$\theta_1, \dots, \theta_c$ and $c-1$ them the unknown $P(\omega_j)$ reduced by the single constraint $\sum\limits_{j=1}^{c} P(\omega_j) = 1$. Thus, the problem is not identifiable if $2c - 1 > m$.

2. PROBLEM NOT YET SOLVED

2. PROBLEM NOT YET SOLVED

Section 10.3

4. We are given that $\mathbf{x}$ is a binary vector and that $P(\mathbf{x}|\boldsymbol{\theta})$ is a mixture of c multivariate Bernoulli distributions:

$$P(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{c} P(\mathbf{x}|\omega_i, \boldsymbol{\theta}) P(\omega_i),$$

where

$$P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \prod_{j=1}^{d} \theta_{ij}^{x_{ij}} (1-\theta_{ij})^{1-x_{ij}}.$$

(a) We consider the log-likelihood

$$\ln P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \sum_{j=1}^{d} \left[x_{ij} \ln \theta_{ij} + (1-x_{ij}) \ln (1-\theta_{ij}) \right],$$

and take the derivative

$$\begin{aligned} \frac{\partial \ln P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)}{\partial \theta_{ij}} &= \frac{x_{ij}}{\theta_{ij}} - \frac{1-x_{ij}}{1-\theta_{ij}} \\ &= \frac{x_{ij}(1-\theta_{ij}) - \theta_{ij}(1-x_{ij})}{\theta_{ij}(1-\theta_{ij})} \\ &= \frac{x_{ij} - x_{ij}\theta_{ij} - \theta_{ij} + \theta_{ij}x_{ij}}{\theta_{ij}(1-\theta_{ij})} \\ &= \frac{x_{ij} - \theta_{ij}}{\theta_{ij}(1-\theta_{ij})}. \end{aligned}$$

We set this to zero, which can be expressed in a more compact form as

$$\sum_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i) \frac{x_k - \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)} = 0.$$

(b) Equation 7 in the text shows that the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_i$ must satisfy

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}_i} \ln P(x_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0.$$

We can write the equation from part (a) in component form as

$$\nabla_{\boldsymbol{\theta}_i} \ln P(x_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = \frac{x_k \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)},$$

and therefore we have

$$\sum_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i) \frac{x_k - \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)} = 0.$$

We assume $\hat{\boldsymbol{\theta}}_i \in (0,1)$, and thus we have

$$\sum_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i)(x_k - \hat{\boldsymbol{\theta}}_i) = 0,$$

which gives the solution

$$\hat{\boldsymbol{\theta}}_i = \frac{\sum\limits_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i) x_k}{\sum\limits_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i)}.$$

(c) Thus $\hat{\boldsymbol{\theta}}_i$, the maximum-likelihood estimate of $\boldsymbol{\theta}_i$, is a weighted average of the x_k's, with the weights being the posteriori probabilities of the mixing weights, i.e., $\hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i)$ for $k = 1, \ldots, n$.

5. We have a c-component mixture of Gaussians with each component of the form

$$p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) \sim N(\boldsymbol{\mu}_i, \sigma_i^2 \mathbf{I}),$$

or more explicitly,

$$p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{d/2}\sigma_i^d} \exp\left[-\frac{1}{2\sigma_i^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)\right].$$

We take the logarithm and find

$$\ln p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = -\frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln \sigma_i^2 - \frac{1}{2\sigma_i^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i),$$

and thus the derivative with respect to the variance is

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)}{\partial \sigma_i^2} &= -\frac{d}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i) \\ &= \frac{1}{2\sigma_i^4}(-d\sigma_i^2 + \|\mathbf{x}-\boldsymbol{\mu}_i\|^2). \end{aligned}$$

The maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_i$ must satisfy Eq. 12 in the text, that is,

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \nabla_{\boldsymbol{\theta}_i} \ln p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0.$$

We set the derivative with respect to σ_i^2 to zero, i.e.,

$$\begin{aligned} \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \frac{\partial \ln p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i)}{\partial \sigma_i^2} &= \\ \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \frac{1}{2\hat{\sigma}_i^4}(-d\hat{\sigma}_i^2 + \|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2) &= 0, \end{aligned}$$

rearrange, and find

$$d\hat{\sigma}_i^2 \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) = \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2.$$

The solution is

$$\hat{\sigma}_i^2 = \frac{\frac{1}{d}\sum\limits_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{\sum\limits_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)},$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)$, the maximum-likelihood estimates of $\boldsymbol{\mu}_i$ and $P(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}_i)$, are given by Eqs. 11–13 in the text.

6. Our c-component normal mixture is

$$p(\mathbf{x}|\alpha) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \alpha)P(\omega_j),$$

and the sample log-likelihood function is

$$l = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\alpha).$$

We take the derivative with respect to α and find

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \sum_{k=1}^{n} \frac{\partial \ln p(\mathbf{x}_k|\alpha)}{\partial \alpha} = \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k, \alpha)} \frac{\partial p(\mathbf{x}_k, \alpha)}{\partial \alpha} \\
&= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k, \alpha)} \frac{\partial}{\partial \alpha} \sum_{l=1}^{c} p(\mathbf{x}_k|\omega_j, \alpha)P(\omega_j) \\
&= \sum_{k=1}^{n}\sum_{j=1}^{c} \frac{p(\mathbf{x}_k|\omega_j, \alpha)P(\omega_j)}{p(\mathbf{x}_k, \alpha)} \frac{\partial}{\partial \alpha} \ln p(\mathbf{x}_k|\omega_j, \alpha) \\
&= \sum_{k=1}^{n}\sum_{j=1}^{c} P(\omega_j|\mathbf{x}_k, \alpha) \frac{\partial \ln p(\mathbf{x}_k|\omega_j, \alpha)}{\partial \alpha},
\end{aligned}$$

where by Bayes' Theorem we used

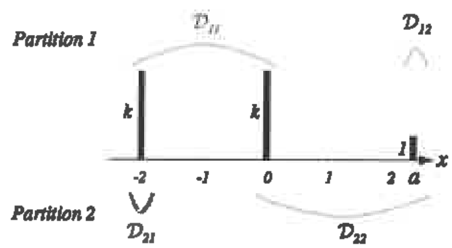
$$P(\omega_j|\mathbf{x}_k, \alpha) = \frac{p(\mathbf{x}_k|\omega_j, \alpha)P(\omega_j)}{p(\mathbf{x}_k|\alpha)}.$$

7. PROBLEM NOT YET SOLVED

8. We are given that θ_1 and θ_2 are statistically independent, i.e., $p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2)$.

