COURSE 02457

Non-Linear Signal Processing: Exercise 2

This exercise is based on C.M. Bishop: *Neural Networks for Pattern Recognition*, chapters 2 and section 8.6. The objective of the exercise is to become familiar with the 2D normal distribution, the notion of covariance, and the projection on eigenvectors as features.

Print and comment on the figures produced by the software main2a.m to main2e.m as outlined below at the four **Checkpoints**. A copy of the "journal" containing these comments and figures is returned to the tutor for evaluation.

Multivariate normal Distribution

Let $\mathbf{x}$ be a d-dimensional variable, i.e. $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. The probability of the variable $\mathbf{x}$ lying in a region, $\mathcal{A}$, which is a subspace of $\mathbf{R}^d$ is given by

$$P(\mathbf{x} \in \mathcal{A}) = \int_{\mathcal{A}} p(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where $p(\mathbf{x})$ is the probability density function of the variable $\mathbf{x}$.

In one dimension, the normal probability density function is given by

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2}$$

where μ and σ^2 are the mean and variance respectively. In d dimensions, the general multivariate normal probability density function is given by

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right), \tag{3}$$

where $\boldsymbol{\mu}$ is a d-dimensional vector, and $\boldsymbol{\Sigma}$ is a $d \times d$ covariance matrix.

2D normal Distribution

Let $\mathbf{x}$ be a 2-dimensional variable, so that $d = 2$ in the above equations. Let $\mathcal{D}$ be a set of N samples from $\mathbf{x}$, so that $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i = (x_{i,1}, x_{i,2})^T$, $i = 1, \ldots, N$.

It is then possible to construct a 2D histogram of the data-set, $\mathcal{D}$, by defining a Cartesian grid of small areas, where $\mathcal{A}_{j,k}$ and $j = 1, \ldots, M_1$, $k = 1, \ldots, M_2$. The histogram is then given by

$$H_{j,k} = \sum_{\mathbf{x}_i \in \mathcal{A}_{j,k}} 1\,, \quad j = 1, \ldots, M_1,\ k = 1, \ldots, M_2 \tag{4}$$

and the normalized histogram is given by

$$\tilde{H}_{j,k} = \frac{H_{j,k}}{\sum_{j',k'} H_{j',k'}}. \tag{5}$$

If the union of all the areas $\mathcal{A}_{j,k}$ includes all the samples in $\mathcal{D}$, equation (5) simplifies to

$$\tilde{H}_{j,k} = \frac{H_{j,k}}{N}. \tag{6}$$

The normalized histogram can be compared with the histogram approximation to the probability density function

$$P_{j,k} = \int_{\mathcal{A}_{j,k}} p(\mathbf{x}) d\mathbf{x}\,, \quad j = 1, \ldots, M_1,\ k = 1, \ldots, M_2. \tag{7}$$

Checkpoint 2.1:

Use the program main2a.m to illustrate a 2-dimensional normal probability density function given by a mean, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$. Create a normalized 2D histogram and compare it with the probability density function.

Discuss the dependence of the histogram on the number of samples, N. Compare this with the results from exercise 1 for a 1D normal distribution and relate this to the curse of dimensionality.

Interpretation of Covariance

A one-dimensional normal distribution is given by its mean, μ, and its variance, σ^2. The variance describes the variation of the variable around its mean.

In two dimensions, each sample consists of two components. Each component has a mean and a variance just as in the onedimensional case. Imagine a sample, $\mathbf{x} = (x_1, x_2)^T$ from a 2D normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$. If the variance of x_1, namely σ_1^2 is large, an individual sample of x_1 may well be quite different from μ_1, and similarly for x_2. However, there may be a trend that whenever x_1 is larger than μ_1, x_2 is also larger than μ_2, and that whenever x_1 is smaller than μ_1, x_2 is also smaller than μ_2. In such a case, x_1 and x_2 are not independent, and they are said to be correlated.

