SD 372 Intro. to Pattern Recognition April 12, 1997

Final Examination

Professor Paul Fieguth

Aids Permitted:	*Two* 8.5x11 (inch!) pages, *No* calculator. (Abacus / fingers are OK.)
Advice:	Read problems carefully before jumping in to calculations. *** Well-drawn sketches / diagrams can be very helpful. *** The grade value for each question is indicated in brackets [ ] next to the question number. I *will* give part marks for relevant statements or insights. Tell me what you know! There are only a few parts of questions which involve nontrivial numerical manipulation: 1(d), 3(c)-iv, 3(c)-v Don't get too distracted by these questions until you have looked at the rest of the exam.

[25%] 4. Feature Selection and Extraction:

a) Define feature selection and feature extraction. State clearly the differences between the two.

b) In what sorts of circumstances might you prefer to use feature extraction over feature selection or vice-versa.

c) Both feature selection and extraction reduce the number of features.

When is having fewer features a good thing?

Under what circumstances might you want more (i.e., additional) features?

d) In lectures we discussed several methods for keeping a good single feature for the two-cluster case:

1. Keep the direction connecting the two cluster means.
2. Keep the minimum intra-class direction.
3. Keep the maximum inter-class direction (i.e., Fisher's discriminant).

For each of the above three methods, do the following:

i) State the underlieing principle behind each method (i.e., what is the approach trying to do?). Write down (**don't** derive) the mathematical solution / definition for the feature direction.

ii) Sketch an example of where each method works well; i.e., sketch two cluster-shapes and the feature direction.

iii) Sketch an example of where each method fails for linearly separable clusters.

[2%] Bonus Question

Do **NOT** waste your time here unless you are happy with your answers to the rest of the exam!!!

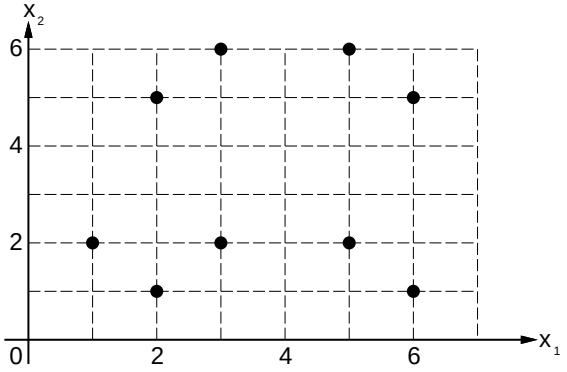
Here's a little probability question which I was asked a few years ago by a friend in graduate school:

Suppose you take a sphere (ball) and draw three points on it at (independent) random locations.

What is the probability that all three points lie in one hemisphere (half a sphere)?

[15%] 2. Unlabeled Clustering:

GIVEN: The nine data points plotted in the figure ...



Recall the definition of cluster distances for the nearest-neighbour and furthest-neighbour hierarchical clustering methods:

$$\text{NN:} \quad d_{\text{NN}}(C_1, C_2) = \min_{\underline{x}_1 \in C_1, \underline{x}_2 \in C_2} d(\underline{x}_1, \underline{x}_2)$$

$$\text{FN:} \quad d_{\text{FN}}(C_1, C_2) = \max_{\underline{x}_1 \in C_1, \underline{x}_2 \in C_2} d(\underline{x}_1, \underline{x}_2)$$

where $d(\underline{x}_1, \underline{x}_2)$ is the usual Euclidean distance between points $\underline{x}_1$ and $\underline{x}_2$.

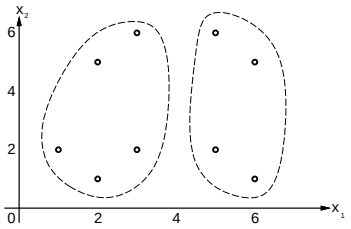
a) Suppose that I tell you that there are $K = 2$ clusters.

   i) Find (i.e., sketch/plot) the two NN clusters from the nine data points.

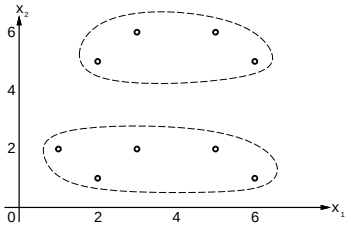
   ii) Find (i.e., sketch/plot) the two FN clusters from the nine data points.

b) We don't really *have* to use a Euclidean distance definition for $d(\underline{x}_1, \underline{x}_2)$. Let's investigate how a non-Euclidean distance might affect clustering behaviour:

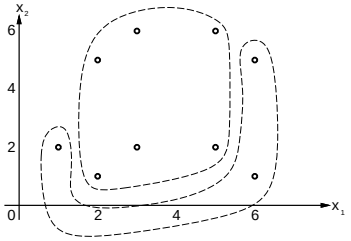
   i) Suggest a function $d(\underline{x}_1, \underline{x}_2)$ such that NN clusters the nine data points into two clusters like this:



   ii) Suggest a function $d(\underline{x}_1, \underline{x}_2)$ such that FN clusters the nine data points into two clusters like this:



   iii) Suggest a function $d(\underline{x}_1, \underline{x}_2)$ such that NN actually clusters the nine data points into two unusual clusters like this:



