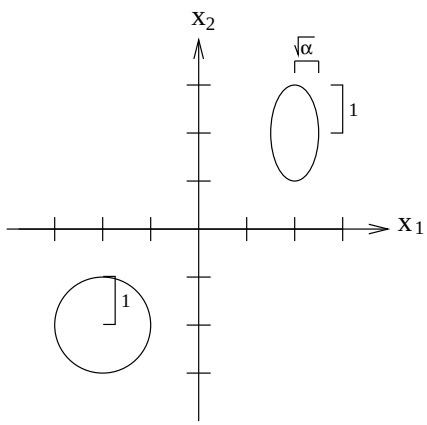
T-61.231 Principles of Pattern Recognition

Answers to exercise 4: 22.10.2001

1. A Gaussian distribution has two parameters: a mean vector $\overline{\mu}$ and covariance matrix Σ.
   $\hat{\overline{\mu_i}} = \overline{m_i}$ and $\hat{\Sigma} = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(\overline{x_k} - \hat{\overline{\mu_i}})(\overline{x_k} - \hat{\overline{\mu_i}})^T$

   a) Lets assume that there are approximately as many samples from both classes ($n_1 \approx n_2$). Now we can draw both distributions in the same picture and use the same scale for both of them.

   The density of class 1 is symmetric as the diagonal elements of S_1 are equal. As for class 2, the density is expanded in the direction of the "width" of the distribution on the x_1-axis depends on α.



   b) $\hat{\overline{w}} = S_W^{-1}(\overline{m_1} - \overline{m_2})$

$$S_W = S_1 + S_2 = \begin{bmatrix} 1+\alpha & 0 \\ 0 & 2 \end{bmatrix} \Rightarrow S_W^{-1} = \frac{1}{2(1+\alpha)}\begin{bmatrix} 2 & 0 \\ 0 & 1+\alpha \end{bmatrix} = \begin{bmatrix} \frac{1}{1+\alpha} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

$$\hat{\overline{w}} = \begin{bmatrix} \frac{1}{1+\alpha} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}\begin{bmatrix} (-2-2) \\ (-2-2) \end{bmatrix} = -2\begin{bmatrix} \frac{2}{1+\alpha} \\ 1 \end{bmatrix}$$

   c) To determine the eigenvectors of $S_W^{-1}S_B$ we first calculate the between-class scatter matrix S_B from $S_B = (\overline{m_1} - \overline{m_2})(\overline{m_1} - \overline{m_2})^T$.

$$S_B = \begin{bmatrix} (-2-2) \\ (-2-2) \end{bmatrix}\begin{bmatrix} (-2-2) & (-2-2) \end{bmatrix} = \begin{bmatrix} 16 & 16 \\ 16 & 16 \end{bmatrix}$$

   Now

$$S_W^{-1}S_B = \begin{bmatrix} \frac{1}{1+\alpha} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}\begin{bmatrix} 16 & 16 \\ 16 & 16 \end{bmatrix} = \begin{bmatrix} \frac{16}{1+\alpha} & \frac{16}{1+\alpha} \\ 8 & 8 \end{bmatrix}$$

   The eigenvectors for matrix A are defined as $A\overline{x} = \lambda\overline{x} \Rightarrow (A - \lambda I)\overline{x} = \overline{0}$. A nontrivial solution exists, if $det(A - \lambda I) = 0$

$$\begin{vmatrix} \frac{16}{1+\alpha} - \lambda & \frac{16}{1+\alpha} \\ 8 & 8-\lambda \end{vmatrix} = 0 \Leftrightarrow (\frac{16}{1+\alpha} - \lambda)(8-\lambda) - 8\frac{16}{1+\alpha} = 0 \Leftrightarrow \lambda^2 - (8 + \frac{16}{1+\alpha})\lambda = 0$$

Thus the eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = 8 + \frac{16}{1+\alpha}$ (Note that $\alpha = \sigma_1^2 > 0$).
The eigenvector corresponding to the larger eigenvalue: $A\overline{e} = \lambda_2 \overline{e}$, where $\overline{e} = [e_1\ e_2]^T$. Thus

$$\begin{bmatrix} \frac{16}{1+\alpha} & \frac{16}{1+\alpha} \\ 8 & 8 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = (8 + \frac{16}{1+\alpha}) \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