Another term is therefore needed to fully describe the variance of the variable, $\mathbf{x}$, namely the covariance between its components, $Cov(x_1, x_2) \equiv \mathcal{E}[(x_1 - \mu_1)(x_2 - \mu_2)]$. The covariance matrix of $\mathbf{x}$ is then given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & Cov(x_1, x_2) \\ Cov(x_2, x_1) & \sigma_2^2 \end{pmatrix} \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}. \tag{8}$$

The terms σ_{12} and σ_{21} are equal since they describe the covariance between the same components. The covariance matrix is therefore always symmetric. The magnitude of the covariance term for a given correlation between the two components also depends on the diagonal variance terms. A useful quantity describing the correlation between the components is the correlation coefficient, ρ. It is the normalized covariance and is given by

$$\rho = \frac{Cov(x_1, x_2)}{\sqrt{\sigma_1^2 \sigma_2^2}} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}. \tag{9}$$

where $\rho \in [-1, 1]$. However, the limiting case $\rho = \pm 1$ corresponds to a perfect linear relationship between x_1 and x_2. In this case the variable, $\mathbf{x}$, is not really 2-dimensional since one component completely defines the other.

Checkpoint 2.2:

Use the program main2b.m to visualize the probability density functions of 2D normal distributions with different covariance matrices. For example, try to fix the variances, σ_1^2 and σ_2^2, while only changing the covariance. Think of an example where there is no correlation between the components and implement this distribution. Comment on the dependence of the orientation and shape of the ellipsoids in the contour plots of quadratic form induced by the covariance matrix.

Coordinate Transformation

For some non-linear signal detection algorithms it is desired that the input should have zero mean, unit variance and zero covariance. The advantage of this is that it is possible to use the same algorithm (and not changing the control parameters of it) for variables of very different origins and covariation.

Geometrically, such a normalization corresponds to a coordinate transformation to the system defined by the eigenvectors of the covariance matrix. Typically, the mean and covariance matrix are not known, and must therefore be estimated from the data-set, $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$:

$$\widehat{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{10}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \widehat{\mathbf{x}})(\mathbf{x}_i - \widehat{\mathbf{x}})^T . \tag{11}$$

The eigenvalue equation for the covariance matrix is

$$\widehat{\boldsymbol{\Sigma}}\mathbf{u}_j = \lambda_j \mathbf{u}_j \, , \quad j = 1, \ldots, d \, , \tag{12}$$

where λ_j is the j'th eigenvalue and $\mathbf{u}_j$ is the corresponding eigenvector of $\widehat{\boldsymbol{\Sigma}}$. The transformed input variables are then given by

$$\tilde{\mathbf{x}}_i = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{x}_i - \widehat{\mathbf{x}}), \tag{13}$$

where

$$\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d) \tag{14}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_d) \, . \tag{15}$$

It can be shown that the transformed data-set, $\tilde{\mathcal{D}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_N\}$ has zero mean and a covariance matrix given by the unit matrix.

Checkpoint 2.3:

Use the program main2c.m to calculate the eigenvalues and eigenvectors of the covariance matrix for different distributions. Comment on the geometrical significance of the eigenvalues and eigenvectors. Compare the transformed data-sets from different distributions. What happens if the term $\boldsymbol{\Lambda}^{-1/2}$ is removed from equation (13)?

Projection on Eigenvectors

In some cases, the measured data is of a lower "true" dimension than the apparent dimension of the data vector. For example, imagine a data-set of a 3-dimensional variable. If all the data are on a straight line, the true dimension of the data is only 1D. If the data-set is transformed to a coordinate system, where the variation of the data is along one of the axes, the two other components can be ignored.

Let $\lambda_1, \ldots, \lambda_d$ be the ordered set of eigenvalues of the covariance matrix, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. If there exists a number m, such that $\lambda_i \gg \lambda_j$, $i = 1, \ldots, m$, and $j = m+1, \ldots, d$, then the data-set can be transformed to a coordinate system, where most of the signal variance is in an m-dimensional linear subspace spanned by the m'th first eigenvectors in the ordered list. This transformation is again given by the eigenvectors of the covariance matrix ($\mathbf{U}$),

$$\tilde{\mathbf{x}}_i = \mathbf{U}^T(\mathbf{x}_i - \widehat{\mathbf{x}}). \tag{16}$$

If we extract only the first m components of the transformed datavector $\tilde{\mathbf{x}}$ we obtain a signal that carries most of the variation of the original signal. Such reduction of the effective dimensionality of the problem is also known as extraction of features.

Checkpoint 2.4:

Use the programs `main2d.m` and `main2e.m` to transform 2D datasets into the eigenvector-space and comment on the "true" dimensionality of the classification problems.

DTU, August 2000,

Karam Sidaros