Continuous Random Variables

Machine Learning

Dr. Neeta Nain

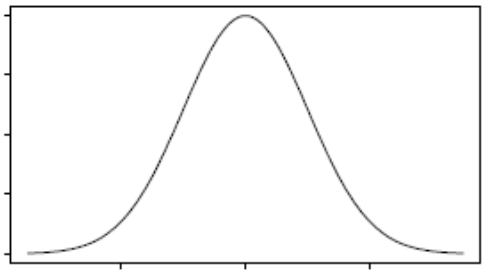
Department of Computer Science and Engineering
Malviya National Institute of Technology, Jaipur

25 July - 5 August 2016

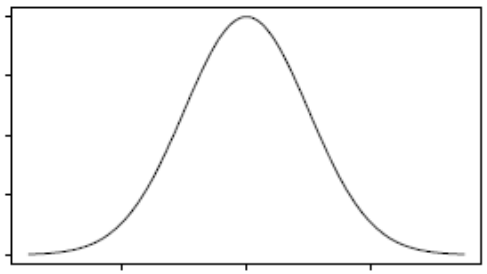
Outline

Random Variable

- A measurement has some uncertainity, can be likened to an experiment whose outcome is not known in advance.
- Random variable is the variable that associates a number with the outcome of a random experiment
- If a random variable can assume any value, not just certain discrete values, it is a continuous random variable, X
- The probability density function (PDF) of X is a function $f(x)$ such that the probability of X taking on value between $x = x_1$ and $x = x_2$ is $P < x_1 \leq x_2 = \int_a^b f(x)dx$



- The cumulative distribution function (CDF) is a function $F(x)$, of a random variable X, defined for a number x by $F(x) = P(X \leq x) = \int_0^x f(u)du$, i.e., for a number x, $F(x)$ is the probability that the observed value will be at most x.
- The mathematical relationship between the *PDF* and the *CDF* is $F(x) = \int_0^x f(s)ds$ where s is a dummy variable.
- Conversely $f(x) = -\frac{d(F(x))}{dx}$. The *CDF* is the integral of the *PDF* and, conversely, the *PDF* is the differential of the *CDF*



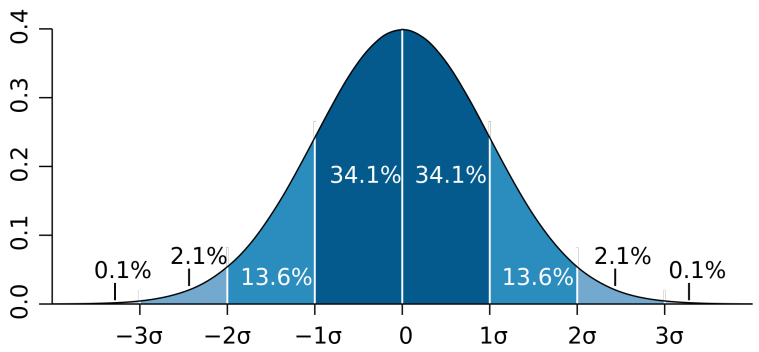
Gaussian Distribution

- An example of a *PDF* is the well known normal (or Gaussian) distribution, for which the *PDF* is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  where μ is the mean, or expected, value of X, denoted as $E[X]$
- The normal distribution is often referred as $N(\mu, \sigma^2)$.
- For a random variable it is the weighted average of all possible values that this random variable can take
- The weights correspond to the probability (p_i) in the case of a discrete random variable, or the probability densities (pdf's) in the case of continuous random variable, i.e.,
  $\mu(= E[X]) = \sum_{i=1}^{N} x_i p_i$ or $\mu(= E[X]) = \int_{-\infty}^{\infty} x f(x) dx$

- σ^2 (or Var[X]) is the variance of X, a measure of the spread of the distribution given by
  $\sigma^2 (= E[(X-\mu)^2]) = \sum_{i=1}^{N}(x_i-\mu)^2 p_i$ for discrete, and
  $\sigma^2 (= E[(X-\mu)^2]) = \int_{-\infty}^{\infty}(x_i-\mu)^2 f(x)dx$ for continuous random variables