(a) We use this assumption to derive

$$p(\theta_1, \theta_2|x_1) = \frac{p(\theta_1, \theta_2, x_1)}{p(x_1)} = \frac{p(x_1|\theta_1, \theta_2)p(\theta_1, \theta_2)}{p(x_1)}$$



and thus our cluster criterion function is

$$\begin{aligned}
J_{e1} &= \sum_{x\in\mathcal{D}_{11}} (x-m_{11})^2 + \sum_{x\in\mathcal{D}_{12}} (x-m_{12})^2 \\
&= \sum_{i=1}^{k}(-2+1)^2 + \sum_{i=1}^{k}(0+1)^2 + (a-a)^2 \\
&= k+k+0 = 2k.
\end{aligned}$$

In **Partition 2**, we have

$$\begin{aligned}
m_{21} &= -2, \\
m_{22} &= \frac{k\cdot 0 + a}{k+1} = \frac{a}{k+1},
\end{aligned}$$

and thus our cluster criterion function is

$$\begin{aligned}
J_{e2} &= \sum_{x\in\mathcal{D}_{21}} (x-m_{21})^2 + \sum_{x\in\mathcal{D}_{22}} (x-m_{22})^2 \\
&= \sum_{x\in\mathcal{D}_{21}} (-2+2)^2 + \sum_{i=1}^{k}\left(0-\frac{a}{k+1}\right)^2 + \left(a-\frac{a}{k+1}\right)^2 \\
&= 0 + \frac{a^2}{(k+1)^2}k + \frac{a^2k^2}{(k+1)^2} \\
&= \frac{a^2k(k+1)}{(k+1)^2} = \frac{a^2k}{k+1}.
\end{aligned}$$

Thus if $J_{e2} < J_{e1}$, i.e., if $a^2/(k+1) < 2k$ or equivalently $a^2 < 2(k+1)$, then the partition that minimizes J_e is Partition 2, which groups the k samples at $x=0$ with the one sample at $x=a$.

(b) If $J_{e1} < J_{e2}$, i.e., $2k < a^2/(k+1)$ or equivalently $2(k+1) > a^2$, then the partition that minimzes J_e is Partition 1, which groups the k-samples at $x=-2$ with the k samples at $x=0$.

28. Our sum-of-square (scatter) criterion is $\mathrm{tr}[\mathbf{S}_W]$. We thus need to calculate $\mathbf{S}_W$, i.e.,

$$\mathbf{S}_W = \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$$

$$\begin{aligned}
&= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}[\mathbf{x}\mathbf{x}^t - \mathbf{m}_i\mathbf{x}^t - \mathbf{x}\mathbf{m}_i^t + \mathbf{m}_i\mathbf{m}^t] \\
&= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}\mathbf{x}^t - \sum_{i=1}^{c} m_1 n_i \mathbf{m}_i^t - \sum_{i=1}^{c} n_i\mathbf{m}_i\mathbf{m}_i^t + \sum_{i=1}^{c} n_i\mathbf{m}_i\mathbf{m}_i^t \\
&= \sum_{k=1}^{4}\mathbf{x}_k\mathbf{x}_k^t - \sum_{i=1}^{c} n_i\mathbf{m}_i\mathbf{m}_i^t,
\end{aligned}$$

where n_i is the number of samples in $\mathcal{D}_i$ and

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}.$$

For the data given in the problem we have the following:

$$\begin{aligned}
\sum_{k=1}^{4}\mathbf{x}_k\mathbf{x}_k^t &= \begin{pmatrix}4\\5\end{pmatrix}(4 \quad 5) + \begin{pmatrix}1\\4\end{pmatrix}(1 \quad 4) + \begin{pmatrix}0\\1\end{pmatrix}(0 \quad 1) + \begin{pmatrix}5\\0\end{pmatrix}(5 \quad 0) \\
&= \begin{pmatrix}16 & 20\\20 & 25\end{pmatrix} + \begin{pmatrix}1 & 4\\4 & 16\end{pmatrix} + \begin{pmatrix}0 & 0\\0 & 1\end{pmatrix} + \begin{pmatrix}25 & 0\\0 & 0\end{pmatrix} \\
&= \begin{pmatrix}42 & 24\\24 & 42\end{pmatrix}.
\end{aligned}$$

Partition 1:

$$\begin{aligned}
\mathbf{m}_1 &= \frac{1}{2}\left(\begin{pmatrix}4\\5\end{pmatrix} + \begin{pmatrix}1\\4\end{pmatrix}\right) = \begin{pmatrix}5/2\\9/2\end{pmatrix}, \\
\mathbf{m}_2 &= \frac{1}{2}\left(\begin{pmatrix}0\\1\end{pmatrix} + \begin{pmatrix}5\\0\end{pmatrix}\right) = \begin{pmatrix}5/2\\1/2\end{pmatrix},
\end{aligned}$$

and thus

$$\mathbf{m}_1\mathbf{m}_1^t = \begin{pmatrix}25/4 & 45/4\\45/4 & 81/4\end{pmatrix} \text{ and } \mathbf{m}_2\mathbf{m}_2^t = \begin{pmatrix}25/4 & 5/4\\5/4 & 1/4\end{pmatrix}.$$