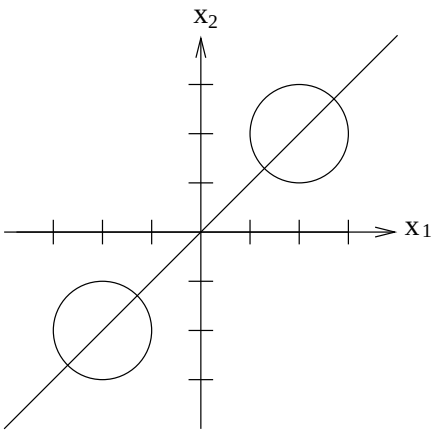
Calculating from the lower row (the same result can be obtained from the upper row, too)

$$8e_1 + 8e_2 = (8 + \frac{16}{1+\alpha})e_2 \Leftrightarrow e_1 = \frac{2}{1+\alpha}e_2 \Rightarrow e = \begin{bmatrix} \frac{2}{1+\alpha} \\ 1 \end{bmatrix}$$

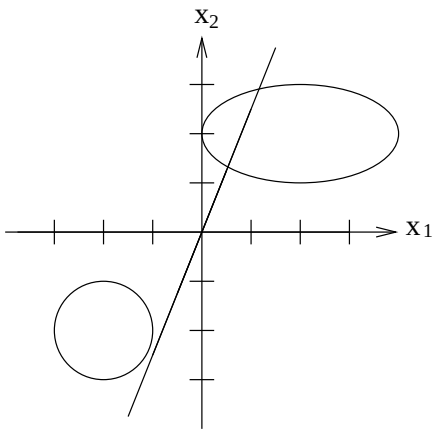
As can be seen, both methods produced a vector of the same orientation, as was to be expected.

d) The Fisher linear discriminant is thus $\hat{\overline{w}} = \begin{bmatrix} \frac{2}{1+\alpha} \\ 1 \end{bmatrix}$, $\underline{m}_1 = (\ -2 \quad -2\ )^T$, $\underline{m}_2 = (\ 2 \quad 2\ )^T$, $\mathbf{S}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{S}_2 = \begin{bmatrix} \alpha & 0 \\ 0 & 1 \end{bmatrix}$.

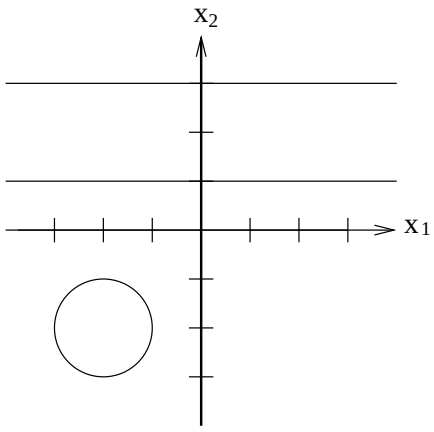
$\alpha = 1 \Rightarrow \hat{\overline{w}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $S_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$:



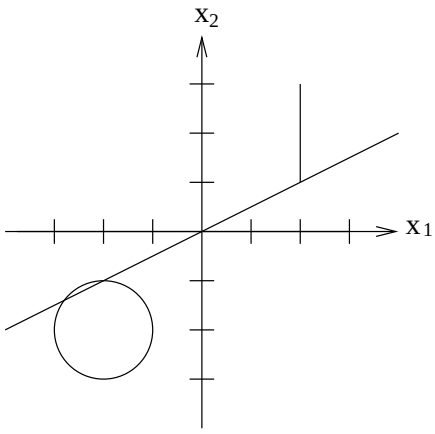
$\alpha = 4 \Rightarrow \hat{\overline{w}} = \begin{bmatrix} 2/5 \\ 1 \end{bmatrix}$ and $S_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$:



$\alpha \to \infty \Rightarrow \hat{\overline{w}} \to \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $S_2 \to \begin{bmatrix} \infty & 0 \\ 0 & 1 \end{bmatrix}$:



$\alpha \to 0 \Rightarrow \hat{\overline{w}} \to \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $S_2 \to \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$:



The discriminants seem quite valid, as it can be seen that by projecting the distributions onto the discriminant in all cases the distributions become well separable.

2. Let's denote the function slightly differently for the proof; let $l = d - 1 \Leftrightarrow d = l - 1$ and $O(N, l) = C(N, d)$, where

$$C(N, d) = 2\sum_{k=0}^{d-1} \binom{N-1}{k}$$

In my opinion this is more illustrative, as we are actually using the dimension of the space d instead of constantly using l to denote $l - 1$ dimensional space. $C(N, d)$ tells us the number of groupings that can be formed by d-dimensional hyperplanes to separate the N points into two classes.

First we need to prove that $C(N + 1, d) = C(N, d) + C(N, d - 1)$.

Let $C(N, d)$ be a separable set of dichotomies X. Let's take a new point x_{N+1} so that $X \cup \{x_{N+1}\}$ is in the general position (well distributed). Let there be a vector w that divides X into two sets $X = \{X^+, X^-\}$ so that $w \cdot x > t \Rightarrow x \in X^+$ and $w \cdot x < t \Rightarrow x \in X^-$, where t is a scalar.