- for *GD*, about 68% of the data values are within one *SD* of the mean, and about 95% are within two *SD*
- The *GD* is very convenient, since the parameters μ and σ are sufficient to uniquely characterize it
- Is the most widespread probability distribution used to describe measurements.
- It arises as the outcome of the Central Limit Theorem, which states that under mild conditions the sum of a large number of random variables will distribute approximately normally
- It has the maximum entropy (randomness) of all distributions having a given mean and variance

- It is the generalization of the $1D$ (univariate) GD to a higher number of dimensions, n
- A random variable is multivariate normally distributed if every linear combination of its components has a univariate normal distribution.
- Used often to describe, atleast approximately, any set of (possibly) correlated real-valued random variables each of which clusters around a mean value, defined as

$$f(X) = \frac{1}{((2\pi)^{n/2}|\Sigma|^{1/2})} exp(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu))$$

X, μ are now vectors of dimension n, and the variance (σ^2), has been replaced by the covariance matrix, Σ, not to be confused with a summation symbol, T indicates the transpose, viz., columns changed to rows

The Covariance Matrix

- The Covariance is a measure of the similarity between two random variables (X_1, X_2),i.e., how they vary with respect to each other
- If there is no similarity, they are independent
- The similarity could be so strong that knowing one determines the other without any uncertainity
- Similarity could be somewhere in between, where knowing one of the variables reduces the uncertainity about the other
- For a pair of random variables X_1 and X_2 it is defined as

$$\sigma_{12}^2 = Cov(X_1, X_2) = E[(X_1 - \mu_1).(X_2 - \mu_2)]$$

where μ_1, μ_2 are the respective means (expected values $E[X_1], E[X_2]$); when $X_1 = X_2$, covariance reduces to variance

The Multivariate Gaussian

- For multiple variables, the covaraince matrix, σ, is $n \times n$ matrix whose $(i,j)^{th}$ entry $Cov(X_i, X_j)$ is square and symmetric by definition
- For three random variables

$$\Sigma = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Cov(X_3, X_3) \end{bmatrix}$$

where the leading diagonal terms are variances σ_1^2, σ_2^2, and σ_3^2, and the off-diagonal terms are the covaraince between pairs of variables(features), σ_{ij}^2 or

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2}^2 & \sigma_{1,3}^2 \\ \sigma_{2,1}^2 & \sigma_2^2 & \sigma_{2,3}^2 \\ \sigma_{3,1}^2 & \sigma_{3,2}^2 & \sigma_3^2 \end{bmatrix}$$

Subtle point

- The variance is the mean of the squares of deviations. However we do not know the mean of the population and use sample values to calculate $\bar{x}$, and then reuse the sample values to calculate the variance.
- This gives the result a bias, which can be removed by using $n-1$ rather than n in the division to get the mean of the squares of the deviations.
- $n-1$ is known in statics as the number of degrees of freedom. In practice, it is not going to make much of a difference if we divide by n or $n-1$, if n is large, say $n > 10$; for small n divide by $n-1$ to get the values of varaince and covariance.

Factorization of Covariance matrix

- The covariance matrix can be factorized as $\Sigma = \Gamma R \Gamma$

$$= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \cdots & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \sigma_n \end{bmatrix}$$

  Γ contains the scales of the features, and R - the correlation matrix retains the essential information of relationship between features

- Terms of the correlation matrix are the (Pearson) correlation coefficients between pairs of variables/features and are equal to the corresponding covariances scaled by the standard deviation

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sigma_i . \sigma_j} = \frac{\sigma_{ij}}{\sigma_i . \sigma_j}$$