Our scatter matrix is therefore

$$\mathbf{S}_W = \begin{pmatrix}42 & 24\\24 & 42\end{pmatrix} - 2\begin{pmatrix}25/4 & 45/4\\45/4 & 81/4\end{pmatrix} - 2\begin{pmatrix}25/4 & 5/4\\5/4 & 1/4\end{pmatrix} = \begin{pmatrix}17 & -1\\-1 & 1\end{pmatrix},$$

and thus our criterion values are the trace

$$\text{tr}[\mathbf{S}_W] = \text{tr}\begin{pmatrix}17 & -1\\-1 & 1\end{pmatrix} = 17 + 1 = 18,$$

and the determinant

$$|\mathbf{S}_W| = 17\cdot 1 - (-1)\cdot(-1) = 16.$$

Partition 2:

$$\begin{aligned}
\mathbf{m}_1 &= \frac{1}{2}\left(\begin{pmatrix}4\\5\end{pmatrix} + \begin{pmatrix}5\\0\end{pmatrix}\right) = \begin{pmatrix}9/2\\5/2\end{pmatrix}, \\
\mathbf{m}_2 &= \frac{1}{2}\left(\begin{pmatrix}1\\4\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\right) = \begin{pmatrix}1/2\\5/2\end{pmatrix},
\end{aligned}$$

and thus

$$\mathbf{m}_1\mathbf{m}_1^t = \begin{pmatrix} 81/4 & 45/4 \\ 45/4 & 25/4 \end{pmatrix} \text{ and } \mathbf{m}_2\mathbf{m}_2^t = \begin{pmatrix} 1/4 & 5/4 \\ 5/4 & 25/4 \end{pmatrix}.$$

Our scatter matrix is therefore

$$\mathbf{S}_W = \begin{pmatrix} 42 & 24 \\ 24 & 42 \end{pmatrix} - 2\begin{pmatrix} 81/4 & 45/4 \\ 45/4 & 25/4 \end{pmatrix} - 2\begin{pmatrix} 1/4 & 5/4 \\ 5/4 & 25/4 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 17 \end{pmatrix},$$

and thus our criterion values are the trace

$$\mathrm{tr}[\mathbf{S}_W] = \mathrm{tr}\begin{pmatrix} 17 & -1 \\ -1 & 1 \end{pmatrix} = 1 + 17 = 18,$$

and the determinant

$$|\mathbf{S}_W| = 1 \cdot 17 - (-1) \cdot (-1) = 16.$$

Partition 3:

$$\begin{aligned} \mathbf{m}_1 &= \frac{1}{3}\left(\begin{pmatrix} 4 \\ 5 \end{pmatrix} + \begin{pmatrix} 1 \\ 4 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 5/3 \\ 3 \end{pmatrix}, \\ \mathbf{m}_2 &= \begin{pmatrix} 5 \\ 0 \end{pmatrix}. \end{aligned}$$

and thus

$$\mathbf{m}_1\mathbf{m}_1^t = \begin{pmatrix} 25/9 & 5 \\ 5 & 9 \end{pmatrix} \text{ and } \mathbf{m}_2\mathbf{m}_2^t = \begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix}.$$

Our scatter matrix is therefore

$$\mathbf{S}_W = \begin{pmatrix} 42 & 24 \\ 24 & 42 \end{pmatrix} - 3\begin{pmatrix} 25/9 & 5 \\ 5 & 9 \end{pmatrix} - 1\begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 26/3 & 22/3 \\ 22/3 & 26/3 \end{pmatrix},$$

and thus our criterion values are

$$\mathrm{tr}\ \mathbf{S}_W = \mathrm{tr}\begin{pmatrix} 17 & -1 \\ -1 & 1 \end{pmatrix} = 26/3 + 26/3 = 17.33,$$

and

$$|\mathbf{S}_W| = 26/3 \cdot 26/3 - 22/3 \cdot 22/3 = 21.33.$$

We summarize our results as

Partition	$\mathrm{tr}[\mathbf{S}_W]$	$\lvert\mathbf{S}_W\rvert$
1	18	16
2	18	16
3	17.33	21.33

Thus for the tr$[\mathbf{S}_W]$ criterion Partition 3 is favored; for the $|\mathbf{S}_W|$ criterion Partitions 1 and 2 are equal, and are to be favored over Partition 3.

24. PROBLEM NOT YET SOLVED

~~26~~. Consider a non-singular transformation of the feature space: $y = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a $d \times d$ non-singular matrix.

(a) If we let $\tilde{\mathcal{D}}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{D}_i\}$ denote the data set transformed to the new space, then the scatter matrix in the transformed domain can be written as

$$\begin{aligned}
\mathbf{S}_W^y &= \sum_{i=1}^{c} \sum_{\mathbf{y}\in\tilde{\mathcal{D}}_i} (\mathbf{y} - \mathbf{m}_i^y)(\mathbf{y} - \mathbf{m}_i^y)^t \\
&= \sum_{i=1}^{c} \sum_{\mathbf{y}\in\tilde{\mathcal{D}}_i} (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i)(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i)^t \\
&= \mathbf{A}\sum_{i=1}^{c} \sum_{\mathbf{y}\in\tilde{\mathcal{D}}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t
\end{aligned}$$

where $\mathbf{A}^t = \mathbf{A}\mathbf{S}_W\mathbf{A}^t$. We also have the between-scatter matrix

$$\begin{aligned}
\mathbf{S}_B^y &= \sum_{i=1}^{c} n_i(\mathbf{m}_i^y - \mathbf{m}^y)(\mathbf{m}_i - \mathbf{m}^y)^t \\
&= \sum_{i=1}^{c} n_i(\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m})(\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m})^t \\
&= \mathbf{A}\left[\sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t\right]\mathbf{A}^t \\
&= \mathbf{A}\mathbf{S}_B\mathbf{A}^t.
\end{aligned}$$

The product of the inverse matrices is

$$\begin{aligned}
[\mathbf{S}_W^y]^{-1}[\mathbf{S}_B^y]^{-1} &= (\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_B\mathbf{A}^t) \\
&= (\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_B\mathbf{A}^t \\
&= (\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{A}^t.
\end{aligned}$$

