T-61.231 Principles of Pattern Recognition

Answers to exercise 1: 1.10.2001

2. Principal Component Analysis (PCA) produces principal components orthogonal to each other. The principal component transformation is closely related to the Karhunen-Loévé and Hotelling transforms, all of which employ the same base idea of eigenvector use to create a linear transform.

   First, let x be a population of random vectors. Let $m_x = E(x)$ be the mean of the random vector population. Again, the covariance matrix of x is defined as $C_x = E(x - m_x)(x - m_x)^T$. Because C_x is real and symmetric, it is always possible to find a set of n orthonormal eigenvectors, and it can be stated that

   $$C_x = A^T \Lambda A$$

   Where A is the matrix whose rows are formed from the eigenvectors of C_x and Λ is a diagonal matrix with the corresponding eigenvalues. Taking A as the transformation matrix, the transformation can be written as

   $$y = A(x - m_x)$$

   Resulting from this transformation, the mean of the y vectors is zero and the covariance matrix of the y's can be written as

   $$C_y = AC_xA^T = AA^T\Lambda AA^T = \Lambda$$

   ie. the covariance matrix of y is a diagonal matrix consisting of the eigenvalues of the original covariance matrix C_x. Thus, as long as none of the eigenvalues are zero (which would indicate that one component had full correlation with some other and contained no additional data), the variance of the y's can easily be scaled to 1. In some cases this is taken a bit further by also decomposing the eigenvalue-matrix Λ and joining it into the transformation matrix A by setting

   $$C_x = A\Lambda A^T = A{\Lambda^{1/2}}^T I \Lambda^{1/2} A^T = A^* I A^{*T}$$

   When using the transformation matrix A^*, it is obvious that the variance of y is I (a diagonal matrix consisting of ones).

   By ordering the eigenvectors and values so that the first row of A corresponds to the largest eigenvalue and the last row to the smallest eigenvalue, or in order of decreasing variance after the transform, the transformation can reduce the amount of needed data. By taking the first m principal components to create A_m for the data projection, a transformation of the form

   $$\hat{y} \approx A_m(x - m_x)$$

   This projection is optimal in the sense that it minimizes the mean square error (MSE) for any approximation with m components. The MSE becomes, for an initially K dimensional set of data,

   $$E[||x - \hat{x}||] = \sum_{i=m}^{K} \lambda_i$$

   Example from Digital Image Processing, Gonzales & Woods, Addison-Wesley 1993; When performing the transformation on data obtained from a six-band multi-spectral scanner, the eigenvalues of the covariance matrixes were calculated. The resulting ordered eigenvalues were

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
Eigenvalue	3210	931.4	118.5	83.88	64.00	13.40
Percentage of total	72.61	21.07	2.68	1.9	1.45	0.3

3. The ambiguity function can be of great use in feature selection, especially in a multi-class situation. The estimation of the required probability functions is in general easy.

$$A = -\sum_i \sum_j P(\Delta_j) P(\omega_i|\Delta_j) \log_M(P(\omega_i|\Delta_j))$$

Completely overlapping distributions: $P(\omega_i|\Delta_j)$ is constant, $P(\omega_i|\Delta_j) = \frac{1}{M} \forall i$. Thus

$$\begin{aligned} A &= -\sum_i \sum_j P(\Delta_j) P(\omega_i|\Delta_j) \log_M(P(\omega_i|\Delta_j)) \\ &= -M \sum_{j=1}^{K} P(\Delta_j) \tfrac{1}{M} \log_M(\tfrac{1}{M}) \\ &= -\sum_{j=1}^{K} P(\Delta_j)(\log_M 1 - \log_M M) \\ &= -\sum_{j=1}^{K} P(\Delta_j)(0-1) \\ &= \sum_{j=1}^{K} P(\Delta_j) \\ &= 1 \end{aligned}$$