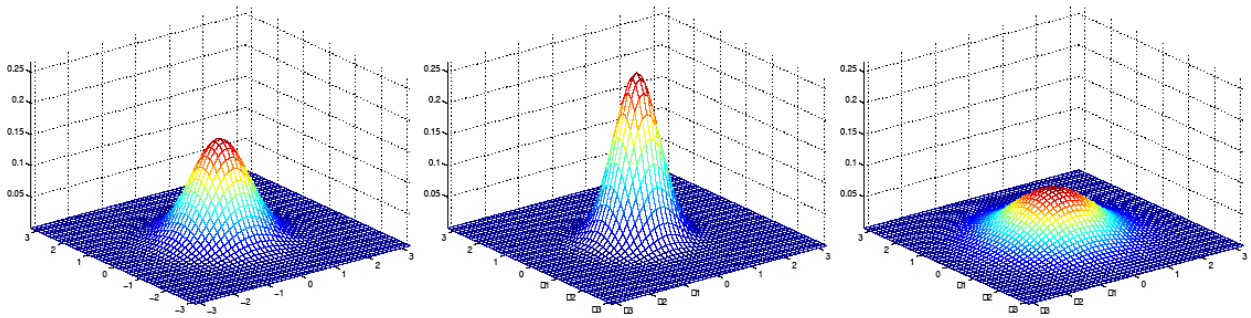
Covariance and Correlation

- Thus the correlation between two random variables is simply the covariance of the corresponding standardized random variables $\frac{(X-\mu)}{\sigma}$
- Both the covariance and the correlation (coefficient) describe the degree of similarity between two random variables and assume a linear link
- Correlation coefficient is dimensionless due to the scaling and assumes a value between -1 and $+1$
- The more dispersed the data points, the lower the value of the correlation coefficient
- Square of correlation coefficient, r^2, known as r-squared (coefficient of determination), gives the proportion of variance (fluctuation) of one variable that is predictable from the other

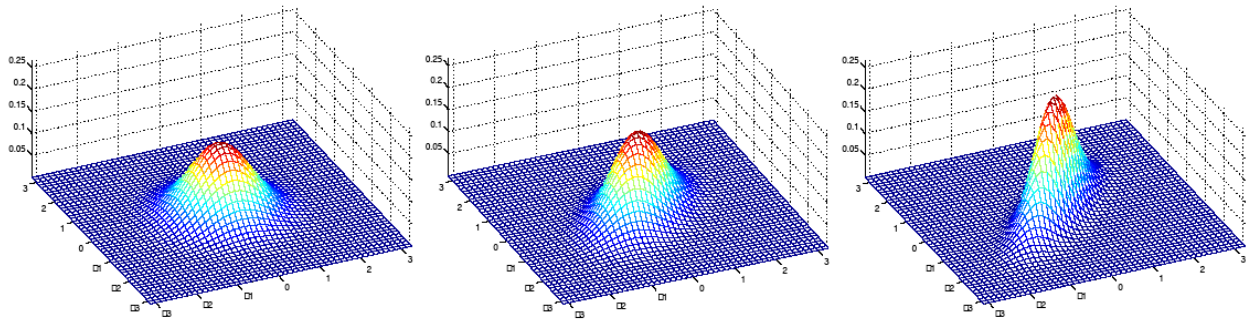
- If $r = 0.6$ between study hours and marks, then the coefficient of determination is 0.36, thus 36% of the marks are directly accounted for by the study hours and vice versa
- We have made a distinction between r, the correlation coefficient measured from a limited sample of X_i, Y_i pairs and ρ, the correlation that exists in the larger population between X and Y, in general
- It is possible, to obtain rather impressive-looking values of r within a sample, even when the correlation between X and Y is zero (when the sample size is small)
- We need to address the question of statistical significance of a given value of r for a particular size, N
- Statistical significance is conventionally set at the 5% level (an observed result is regarded as statistically significant only if it had a 5% or smaller likelihood of occuring by mere chance, otherwise it is non significant)

Returning to 2D Gaussian

- If Σ is diagonal the variables are uncorrelated
- If Σ is a multiple of the identity matrix, then the isocontours of the Gaussian [slices through the Gaussian at varying heights/probabilities, centered on the mean value (μ_1, μ_2)] are circular
- Changing the values on the diagonal of the covariance matrix simultaneously gives a narrower or broader Gaussian as shown (a) $\Sigma = I$, (b) $\Sigma = 2I$, (c) $\Sigma = 0.6I$