We let λ_i denote the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$ for $i = 1, \ldots, d$. There exist vectors $\mathbf{z}_i, \ldots, \mathbf{z}_d$ such that

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{z}_i = \lambda_i\mathbf{z}_i,$$

for $i = 1, \ldots, d$, and this in turn implies

$$\begin{aligned}
(\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{A}^t\mathbf{A}^t)^{-1}\mathbf{z}_i &= \lambda_i(\mathbf{A}^t)^{-1}\mathbf{z}_i, \text{or} \\
\mathbf{S}_W^{y-1}\mathbf{S}_B^y\mathbf{u}_i &= \lambda_i\mathbf{u}_i,
\end{aligned}$$

where $\mathbf{u}_i = (\mathbf{A}^t)^{-1}\mathbf{z}_i$. This implies that $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{S}_W^{y-1}\mathbf{S}_B^y$, and finally that $\lambda_1, \ldots, \lambda_d$ are invariant to non-singular linear transformation of the data.

(b) Our total scatter matrix is $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$, and thus

$$\begin{aligned}\mathbf{S}_T^{-1}\mathbf{S}_W &= (\mathbf{S}_B + \mathbf{S}_W)^{-1}\mathbf{S}_W \\ &= [\mathbf{S}_W^{-1}(\mathbf{S}_B + \mathbf{S}_W)]^{-1} \\ &= [\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}.\end{aligned}$$

If $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$ and the $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are the corresponding eigenvectors, then $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for $i = 1, \ldots, d$ and hence

$$\mathbf{u}_i + \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u}_i = \mathbf{u}_i + \lambda_i\mathbf{u}_i.$$

This equation implies

$$[\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]\mathbf{u}_i = (1 + \lambda_i)\mathbf{u}_i.$$

We multiply both sides of the equation by $(1+\lambda_i)^{-1}[\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}$ and find

$$(1+\lambda_i)^{-1}\mathbf{u}_i = [\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}\mathbf{u}_i$$

and this implies $\nu_i = 1/(1+\lambda_i)$ for $i = 1, \ldots, d$ are eigenvalues of $\mathbf{I}+\mathbf{S}_W^{-1}\mathbf{S}_B$.

(c) We use our result from part (??) and find

$$J_d = \frac{|\mathbf{S}_W|}{|\mathbf{S}_T|} = |\mathbf{S}_T^{-1}\mathbf{S}_W| = \prod_{i=1}^{d}\nu_i = \prod_{i=1}^{d}\frac{1}{1+\lambda_i},$$

which is invariant to non-singular linear transformations described in part (??).

26. Problem not yet solved

27. Equation 68 in the text defines the criterion $J_d = |\mathbf{S}_W| = \left|\sum_{i=1}^{c}\mathbf{S}_i\right|$, where

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

is the scatter matrix for categoryω_i defined in Eq. 61 in the text. We let $\mathbf{T}$ be a non-singular matrix and consider the change of variables $\mathbf{x}' = \mathbf{T}\mathbf{x}$.

(a) From the conditions stated, we have

$$\mathbf{m}_i' = \frac{1}{n_i}\sum_{\mathbf{x}'\in\mathcal{D}_{i'}}\mathbf{x}'$$

where n_i is the number of points in category ω_i. Thus we have

$$\mathbf{m}_i' = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{T}\mathbf{x} = \mathbf{T}\mathbf{m}_i.$$

Furthermore, we have

$$\begin{aligned}\mathbf{S}_{i'} &= \sum_{\mathbf{x}'\in\mathcal{D}_i'}(\mathbf{x}' - \mathbf{m}_i')(\mathbf{x}' - \mathbf{m}_i')^t \\ &= \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{T}\mathbf{x} - \mathbf{T}\mathbf{m}_i)(\mathbf{T}\mathbf{x} - \mathbf{T}\mathbf{m}_i)^t \\ &= \mathbf{T}\left[\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t\right]\mathbf{T}^t = \mathbf{T}\mathbf{S}_i\mathbf{T}^t.\end{aligned}$$

(b) From the conditions stated by the problem, the criterion function of the transformed data must obey

$$\begin{aligned} J'_d = |\mathbf{S}'_W| &= \left|\sum_{i=1}^{c}\mathbf{S}'_i\right| = \left|\sum_{i=1}^{c}\mathbf{T}\mathbf{S}_i\mathbf{T}^t\right| = \left|\mathbf{T}\left(\sum_{i=1}^{c}\mathbf{S}_i\right)\mathbf{T}^t\right| \\ &= |\mathbf{T}||\mathbf{T}^t|\left|\sum_{i=1}^{c}\mathbf{S}_i\right| \\ &= |\mathbf{T}|^2 J_d. \end{aligned}$$

Therefore, J'_d differs from J_d only by an overall non-negative scale factor $|\mathbf{T}|^2$.

(c) Since J'_d differs from J_d only by a scale factor of $|\mathbf{T}|^2$ (which does not depend on the partitioning into clusters) J'_d and J_d will rank partitions in the same order. Hence the optimal clustering based on J_d is always the optimal clustering based on J'_d. Optimal clustering are invariant to non-singular linear transformations of the data .

