpaydin's Sec 5.10, ex7
z1 = x1;
z2 = x2;
z3 = x1.*x2;
z4 = x1.*x1;
z5 = x2.*x2;
Z = [z1 z2 z3 z4 z5];
%Solution 1 (the lazy way!) - use the Matlab routine
%mdl = fitlm(Z,r)
%OR
%Solution 2 (hardcore!) - implement the equation!
```

```
%We need to make up the offset/w0 "data"
z0 = ones(100,1);
Z = [z0 Z];
w = inv(Z'*Z)*Z'*r
```

When this is run:

```
>> hw3q4

w =

    0.0604
   -0.2570
    0.0513
    0.1381
    1.1423
    0.8996

>>
```

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 (include mixture coefficients but do not consider the overall class priors in Bayes Rule).

   There are 7 mixture models. Each one has three components. Each component has 3 mixing coefficients, plus a 9-D mean vector and 9 variance terms. $7 \times (3 + 3 \times (9 + 9)) = 399$.

   An argument could be made that only 2 mixing coefficients are needed because they sum to one, so this is also acceptable. $7 \times (2 + 3 \times (9 + 9)) = 392$.