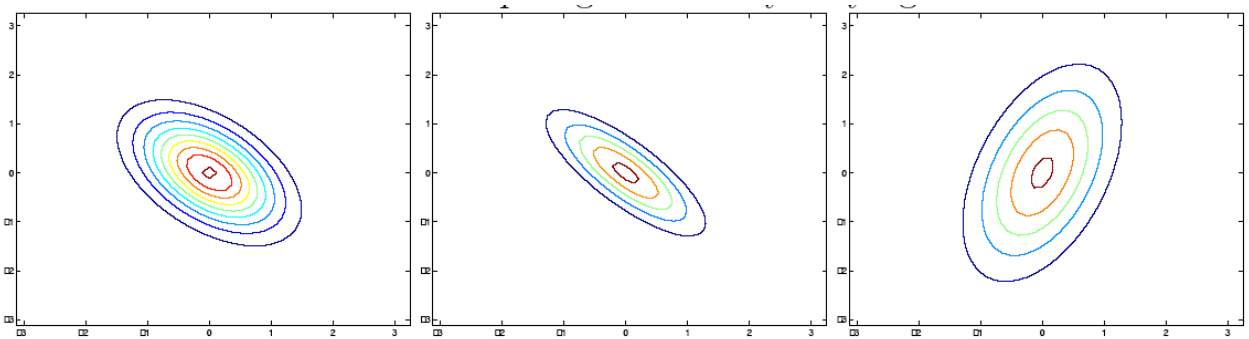


- If the values along the diagonal of the covariance matrix are not equal, then the isocontours of the Gaussian are elliptical, elongated along the feature axis with the larger variance
- If there are off-diagonal elements in the covaraince matrix, then the features are correlated. Increasingly large off-diagonal elements within Σ reflect increasing correlation between the variables (features) as shown (a) $\Sigma = I$, (b) $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, (c) $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$



Isocontours

- If there are off-diagonal elements in the covaraince matrix, the isocontours of the Gaussian are elliptical, but are not aligned with the feature axes, e.g, Isocontours of 2D Gaussian with (a) $\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$, (b) $\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$, (c) $\Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}$,



In PCA, the covariance matrix is diagonalized to remove the off-diagonal, correlated elements

Isocontours

- Consider the case of zero mean random vector with a diagonal covaraince matrix
- The isocontours are obtained by computing the curves of constant values for the exponent, that is

$$X^T \Sigma^{-1} X = [x_1, x_2] \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = C$$

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = C$$

This is the equation of an ellipse whose axes are determined by the variances of the respective features

For non-zero off diagonal elements in the covariance matrix, the ellipses are rotated relative to the feature axes, angle of rotation depends on eigenvectors/eigenvalues of the Σ

Eigenvalues and Eigenvectors

- Find the eigenvalues and eigenvectors of the matrix (real symmetric matrix, typical of covariance matrix) $\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$
- $Ax = \lambda x$, x is an eigenvector of A and λ its corresponding eigenvalue. We can re-write this as $Ax = \lambda Ix$, where I is the identity matrix, or as $(\lambda I - A)x = 0$
- For λ to be an eigenvalue, there must be a $\neq 0$ solution of this, which occurs when $\det(\lambda I - A) = 0$

$$\det(\lambda I - A) = \begin{bmatrix} \lambda - 5 & -3 \\ -3 & \lambda - 5 \end{bmatrix} = 0$$

$\lambda^2 - 10\lambda + 16 = 0$; $(\lambda - 2)(\lambda - 8) = 0$; the eigenvalues of A are $\lambda_1 = 2$ and $\lambda_2 = 8$

Eigenvectors

- Substituting $\lambda_1 = 2$ into $Ax = \lambda x$ we get

$$\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

  which gives $3x_1 + 3x_2 = 0$ from which we deduce, $e_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. In a similar manner, for eigenvalue $\lambda_2 = 8$, we get $-3x_1 + 3x_2 = 0$ from which, $e_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

- The eigenvectors are orthogonal (which only happens for a real symmetric matrix) and in this case are rotated $\pi/4$ rad from the original axes.
- In the coordinate system of these new principal axes the isocontours will be ellipses and the ellipse corresponding to one standard deviation will be $\frac{u^2}{8^2} + \frac{v^2}{2^2} = 1$

Example

- Find the eigenvalues and eigenvectors of the matrix
$$\begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}$$
-
$$\det(A - \lambda I) = \begin{bmatrix} 0.8 - \lambda & .3 \\ .2 & .7 - \lambda \end{bmatrix} = 0$$

Example

- Find the eigenvalues and eigenvectors of the matrix $\begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}$
- $$\det(A - \lambda I) = \begin{bmatrix} 0.8 - \lambda & .3 \\ .2 & .7 - \lambda \end{bmatrix} = 0$$
- $\lambda^2 - \frac{3}{2}\lambda + \frac{1}{2} = 0$; $(\lambda - 1)(\lambda - \frac{1}{2}) = 0$; the eigenvalues of A are $\lambda_1 = 1$ and $\lambda_2 = \frac{1}{2}$