28. PROBLEM NOT YET SOLVED

29. PROBLEM NOT YET SOLVED

Section 10.8

28 **~~30~~.** Our generalization of the basic minimum-squared-error procedure uses the criterion function

$$J_T = \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i).$$

(a) We consider a non-singular transformation of the feature space of the form $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a $d\times d$ non-singular matrix. We let $\tilde{\mathcal{D}}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x}\in\mathcal{D}_i\}$ denote the transformed data set. Then, we have

$$\begin{aligned} J_T^y &= \sum_{i=1}^{c}\sum_{\mathbf{y}\in\tilde{\mathcal{D}}_i}(\mathbf{y}-\mathbf{m}_i^y)^t\mathbf{S}_T^{y-1}(\mathbf{y}-\mathbf{m}_i^y) \\ &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{m}_i)^t\mathbf{S}_T^{y-1}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{m}_i). \end{aligned}$$

As mentioned in the solution to Problem 25, the within- and between-scatter matrixes transform according to:

$$\begin{aligned} \mathbf{S}_W^y &= \mathbf{A}\mathbf{S}_W\mathbf{A}^t \quad \text{and} \\ \mathbf{S}_B^y &= \mathbf{A}\mathbf{S}_B\mathbf{A}^t. \end{aligned}$$

Thus the scatter matrix in the transformed coordinates is

$$\mathbf{S}_T^y = \mathbf{S}_W^y + \mathbf{S}_B^y = \mathbf{A}(\mathbf{S}_W+\mathbf{S}_B)\mathbf{A}^t = \mathbf{A}\mathbf{S}_T\mathbf{A}^t,$$

and this implies

$$[\mathbf{S}_T^y]^{-1} = (\mathbf{A}\mathbf{S}_T\mathbf{A}^t)^{-1} = (\mathbf{A}^t)^{-1}\mathbf{S}_T^{-1}\mathbf{A}^{-1}.$$

Therefore, we have

$$\begin{aligned}
J_T^y &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}\sum(\mathbf{A}(\mathbf{x}-\mathbf{m}_i))^t(\mathbf{A}^t)^{-1}\mathbf{S}_T^{-1}\mathbf{A}^{-1}(\mathbf{A}(\mathbf{x}-\mathbf{m}_i))\\
&= \sum_{i=1}^{c}(\mathbf{x}-\mathbf{m}_i)^t(\mathbf{A}^t)(\mathbf{A}^t)^{-1}\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i)\\
&= \sum_{i=1}^{c}(\mathbf{x}-\mathbf{m}_i)\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i) = J_T.
\end{aligned}$$

In short, then, J_T is invariant to non-singular linear transformation of the data.

(b) We consider sample $\hat{\mathbf{x}}$ being transferred from $\mathcal{D}_i$ to $\mathcal{D}_j$. Recall that the total scatter matrix $\mathbf{S}_T = \sum_{\mathbf{x}}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t$, given by Eq. 64 in the text, does not change as a result of changing the partition. Therefore

$$J_T^* = \sum_{k=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_k^*}(\mathbf{x}-\mathbf{m}_k^*)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_k^*),$$

where

$$\mathcal{D}_k^* = \begin{cases} H_k & \text{if } k\neq i,j\\ \mathcal{D}_i - \{\hat{\mathbf{x}}\} & \text{if } k=i\\ \mathcal{D}_j + \{\hat{\mathbf{x}}\} & \text{if } k=j.\end{cases}$$

We note the following values of the means after transfer of the point:

$$\begin{aligned}
\mathbf{m}_k^* &= \mathbf{m}_k \text{ if } k\neq i,j,\\
\mathbf{m}_i^* &= \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_i^*}\mathbf{x}}{\sum\limits_{\mathbf{x}\in\mathcal{D}_i^*}1} = \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}-\hat{\mathbf{x}}}{n_i-1}\\
&= \frac{n_i m_i - \hat{\mathbf{x}}}{n_i-1} = \frac{(n_i-1)\mathbf{m}_i - (\hat{\mathbf{x}}-\mathbf{m}_i)}{n_i-1}\\
&= \mathbf{m}_i - \frac{\hat{\mathbf{x}}-\mathbf{m}_i}{n_i-1},\\
\mathbf{m}_j^* &= \frac{\sum\limits_{\mathbf{x}\in H_i}\mathbf{x}+\hat{\mathbf{x}}}{n_i+1}\\
&= \frac{n_j m_j + \hat{\mathbf{x}}}{n_j+1} = \frac{(n_j+1)\mathbf{m}_j + (\hat{\mathbf{x}}-\mathbf{m}_j)}{n_j+1}\\
&= \mathbf{m}_j + \frac{\hat{\mathbf{x}}-\mathbf{m}_j}{n_j+1}.
\end{aligned}$$

Thus our criterion function is

$$\begin{aligned}
J_T^* &= \sum_{k=1,k\neq i,j}^{c}(\mathbf{x}-\mathbf{m}_k)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_k) + \sum_{\mathbf{x}\in\mathcal{D}_i^*}(\mathbf{x}-\mathbf{m}_i^*)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i^*)\\
&\quad + \sum_{\mathbf{x}\in\mathcal{D}_j^*}(\mathbf{x}-\mathbf{m}_j^*)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_j). \qquad (*)
\end{aligned}$$