If $\{X^+, X^-\}$ is separable, must also either $\{X^+ \cup \{x_{N+1}\}, X^-\}$ or $\{X^+, X^- \cup \{x_{N+1}\}\}$ be separable. However, they both are separable if and only if $\exists w$ that is a vector that separates $\{X^+, X^-\}$ in a $(d-1)$ dimensional space and is orthogonal to x_{N+1}

To prove the prior statement regarding w, let the set of separating vectors $W = \{w : w \cdot x > t, x \in X^+; w \cdot x < t, x \in X^-\}$. The set $\{X^+ \cup \{x_{N+1}\}, X^-\}$ is homogeneously separable if and only if $\exists w \in W$ so that $w \cdot x_{N+1} > t$, and equivalently $\{X^+, X^- \cup \{x_{N+1}\}\}$ is homogeneously separable if and only if $\exists w \in W$ so that $w \cdot x_{N+1} < t$ Let the sets be linearly separable with w_1 and w_2, respectively. Then $w^* = (-w_2 \cdot x_{N+1})w_1 + (w_1 \cdot x_{N+1})w_2$ separates $\{X^+, X^-\}$ by the hyperplane $\{x : w^* \cdot x = t\}$ passing through x_{N+1}. Conversely, if the sets $\{X^+, X^-\}$ are homogeneously linearly separable by a hyperplane containing x_{N+1}, then $\exists w^* \in W$ so that $w^* \cdot x = t$. Since W is an open set, $\exists \epsilon > 0$ so that $w * +\epsilon x_{N+1}$ and $w * -\epsilon x_{N+1}$ are in W. Hence $\{X^+ \cup \{x_{N+1}\}, X^-\}$ and $\{X^+, X^- \cup \{x_{N+1}\}\}$ are homogeneously linearly separable by $w * +\epsilon x_{N+1}$ and $w * -\epsilon x_{N+1}$, respectively.

So the set can be separated if and only if $\exists w$ so that the projection onto a $(d-1)$ dimensional subspace is separable. By the induction hypothesis there are $C(N, d - 1)$ such separable dichotomies. Hence,

$$C(N + 1, d) = C(N, d) + C(N, d - 1)$$

By repeatedly applying of this to the terms on the right we obtain

$$C(N, d) = \sum_{k=0}^{N-1} \binom{N-1}{k} C(1, d - k)$$

Now, as it is obvious that one point can be separated in two ways if the dimension is greater or equal to 1 and no separation can be made when the dimension is below one, or

$$C(1, m) = \begin{cases} 2, m \geq 1 \\ 0, m < 1 \end{cases}$$

The original theorem follows by separating the part of the sum where $d - k < 1 \Leftrightarrow k > d - 1$:

$$C(N, d) = 2\sum_{k=0}^{d-1} \binom{N-1}{k} + 0 \cdot \sum_{k=d}^{N-1} \binom{N-1}{k} = 2\sum_{k=0}^{d-1} \binom{N-1}{k} \Leftrightarrow O(n, l) = 2\sum_{k=0}^{l} \binom{N-1}{k}$$

3. The SVM optimal hyperplane separates the space so that

$$\omega^T x_i + \omega_0 \geq +1, \text{ if } x_i \in \omega_1$$
$$\omega^T x_i + \omega_0 < -1, \text{ if } x_i \in \omega_2$$

Let ω_1 be on the positive side of the optimal hyperplane and ω_2 on the negative side, and d_+ and d_- be the distances from the optimal hyperplane and the nearest point in classes ω_1 and ω_2, respectively. Let $g(x) = \omega^T x_i + \omega_0$ be the distance from the optimal hyperplane ω. It can also be stated that

$$x = x_p + r\frac{\omega}{||\omega||}$$

where x_p is the projection of x onto the optimal hyperplane and r is the distance from the hyperplane. Since $g(x_p) = 0$ by definition (the point x_p lies on the optimal hyperplane),

$$g(x) = \omega^T x + \omega_0 = r||w|| \Leftrightarrow r = \frac{g(x)}{||\omega||}$$

Thus the algebraic distance for the support vectors is

$$r = \frac{g(x)}{||\omega||} = \begin{cases} \frac{1}{||\omega||} = d_+, \text{ when } x \text{ is the nearest point of } \omega_1 \\ -\frac{1}{||\omega||} = d_-, \text{ when } x \text{ is the nearest point of } \omega_2 \end{cases}$$

Here the negative sign denotes being on the negative side of the hyperplane. Thus the distance between the two points is $\frac{2}{||\omega||}$.