Example

- Find the eigenvalues and eigenvectors of the matrix $\begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}$
- $$\det(A - \lambda I) = \begin{bmatrix} 0.8 - \lambda & .3 \\ .2 & .7 - \lambda \end{bmatrix} = 0$$
- $\lambda^2 - \frac{3}{2}\lambda + \frac{1}{2} = 0$; $(\lambda - 1)(\lambda - \frac{1}{2}) = 0$; the eigenvalues of A are $\lambda_1 = 1$ and $\lambda_2 = \frac{1}{2}$
- $(A - I)x_1 = 0 \Longrightarrow Ax_1 = x_1$; the first eigenvector $= (0.3, 0.2)$
  $(A - \frac{1}{2}I)x_2 = 0 \Longrightarrow Ax_2 = \frac{1}{2}x_2$; the second eigenvector $= (1, -1)$

The Mahalanobis Distance

- The Mahalanobis distance is a generalization of Euclidean distance. It is an appropriate measure when variables have different scales and are correlated, but still are approximately Gaussian distributed.
- Mahalanobis distance between two objects, D, in feature space is given as

$$D_m(x, y) = sqrt((x - y)\Sigma^{-1}(x - y)^T)$$

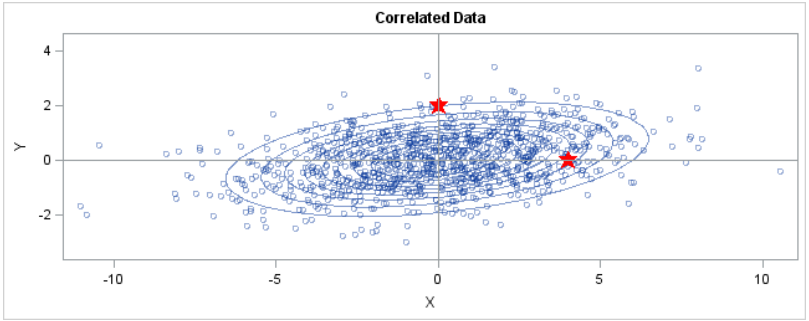
  where Σ^{-1} is the inverse of the covariance matrix
- It is common to work with square of Mahalanobis distance as it appears within the exponential of the multivariate Gaussian
- The isocontours of the multivariate Gaussian are ellipsoids of constant Mahalanobis distance

Estimating the class of a test data point

- For estimating the probability that a test data point in N-Dimensional feature space belongs to a labeled class
- Intuitively, the closer the point in question is to the center of mass μ, the more likely it is to belong to the class
- In statistics, we sometimes measure "nearness" or "farness" in terms of the sacle of the data. We need to know if the class is spread out over a large range or a small range
- Simplistic approach would be to estimate the width of the class distribution, i.e., σ of the distances of the sample points from μ. It could be made scale-independent or a dimensionless quantity by using Normalized distances, i.e., $\frac{x-\mu}{\sigma}$, called z-score of x, that is interpreted as the number of standard deviations that data is from the μ, however, this assumes that the sample points are distributed about the center of mass in a spherical manner

- This is not the case for a multivariate Gaussian distribution: the isocontours are elliptical (in 2D) or ellipsoidal (in 3D)
- The probability of the test point belonging to the class will depend not only on the distance from the center of the mass, but also on the direction
- Where the ellipsoid has a short axis, the test point must be closer, while in those where the axis is long the test point can be further away from the center
- The isocontours can be estimated from the covariance matrix
- The Mahalanobis distance is then the distance of the test point from μ, normalized by the width of the ellipsoid in the direction of the test point, it provides a way to measure distances that takes into account the scale of the data.

- Figure shows two points in feature space which are at the same Mahalanobis distance from the center of the distribution



The square of their Mahalanobis distance is given by

$$D_m^2(x, y) = (x - \mu)\Sigma^{-1}(x - \mu)^T$$

- If the covariance matrix is diagonal, the Mahalanobis distance reduces to the normal Euclidean distance
- If the covariance matrix is identity matrix, the Mahalanobis distance reduces to the (standard) Euclidean distance

Mahalanobis distance is widely used in supervised classification techniques, e.g., Fisher's Linear Discriminant Analysis (LDA) and in cluster analysis.