We expand

$$
\begin{aligned}
& \sum_{\mathbf{x}\in\mathcal{D}_i^*} (\mathbf{x}-\mathbf{m}_i^*)^t \mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i^*) + \sum_{\mathbf{x}\in\mathcal{D}_j^*} (\mathbf{x}-\mathbf{m}_j^*)^t \mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_j^*) \\
= & \sum_{\mathbf{x}\in\mathcal{D}_i^*} \mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x} - n_i^*\mathbf{m}_i^*\mathbf{m}_i^{*t} + \sum_{\mathbf{x}\in\mathcal{D}_j^*} \mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x} - n_j^*\mathbf{m}_j^*\mathbf{m}_j^{*t}\mathbf{S}_T^{-1}\mathbf{m}_j^* \\
= & \sum_{\mathbf{x}\in\mathcal{D}_i^*} \mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x} - \hat{\mathbf{x}}\mathbf{S}_T^{-1}\hat{\mathbf{x}} - (n_i-1)\left(\mathbf{m}_i - \frac{\hat{\mathbf{x}}-\mathbf{m}_i}{n_i-1}\right)^t \mathbf{S}_T^{-1}\left(\mathbf{m}_i - \frac{\hat{\mathbf{x}}-\mathbf{m}_i}{n_i-1}\right) \\
& + \sum_{\mathbf{x}\in\mathcal{D}_j^*} \mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x} + \hat{\mathbf{x}}\mathbf{S}_T^{-1}\hat{\mathbf{x}} - (n_j+1)\left(\mathbf{m}_j + \frac{\hat{\mathbf{x}}-\mathbf{m}_j}{n_j+1}\right)^t \mathbf{S}_T^{-1}\left(\mathbf{m}_j + \frac{\hat{\mathbf{x}}-\mathbf{m}_j}{n_j+1}\right) \\
= & \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i) - \hat{\mathbf{x}}^t\mathbf{S}_T^{-1}\hat{\mathbf{x}} + \mathbf{m}_i\mathbf{S}_T^{-1}\mathbf{m}_i + 2\mathbf{m}_i^t\mathbf{S}_T^{-1}\hat{\mathbf{x}} - 2\mathbf{m}_i^t\mathbf{S}_T^{-1}\mathbf{m}_i \\
& - \frac{1}{n_i-1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i) \\
& + \sum_{\mathbf{x}\in\mathcal{D}_j} (\mathbf{x}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_j) + \hat{\mathbf{x}}^t\mathbf{S}_T^{-1}\hat{\mathbf{x}} - \mathbf{m}_j\mathbf{S}_T^{-1}\mathbf{m}_j + 2\mathbf{m}_j^t\mathbf{S}_T^{-1}\hat{\mathbf{x}} + 2\mathbf{m}_j^t\mathbf{S}_T^{-1}\mathbf{m}_j \\
& - \frac{1}{n_j+1}(\hat{\mathbf{x}}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_j) \\
= & \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i) - \frac{n_i}{n_i+1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i) \\
& + \frac{n_j}{n_j+1}(\hat{\mathbf{x}}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_j).
\end{aligned}
$$

We substitute this result in $(*)$ and find

$$
\begin{aligned}
J_T^* &= \sum_{k=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_k} (\mathbf{x}-\mathbf{m}_k)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_k) \\
& \quad + \left[\frac{n_j}{n_j+1}(\mathbf{x}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_j) - \frac{n_i}{n_i-1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i)\right] \\
&= J_T + \left[\frac{n_j}{n_j+1}(\hat{\mathbf{x}}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_j) - \frac{n_i}{n_i-1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i)\right].
\end{aligned}
$$

(c) If we let $\mathcal{D}$ denote the data set and n the number of points, the algorithm is:

Algorithm 3 (Minimizing J_T)

1 **begin initialize** $\mathcal{D}, c$
2 Compute c means $\mathbf{m}_1, \ldots, \mathbf{m}_c$
3 Compute J_T
4 **do** Randomly select a sample; call it $\hat{\mathbf{x}}$
5 Determine closest mean to $\hat{\mathbf{x}}$; call it $\mathbf{m}_j$
6 **if** $n_i = 1$ **then** go to line 10
7 **if** $j \neq i$ **then** $\rho_j \leftarrow \frac{n_j}{n_j+1}(\hat{\mathbf{x}}-\mathbf{m}_j)^t S_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_j)$
8 **if** $j = 1$ **then** $\rho_j \leftarrow \frac{n_i}{n_i-1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t S_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i)$
9 **if** $\rho_k \leq \rho_j$ forall j **then** transfer $\hat{\mathbf{x}}$ to $\mathcal{D}_k$

Update $J_T, \mathbf{m}_i$, and $\mathbf{m}_k$
until J_T has not changed in n attempts
end

29 **31.** The total scatter matrix, $\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$, given by Eq. 64 in the text, does not change. Thus our criterion function is

$$J_e = \text{tr}[\mathbf{S}_W] = \text{tr}[\mathbf{S}_T - \mathbf{S}_B] = \text{tr}[\mathbf{S}_T] - \text{tr}[\mathbf{S}_B].$$

We let J_e^* be the criterion function which results from tranferring a sample $\hat{\mathbf{x}}$ from $\mathcal{D}_i$ to $\mathcal{D}_j$. Thus we have

$$\begin{aligned}
J_e^* &= \text{tr}[\mathbf{S}_T] - \text{tr}[\mathbf{S}_B^*] \\
&= \text{tr}[\mathbf{S}_T] - \sum_k n_k^* \|\mathbf{m}_k^* - \mathbf{m}\|^2 \\
&= \text{tr}[\mathbf{S}_T] - \sum_{k \neq i,j} n_k^* \|\mathbf{m}_k^* - \mathbf{m}\|^2 - \sum_{k=i,j} n_k^* \|\mathbf{m}_k^* - \mathbf{m}\|^2 \\
&= \text{tr}[\mathbf{S}_T] - \sum_{k \neq i,j} n_k \|\mathbf{m}_k - \mathbf{m}\|^2 - n_i^* \|\mathbf{m}_i^* - \mathbf{m}\|^2 - n_j^* \|\mathbf{m}_j^* - \mathbf{m}\|^2.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
n_i^* \|\mathbf{m}_i^* - \mathbf{m}\|^2 + n_j^* \|\mathbf{m}_j^* - \mathbf{m}\|^2 &= (n_i - 1) \left\| \mathbf{m}_i - \frac{\hat{\mathbf{x}} - \mathbf{m}_i}{n_i - 1} - \mathbf{m} \right\|^2 \\
&\quad + (n_j + 1) \left\| \mathbf{m}_j + \frac{\hat{\mathbf{x}} - \mathbf{m}_j}{n_j + 1} - \mathbf{m} \right\|^2,
\end{aligned}$$

as shown in Problem 30. We thus find that the means change by

$$\begin{aligned}
\mathbf{m}_i^* &= \mathbf{m}_i - \frac{\hat{\mathbf{x}} - \mathbf{m}_i}{n_i - 1}, \\
\mathbf{m}_j^* &= \mathbf{m}_j + \frac{\hat{\mathbf{x}} - \mathbf{m}_j}{n_j + 1}.
\end{aligned}$$

