COMS21202 Symbols, Patterns and Signals

Problem sheet: Classification and Clustering

1. (**Naive Bayes, revision of COMS10003**) Suppose a naive Bayesian spam filter uses a vocabulary consisting of the words 'Viagra', 'CONFIDENTIAL', 'COMS21202' and 'Gaussian', and has estimated the class-conditional likelihoods of these words occurring in spam and non-spam emails as in Table 1.

Table 1: Class-conditional likelihoods for words in the vocabulary.

word	$P(word\|spam)$	$P(word\|\neg spam)$
Viagra	0.20	0.01
CONFIDENTIAL	0.30	0.05
COMS21202	0.02	0.20
Gaussian	0.05	0.10

Consider three test emails as follows: A contains the word 'Viagra' but none of the others; B contains the word 'CONFIDENTIAL' but none of the others; C contains the words 'COMS21202' and 'Gaussian' but none of the others.

   (a) Determine the most likely class of each of these emails by calculating the likelihood ratios $\frac{P(email|spam)}{P(email|\neg spam)}$.

   (b) Now assume that typically 10% of your emails are spam. Using MAP estimation, investigate how this affects your predictions.

2. (**Mahalanobis distance**) Given a covariance matrix Σ, the Mahalanobis distance between vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathrm{Dis}_M(\mathbf{x},\mathbf{y}|\Sigma) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mathbf{y})}$$

   (a) Show that in the 2-D case, points with the same Mahalanobis distance D to a fixed $\mathbf{y} = \mu = (\mu_1\ \mu_2)^{\mathrm{T}}$ describe an ellipse. Use $\Sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with inverse $\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.
   (Hint: the general form of an ellipse is $Ax_1^2 + Bx_1x_2 + Cx_2^2 + Dx_1 + Ex_2 + F = 0$, but in our case it is easier to use $(x_1 - \mu_1)$ and $(x_2 - \mu_2)$ as variables.)

   (b) Derive the ellipse equation for $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

   (c) Derive the ellipse equation for $\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$.

   (d) Derive the ellipse equation for $\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$.

3. (**Maximum-likelihood decision boundaries using Mahalanobis distance**) The multivariate normal distribution can be expressed in terms of Mahalanobis distance as follows:

$$P(\mathbf{x}|\mu,\Sigma) = \frac{1}{E_d}\exp\left(-\frac{1}{2}\left(\mathrm{Dis}_M(\mathbf{x},\mu|\Sigma)\right)^2\right), \quad E_d = (2\pi)^{d/2}\sqrt{|\Sigma|}$$

Suppose we have two sets of bivariate normally distributed data points whose covariance matrices have the same determinant, then $P(\mathbf{x}|\mu_1,\Sigma_1) = P(\mathbf{x}|\mu_2,\Sigma_2)$ if and only if $\mathrm{Dis}_M(\mathbf{x},\mu_1|\Sigma_1) = \mathrm{Dis}_M(\mathbf{x},\mu_2|\Sigma_2)$. In other words, the decision boundary is formed by the set of points that have the same Mahalanobis distance to both means.

Use the results of the previous question to derive equations for the decision boundary in each of the cases below, using means $\mu_1 = (1,1)$ and $\mu_2 = (-1,-1)$.

(Sanity check: the first two should give a straight line through the mid-point between the two means, which is $(0,0)$.)

(a) $\Sigma_1 = \Sigma_2 = \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ – i.e., the real-valued equivalent of the naive Bayes assumption.

(b) $\Sigma_1 = \Sigma_2 = \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$.

(c) $\Sigma_1 = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$.

4. (**Decision trees**) Suppose we have a training set of 32 spam emails and 32 non-spam emails, and the numbers of emails containing particular words are as in Table 2. We want to build a decision tree using these words as boolean features: if the word occurs in an email the feature is true, else it is false. Which feature results in the best split, as measured by information gain?

Table 2: Numbers of spam and non-spam emails containing particular words.

word	spam	non-spam
Viagra	15	1
CONFIDENTIAL	28	4
COMS21202	1	15
Gaussian	4	12

5. (**Distance metrics**) Calculate the pairwise distances between the following points using the L_k norm with $k = 1$, $k = 2$ and $k = \infty$:

(a) $x = 1$, $y = 2$ and $z = 4$.

(b) $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$.

(c) $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$.

6. (**Nearest-neighbour classification**) Assume $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ are three training instances labelled $+$, $+$ and $-$, respectively. Derive the k-nearest neighbour prediction for the test points $\mathbf{p} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\mathbf{q} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$, using Euclidean distance, for $k = 1$ and $k = 3$.

7. (K-**means clustering**) Assume $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ form a cluster. Using Euclidean distance:

(a) calculate the sum of the squared distances of cluster points to the mean (within-cluster scatter); and

(b) calculate the pairwise squared distance between cluster points, summed over all pairs.

8. (***K*-means clustering**) Let $\mathbf{X} = \begin{bmatrix} 1 & 2 & 2 & 2 & 3 \\ 2 & 3 & 4 & 1 & 2 \end{bmatrix}$ be a data matrix, with data points in columns. Calculate new centroids $\mu_i(1)$ after one iteration of K-means for each of the following initial centroids:

   (a) $\mu_1(0) = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$;

   (b) $\mu_1(0) = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$;

   (c) $\mu_1(0) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$.

   Also indicate in each of these cases whether K-means has converged. Which of these sets of centroids are optimal?

9. (**Gaussian mixtures**) Using the same data as in the previous question, use Gaussian mixtures and one iteration of Expectation-Maximisation to calculate new centroids from $\mu_1(0) = \begin{bmatrix} 2 \\ 2.15 \end{bmatrix}$ and $\mu_2(0) = \begin{bmatrix} 2 \\ 2.65 \end{bmatrix}$.