Exercise

- In a two class, two feature classification problem, the feature vectors are described by two normal distributions having same covariance $\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$; mean vectors: $\mu_1 = [0 \quad 0]^T, \mu_2 = [3 \quad 3]^T$. Classify vector $\mu_1 = [1.0 \quad 2.2]^T$

Exercise

- In a two class, two feature classification problem, the feature vectors are described by two normal distributions having same covariance $\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$; mean vectors: $\mu_1 = [0 \quad 0]^T, \mu_2 = [3 \quad 3]^T$. Classify vector $\mu_1 = [1.0 \quad 2.2]^T$
- Compute the Mahalanobis distance from the two means $D_m^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$

$$D_m^2(\mu_1, x) = [1.0 \quad 2.2] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$$

$$D_m^2(\mu_2, x) = [-2.0 \quad -0.8] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

The vector is assigned to class 1, since it is closer to μ_1

Exercise

- Compute the principal axes of the ellipse centered at $[0 \quad 0]^T$ that corresponds to a constant Mahalanobis distance $D_m = \sqrt{2.952}$ from the center
- Calculating the eigenvalues of Σ as

Exercise

- Compute the principal axes of the ellipse centered at $[0 \quad 0]^T$ that corresponds to a constant Mahalanobis distance $D_m = \sqrt{2.952}$ from the center
- Calculating the eigenvalues of Σ as

$$\det \begin{bmatrix} 1.1 - \lambda & 0.3 \\ 0.3 & 1.9 - \lambda \end{bmatrix} = \lambda^2 - 3\lambda + 2 = 0$$

gives $\lambda_1 = 1$ and $\lambda_2 = 2$

- The unit norm eigenvectors are

Exercise

- Compute the principal axes of the ellipse centered at $[0 \quad 0]^T$ that corresponds to a constant Mahalanobis distance $D_m = \sqrt{2.952}$ from the center
- Calculating the eigenvalues of Σ as

$$\det \begin{bmatrix} 1.1 - \lambda & 0.3 \\ 0.3 & 1.9 - \lambda \end{bmatrix} = \lambda^2 - 3\lambda + 2 = 0$$

gives $\lambda_1 = 1$ and $\lambda_2 = 2$

- The unit norm eigenvectors are $\begin{bmatrix} \frac{3}{\sqrt{10}} \\ \frac{-1}{\sqrt{10}} \end{bmatrix}$ and $\begin{bmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{bmatrix}$
- The eigenvectors give the axes of the isocontours and the half-lengths of the axes are proportional to the square roots of the corresponding eigenvalues.

- Mahalanobis distance can also be used to detect outliers, as samples that have a significantly greater Mahalanobis distance from the mean than the rest of the samples
- It can also be used to measure the separation between two classes, their dissimilarity, by measuring the distance between their respective centers
- Each class may have different number of samples
- For two classes with identical covariance matrices, the Mahalanobis distance is

$$D_m(1,2) = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$

- A Mahalanobis distance > 3 would indicate that the overlap between the two classes is small (two centers differ by more than three standard deviations)

Principal Component Analysis

- In projection methods we want to find a mapping from the original d-dim space to a new $k(< d)$-dim space, with minimum loss of information
- An optimal mapping would be one that results in no increase in the minimum probability of error
- In general, optimal mapping will require a nonlinear function, but we restrict ourselves to linear mapping, where two techniques are in common use: Principal Component Analysis (PCA) and Linear Discriminant Analysis(LDA)
- Principal component analysis (PCA) is a statistical technique that uses an orthogonal transformation to convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components

- PCA was invented in 1901 by Karl Pearson, it was later independently developed (and named) by Harold Hotelling in the 1930s.
- Depending on the field of application, it is also named the discrete Kosambi-KarhunenLove transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, singular value decomposition (SVD), eigenvalue decomposition (EVD) of XTX in linear algebra, etc
- The goal of the PCA is to represent data accurately in a lower dimensional space
- As much randomness (variance) in the higher dim space as possible should be preserved
- This is achieved by a transformation that centers the data and rotates the axes to line up with the direction of the highest variance

- Each principal component $(PC_1, PC_2, \cdots)$ is a linear combination of the original variables, there are as many PCs as original variables
- The first PC has as high a variance as possible, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (uncorrelated) the proceeding components
- Usually, the variance of the original data can be explained by the first few components and the rest can be ignored
- Using the PCs reduces the dimensionality of the data, making it more amenable to visual inspection, clustering, and pattern recognition efforts