We substitute these above and find

$$\begin{aligned}
J_e^* &= (n_i - 1) \left[\|\mathbf{m}_i - \mathbf{m}\|^2 + \frac{1}{(n_i - 1)^2} \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 - \frac{2}{n_i - 1} (\hat{\mathbf{x}} - \mathbf{m}_i)^t (\mathbf{m}_i - \mathbf{m}) \right] \\
&\quad + (n_j - 1) \left[\|\mathbf{m}_j - \mathbf{m}\|^2 + \frac{1}{(n_j + 1)^2} \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 - \frac{2}{n_j + 1} (\hat{\mathbf{x}} - \mathbf{m}_j)^t (\mathbf{m}_j - \mathbf{m}) \right] \\
&= n_i \|\mathbf{m}_i - \mathbf{m}\|^2 - \|\mathbf{m}_i - \mathbf{m}\|^2 - 2(\hat{\mathbf{x}} - \mathbf{m}_i)^t (\mathbf{m}_i - \mathbf{m}) + \frac{1}{n_i - 1} \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 \\
&\quad + n_j \|\mathbf{m}_j - \mathbf{m}\|^2 - \|\mathbf{m}_j - \mathbf{m}\|^2 + 2(\hat{\mathbf{x}} - \mathbf{m}_j)^t (\mathbf{m}_j - \mathbf{m}) + \frac{1}{n_j + 1} \hat{\mathbf{x}} - \mathbf{m}_j\|^2 \\
&= n_i \|\mathbf{m}_i - \mathbf{m}\|^2 + n_j \|\mathbf{m}_j - \mathbf{m}\|^2 + \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 - \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 \\
&\quad + \frac{1}{n_i - 1} \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 + \frac{1}{n_j + 1} \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 \\
&= n_i \|\mathbf{m}_i - \mathbf{m}\|^2 + n_j \|\mathbf{m}_j - \mathbf{m}\|^2 + \frac{n_i}{n_i - 1} \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 - \frac{n_j}{n_j + 1} \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2.
\end{aligned}$$

Therefore our criterion function is

$$\begin{aligned}
J_e^* &= \mathrm{tr}[\mathbf{S}_T] - \mathrm{tr}[\mathbf{S}_B^*] \\
&= \mathrm{tr}[\mathbf{S}_T] - \sum_k n_k\|\mathbf{m}_k - \mathbf{m}\|^2 - \frac{n_i}{n_i - 1}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 + \frac{n_j}{n_j + 1}\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 \\
&= J_e + \frac{n_j}{n_j + 1}\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 - \frac{n_i}{n_i - 1}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2.
\end{aligned}$$

Section 10.9

32. Our similarity measure is

$$s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^t\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}.$$

(a) We have that $\mathbf{x}$ and $\mathbf{x}'$ are d-dimensional vectors with $x_i = 1$ if $\mathbf{x}$ possesses the ith feature and $x_i = -1$ otherwise. The Euclidean length of the vectors obeys

$$\|\mathbf{x}\| = \|\mathbf{x}'\| = \sqrt{\sum_{i=1}^d x_i^2} = \sqrt{\sum_{i=1}^d 1} = \sqrt{d},$$

and thus we can write

$$\begin{aligned}
s(\mathbf{x}, \mathbf{x}') &= \frac{\mathbf{x}^t\mathbf{x}'}{\sqrt{d}\sqrt{d}} = \frac{1}{d}\sum_{i=1}^d x_i x_i' \\
&= \frac{1}{d}\ [\text{number of common features} - \text{number of features not common}] \\
&= \frac{1}{d}[\text{number of common features} - (d - \text{number of common features})\ ] \\
&= \frac{2}{d}\ (\text{number of common features})\ - 1.
\end{aligned}$$

(b) The length of the difference vector is

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}'\|^2 &= (\mathbf{x} - \mathbf{x}')^t(\mathbf{x} - \mathbf{x}') \\
&= \mathbf{x}^t\mathbf{x} + \mathbf{x}'^t\mathbf{x}' - 2\mathbf{x}^t\mathbf{x}' \\
&= \|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2s(\mathbf{x}, \mathbf{x}')\|\mathbf{x}\|\|\mathbf{x}'\| \\
&= d + d - 2s(\mathbf{x}, \mathbf{x}')\sqrt{d}\sqrt{d} \\
&= 2d[1 - s(\mathbf{x}, \mathbf{x}')],
\end{aligned}$$

where, from part (a), we used $\|\mathbf{x}\| = \|\mathbf{x}'\| = \sqrt{d}$.

33. Consider the following candidates for metrics or pseudometerics.

(a) Squared Euclidean distance:

$$s(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2 = \sum_{i=1}^d (x_i - x_i')^2.$$

$$
\begin{aligned}
d_{hk} &= \Big\|\mathbf{m}_h - \frac{n_i}{n_i+n_j}\mathbf{m}_i - \frac{n_j}{n_i+n_j}\mathbf{m}_j\Big\|^2 \\
&= \Big\|\frac{n_i}{n_i+n_j}\mathbf{m}_h - \frac{n_i}{n_i+n_j}\mathbf{m}_i + \frac{n_j}{n_i+n_j}\mathbf{m}_h - \frac{n_j}{n_i+n_j}\mathbf{m}_j\Big\|^2 \\
&= \Big\|\frac{n_i}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_i)\Big\|^2 + \Big\|\frac{n_j}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_j)\Big\|^2 \\
&\quad +2\frac{n_i}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_i)^t\frac{n_j}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_j) \\
&= \frac{n_i^2+n_in_j}{(n_i+n_j)^2}\|\mathbf{m}_h - \mathbf{m}_i\|^2 + \frac{n_j^2+n_in_j}{(n_i+n_j)^2}\|\mathbf{m}_h - \mathbf{m}_j\|^2 \\
&\quad + \frac{n_in_j}{(n_i+n_j)^2}[(m_h - m_i)^t(\mathbf{m}_i - \mathbf{m}_j) - (m_h - m_j)^t(\mathbf{m}_i - \mathbf{m}_j)] \\
&= \frac{n_i}{n_i+n_j}\|\mathbf{m}_h - \mathbf{m}_i\|^2 + \frac{n_j}{n_i+n_j}\|\mathbf{m}_h - \mathbf{m}_j\|^2 - \frac{n_in_j}{(n_i+n_j)^2}\|\mathbf{m}_h - \mathbf{m}_j\|^2 \\
&= \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma|d_{hi} - d_{hj}|,
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_i &= \frac{n_i}{n_i+n_j} \\
\alpha_j &= \frac{n_j}{n_i+n_j} \\
\beta &= -\frac{n_in_j}{(n_i+n_j)^2} = -\alpha_i\alpha_j \\
\gamma &= 0.
\end{aligned}
$$