Completely separate distributions: $P(\omega_i|\Delta_j) = 1 \Rightarrow P(\omega_k|\Delta_j) = 0 \ \forall k \neq i$. Thus for all other distributions at each point the term $P(\Delta_j) P(\omega_i|\Delta_j) \log_M(P(\omega_i|\Delta_j)) = 0$, and

$$\begin{aligned} A &= -\sum_i \sum_j P(\Delta_j) P(\omega_i|\Delta_j) \log_M(P(\omega_i|\Delta_j)) \\ &= 0 + -\sum_{j=1}^{K} P(\Delta_j) \cdot 1 \cdot \log_M(1) \\ &= -\sum_{j=1}^{K} P(\Delta_j) \cdot 0 \\ &= 0 \end{aligned}$$

And for those concerned with the expression $P(\Delta_j) P(\omega_i|\Delta_j) \log_M(P(\omega_i|\Delta_j)) = P(\Delta_j) 0 \cdot \log_M(0) = 0$, it is a commonly accepted convention (see, for example, definitions of Entropy in books that care to define entropy "properly" also for zero probabilities - which is surprisingly often not the case, but for example *ftp://wol.ra.phy.cam.ac.uk/pub/mackay/info-theory/l1.pdf*) that $0 \cdot \log_k(0) = 0$, since $\lim_{x \to 0^+} x \log_k(x) = 0$ for all k.

4. The motivation for this to prove that variance for the estimated classification error can be calculated using formula 10.7 (Theodorodis p.339, Error Counting Approach). A help in solving this exercise is provided by knowledge of the binomial formula,

$$\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n$$

The expectation value can be written as

$$\begin{aligned} E[k] &= \sum_{k=0}^{N} \tbinom{N}{k} k P^k (1-P)^{N-k} \\ &= \sum_{k=0}^{N} \tfrac{N!}{k!(N-k)!} k P^k (1-P)^{N-k} \\ &= \sum_{k=1}^{N} \tfrac{N!}{k!(N-k)!} k P^k (1-P)^{N-k} \\ &= PN \sum_{k=1}^{N} \tfrac{(N-1)! P^{(k-1)} (1-P)^{(N-1-(k-1))}}{(k-1)!(N-1-(k-1))!} \\ &= PN \sum_{k=1}^{N} \tbinom{N-1}{k-1} P^{(k-1)} (1-P)^{N-1-(k-1)} \\ &= PN(P+1-P)^{(N-1)} \\ &= PN \end{aligned}$$

Since the variance can be stated as $var[k] = E[k^2] - E[k]^2$, we need to calculate $E[k^2]$.

$$\begin{aligned}
E[k^2] &= \textstyle\sum_{k=0}^{N} \binom{N}{k} k^2 P^k (1-P)^{N-k} \\
&= \textstyle\sum_{k=0}^{N} \frac{N!}{k!(N-k)!} k^2 P^k (1-P)^{N-k} \\
&= PN \textstyle\sum_{k=0}^{N} \frac{(N-1)! k P^{k-1} (1-P)^{(N-1-(k-1))}}{(k-1)!(N-1-(k-1))!} \\
&= PN \textstyle\sum_{l=k-1=0}^{L=N-1} \frac{(L)!(l+1) P^{l} (1-P)^{(L-l)}}{(l)!(L-l)!} \\
&= PN[PL \textstyle\sum_{l=0}^{L} \frac{(L-1)! P^{l} (1-P)^{(L-1-(l-1))}}{(l-1)!(L-1-(l-1))!} + 1] \\
&= PN[P(N-1) \cdot 1 + 1] \\
&= PN(PN - P + 1) \\
&= P^2N^2 - P^2N + PN
\end{aligned}$$

And thus

$$\begin{aligned}
var[k] &= E[k^2] - E[k]^2 \\
&= P^2N^2 - P^2N + PN - P^2N^2 \\
&= NP(1-P)
\end{aligned}$$