- PCA can reveal the internal structure of the data in a way that best explains its variance, but it treats the total data and does not take into account class labels (it is unsupervised)
- There is no guarantee that the directions of maximum variance will make good features for discrimination
- The *PC*s are obtained by diagonalizing the covariance matrix of the original data
- Their directions and magnitudes are given by the eigenvectors and eigenvalues, respectively of the original covariance matrix

$$\begin{aligned} \Sigma \mathbf{x} &= \lambda I \mathbf{x} \\ (\Sigma - \lambda I)\mathbf{x} &= 0 \end{aligned}$$

where the λ's are the eigenvalues of Σ, and the $\mathbf{x}$'s are the corresponding eigenvectors

Solution of eigenvalue problem for an $n \times n$ matrix

- Compute the det $|A - \lambda I|$. With λ subtracted along the diagonal, this determinant starts with λ^n or $-\lambda^n$. It is a polynomial in λ of degree n.
- Find the roots of this polynomial, by solving det $|A - \lambda I| = 0$. The n roots are the n eigenvalues of A. They make $A - \lambda I$ singular.
- For each eigenvalue λ, solve $(A - \lambda I)x = 0$ to find for eigenvector x.
- Eigenvalues have their greatest importance in dynamic problems. Linear equations $Ax = b$ come from steady state problems. The solution $\frac{\partial \mathbf{u}}{\partial t} = A\mathbf{u}$ is changing with time - growing or decayoing or oscillating. We cannot find it by oscillations. We solve it by a new type of linear algebra, based on $A\mathbf{x} = \lambda\mathbf{x}$.
- Almost all vectors change direction, when they are multiplied by A - square matrix. Certain exceptional vectors $\mathbf{x}$ are in the same direction as $A\mathbf{x}$. Those are eigenvectors. Multiply an eigenvector $\mathbf{x}$ by A, the vector $A\mathbf{x}$ is a number λ times the original $\mathbf{x}$. The basic eqn is $A\mathbf{x} = \lambda\mathbf{x}$, where λ is the eigenvalue of A
- The eigenvalue tells whether the special vector $\mathbf{x}$ is stretched or shrunk or reversed or left unchanged

Exercise: PCA projection and recovery

- Find the principle components of the following dataset:
  $$\mathbf{x} = \begin{bmatrix} 1 & 3 & 3 & 3 & 3 & 4 & 4 & 5 & 5 & 6 & 8 & 9 \\ 2 & 3 & 5 & 6 & 7 & 4 & 5 & 4 & 6 & 5 & 7 & 8 \end{bmatrix}$$
- $\bar{\mathbf{x}} = [4.5, 5.1]$
- Zero mean data: $\mathbf{x}_s = \mathbf{x} - \bar{\mathbf{x}}$

Exercise: PCA projection and recovery

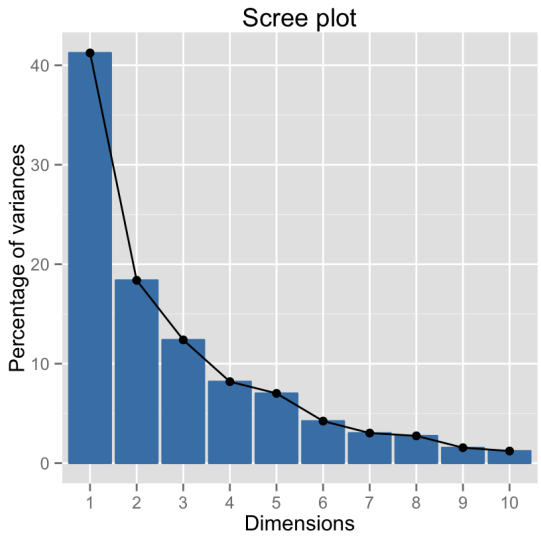
- Find the principle components of the following dataset:
  $\mathbf{x} = \begin{bmatrix} 1 & 3 & 3 & 3 & 3 & 4 & 4 & 5 & 5 & 6 & 8 & 9 \\ 2 & 3 & 5 & 6 & 7 & 4 & 5 & 4 & 6 & 5 & 7 & 8 \end{bmatrix}$
- $\bar{\mathbf{x}} = [4.5, 5.1]$
- Zero mean data: $\mathbf{x}_s = \mathbf{x} - \bar{\mathbf{x}}$
- The covariance matrix is $\Sigma = \begin{bmatrix} 5.1818 & 2.7273 \\ 2.7273 & 3.0606 \end{bmatrix}$