33 35. The sum-of-squared-error criterion is given by Eq. 72 in the text:

$$
\begin{aligned}
J_e &= \sum_{i'=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_{i'}} \|\mathbf{x} - \mathbf{m}_{i'}\|^2 \\
&= \sum_{i'=1}^{c}\left[\sum_{\mathbf{x}\in\mathcal{D}_{i'}} \mathbf{x}^t\mathbf{x} - n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_{i'}\right] \\
&= \sum_{\mathbf{x}}\mathbf{x}^t\mathbf{x} - \sum_{i'=1}^{c} n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_i.
\end{aligned}
$$

We merge $\mathcal{D}_i$ and $\mathcal{D}_j$ into $\mathcal{D}_k$ and find an increase in the criterion function J_e of

$$
\begin{aligned}
\Delta &\equiv J_e^* - J_e = \sum_{\mathbf{x}}\mathbf{x}^t\mathbf{x} - \sum_{\substack{i'=1\\ i\neq k}}^{c} n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_{i'} - n_k\mathbf{m}_k^t\mathbf{m}_k \qquad \text{NEW} \\
&\quad - \left[\sum_{\mathbf{x}}\mathbf{x}^t\mathbf{x} - \underbrace{\sum_{\substack{i'=1\\ i'\neq i<j}}^{c} n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_i - n_i\mathbf{m}_i^t\mathbf{m}_i - n_j\mathbf{m}_j^t\mathbf{m}_j}\right] \\
&= n_i\mathbf{m}_i^t\mathbf{m}_i + n_j\mathbf{m}_j^t\mathbf{m}_j - n_k\mathbf{m}_k^t\mathbf{m}_k,
\end{aligned}
$$

where

$$
n_k = n_i + n_j
$$

$$\begin{aligned}
\mathbf{m}_k &= \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_k} \mathbf{x}}{\sum\limits_{\mathbf{x}\in\mathcal{D}_k} 1} = \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_i} \mathbf{x} + \sum\limits_{\mathbf{x}\in\mathcal{D}_j} \mathbf{x}}{n_i + n_j} = \frac{n_i\mathbf{m}_i + n_j\mathbf{m}_j}{n_i + n_j} \\
&= \frac{n_i}{n_i+n_j}\mathbf{m}_i + \frac{n_j}{n_i+n_j}\mathbf{m}_j \\
nn_k\mathbf{m}_k^t\mathbf{m}_k &= (n_i+n_j)\left[\frac{n_i}{n_i+n_j}\mathbf{m}_i + \frac{n_j}{n_i+n_j}\mathbf{m}_j\right]^t\left[\frac{n_i}{n_i+n_j}\mathbf{m}_i + \frac{n_j}{n_i+n_j}\mathbf{m}_j\right] \\
&= \frac{n_i^2}{n_i+n_j}\mathbf{m}_i^t\mathbf{m}_i + \frac{n_j^2}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_j + \frac{2n_in_j}{n_i+n_j}\mathbf{m}_i^t\mathbf{m}_j,
\end{aligned}$$

and hence

$$\begin{aligned}
\Delta &= \left(n_i - \frac{n_i^2}{n_i+n_j}\right)\mathbf{m}_i^t\mathbf{m}_i + \left(n_j - \frac{n_i^2}{n_i+n_j}\right)\mathbf{m}_j^t\mathbf{m}_j - \frac{2n_in_j}{n_i+n_j}\mathbf{m}_i^t\mathbf{m}_j \\
&= \frac{n_in_j}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_i + \frac{n_in_j}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_j - \frac{2n_in_j}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_j \\
&= \frac{n_in_j}{n_i+n_j}(\mathbf{m}_i^t\mathbf{m}_i + \mathbf{m}_j^t\mathbf{m}_j - 2\mathbf{m}_i^t\mathbf{m}_j) \\
&= \frac{n_in_j}{n_i+n_j}\|\mathbf{m}_i - \mathbf{m}_j\|^2.
\end{aligned}$$

The smallest increase in J_e corresponds to the smallest value of Δ, and this arises from the smallest value of

$$\frac{n_in_j}{n_i+n_j}\|\mathbf{m}_i - \mathbf{m}_j\|^2.$$

36. PROBLEM NOT YET SOLVED

37. PROBLEM NOT YET SOLVED

38. PROBLEM NOT YET SOLVED

39. PROBLEM NOT YET SOLVED

Section 10.10

40. PROBLEM NOT YET SOLVED

41. As given in Problem 35, the change in J_e due to the transfer of one point is

$$J_e(2) = J_e(1) - \frac{n_1n_2}{n_1+n_2}\|\mathbf{m}_1 - \mathbf{m}_2\|^2.$$

We calculate the expected value of $J_e(1)$ as:

$$\begin{aligned}
\mathcal{E}(J_e(1)) &= \mathcal{E}\left(\sum_{\mathbf{x}\in\mathcal{D}}\|\mathbf{x}-\mathbf{m}\|^2\right) \\
&= \sum_{i=1}^d \mathcal{E}\left(\sum_{\mathbf{x}\in\mathcal{D}}(x_i - m_i)^2\right) \\
&= \sum_{i=1}^d (n-1)\sigma^2,
\end{aligned}$$