[30%] 3. Linear Discriminants:

GIVEN: Two clusters C_1, C_2 with four data points each:

$$C_1 = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}$$

$$C_2 = \left\{ \begin{bmatrix} -1+\alpha \\ \alpha \end{bmatrix} \quad \begin{bmatrix} -2+\alpha \\ \alpha \end{bmatrix} \quad \begin{bmatrix} -1+\alpha \\ 1+\alpha \end{bmatrix} \quad \begin{bmatrix} -1+\alpha \\ -1+\alpha \end{bmatrix} \right\}$$

where α is a scalar.

a) For what values of α will the Perceptron algorithm produce a solution?

b) For what values of α will the MSE algorithm produce a solution?

c) For the remainder of the problem, use $\alpha = 2$.

Recall: a linear discriminant $\underline{w}^T \underline{x} + w_o$ is a hyperplane which separates clusters C_1 and C_2. Let's define the *quality* Q of a discriminant to be the smallest distance from the hyperplane (line) to any point in C_1 or C_2; i.e.,

$$Q(\underline{w}, w_o) = \min_{y \in C_1 \cup C_2} (\text{distance from the point } y \text{ to the line } \underline{w}^T \underline{x} + w_o = 0)$$

If the line $\underline{w}^T \underline{x} + w_o = 0$ fails to separate C_1 and C_2 then we'll define $Q(\underline{w}, w_o) = 0$. A larger Q is clearly better than a small one.

i) What is the "best" linear discriminant (i.e., what are the $\underline{w}, w_o$ which maximize $Q()$)? No proof is required, just draw a sketch and write down the best $\underline{w}, w_o$.

ii) Let the discriminant normal vector $\underline{w}$ be given by the difference of the cluster sample means. What is $\underline{w}$? Sketch the best discriminant having this normal vector.

iii) Compute $\underline{w}$ using Fisher's discriminant method. What is $\underline{w}$? Sketch the best discriminant having this normal vector.

iv) Let $\underline{w}$ be the principal component of the total sample; that is, let $\underline{w}$ be the eigenvector corresponding to the largest eigenvalue of the total sample covariance. Find $\underline{w}$. Sketch the best discriminant having this normal.

v) Apply the Perceptron to the data (I recommend you use the sequential algorithm – it is the easiest; as usual, start with $\underline{a}_o = [0\ 0\ 0]^T$ and use a multiplier $\rho_k = 1$). What is the resulting discriminant? Sketch it.

Which of the methods in (ii) – (v) came the closest to your best discriminant in (i)? (no math / proof needed, just state your answer).

[30%] 1. MICD / MAP ...

GIVEN: We have three clusters, C_1, C_2, C_3. We have only one feature dimension (i.e., x is a scalar); the statistics $p(x|C_i)$ for each cluster are Gaussian, where

$$\begin{array}{ccc} \mu_1 = -2 & \mu_2 = 0 & \mu_3 = +2 \\ \sigma_1^2 = 1 & \sigma_2^2 = 1 & \sigma_3^2 = 1 \end{array}$$

The prior probabilities of the three clusters are

$$P(C_1) = 0.5, \quad P(C_2) = 0.2, \quad P(C_3) = 0.3$$

You may find the following values useful:

$$\ln\,(0.2) = -1.6, \quad \ln\,(0.3) = -1.2, \quad \ln\,(0.5) = -0.7$$

a) Find the MICD classifier for this problem.
What is the probability of error $P_{\text{MICD}}(\epsilon)$ for this classifier?

b) Find the MAP classifier for this problem.
What is the probability of error $P_{\text{MAP}}(\epsilon)$?

c) Which is greater, $P_{\text{MICD}}(\epsilon)$ or $P_{\text{MAP}}(\epsilon)$?

d) Suppose we collect samples $x_1, x_2, \ldots$ from the clusters. We'll label the data (i.e., we want to keep track of which cluster x_i came from), but our labeler makes an error: both clusters #2 and #3 will be reported as cluster #2.

So now we'll have only two clusters: C_1 and C_{23}.

The statistics for cluster C_1 will be unchanged from before, but what are μ_{23} and σ_{23}^2, the mean and variance of cluster C_{23}? (you can leave your answer in equation form)

e) Regardless of your answer from part (d), let's say $\mu_{23} = 1.2, \sigma_{23}^2 = 2$.

How many MICD boundaries are there now? Sketch (don't calculate) the approximate positions of the boundaries.

Is the MICD boundary between clusters C_1 and C_2 in (a) less than, equal to, or greater than the MICD boundary between C_1 and C_{23}?