Exercise: PCA projection and recovery

- Find the principle components of the following dataset:
  $\mathbf{x} = \begin{bmatrix} 1 & 3 & 3 & 3 & 3 & 4 & 4 & 5 & 5 & 6 & 8 & 9 \\ 2 & 3 & 5 & 6 & 7 & 4 & 5 & 4 & 6 & 5 & 7 & 8 \end{bmatrix}$
- $\bar{\mathbf{x}} = [4.5, 5.1]$
- Zero mean data: $\mathbf{x}_s = \mathbf{x} - \bar{\mathbf{x}}$
- The covariance matrix is $\Sigma = \begin{bmatrix} 5.1818 & 2.7273 \\ 2.7273 & 3.0606 \end{bmatrix}$
- Eigenvectors are $\mathbf{v}_1 = \begin{bmatrix} 0.5646 \\ -0.8254 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} -0.8254 \\ -0.5646 \end{bmatrix}$
- Project data: $\mathbf{y} = \mathbf{v}\mathbf{x}_s =$
  $\begin{bmatrix} .6375 & .9414 & -0.7093 & -1.5347 & -2.3601 & 0.6806 & -0.1447 & 1.2452 & -0.4055 & 0.9845 \\ 4.667 & 2.4614 & 1.3321 & 0.7675 & 0.2029 & 1.0714 & 0.5068 & 0.246 & -0.8832 & -1.1439 \end{bmatrix}$
- $\mathbf{x}_s = \mathbf{v}^{-1}\mathbf{y}$
- $\mathbf{x} = \mathbf{x}_s + \bar{\mathbf{x}}$

- If we order the eigenvalues in descending order, then the first eigenvector will be the directional cosines of the first PC axis and the first eigenvalue will be the variance along this axis, and so on
- Total variance is conserved:
  $\sum_{n=1}^{N} \lambda_n = \text{trace}(\Sigma) = a_{11} + a_{22} + \cdots + a_{nn}$
- The contribution of any eigenvalue, λ_i, to the total variance is $\frac{\lambda_i}{trace(\Sigma)}$
- PCA only helps if the original variables are correlated
- If they are highly correlated, there will be a small number of eigenvectors with large eigenvalues; a large reduction in dimensionality can be obtained by keeping only the k largest PCs

- The % of the total (var) accounted for by each of the PCs is proportional to the value of the individual eigenvalues
- Eigenvalues plotted vs the no. of PCs is the scree plot; useful in determining no. of PCs needed to capture most of the data



- The first few PCs captures $> 90\%$ of the variation, keeping only them will perserve most of the data variability and reduce the dimensionality

- PCA is sensitive to outliers (which should be discarded on the basis of their Mahalanobis distances to the centroids)
- PCA is a one class procedure and therefore cannot help in separating classes
- The Karhunen-Loeve expansion allows the use of class information; instead of using the covariance matrix of the whole sample, it estimates separate class covariance matrices, take their average (weighted by their priors) and use its eigenvectors
- PCA is limited to finding linear combinations of the original features, which is sufficient in many applications but may result in loss of too much information
- To retain such information, a nonlinear mapping method such as multi-dimensional scaling (MDS) is needed

- λ is an eigenvalue of A if and only if $(A - \lambda I)$ is singular; $\det(A - \lambda I) = 0$
- The eigenvalue λ could be zero! Then $A\mathbf{x} = 0\mathbf{x} = 0$ means that this eigenvector is in the nullspace
- If A is invertible, zero is not an eigenvalue
- If $A = I$, every vector has $A\mathbf{x} = \mathbf{x}$. All vectors are eigenvectors of I, where all $\lambda = 1$
- If each column of A adds to 1, then $\lambda = 1$ is an eigenvalue
- If A is singular, then $\lambda = 0$ is an eigenvalue
- If A is symmetric, then its eigenvectors are perpendicular
- Product of n eigenvalues equals the determinant: $\prod \lambda_n = \det |A|$
- Sum of the n eigenvalues equals the sum of the n diagonal entries: $\sum \lambda_n = \sum A_{nn}$

PCA

- If we are projecting from M to D dim, PCA will define D vectors, $\mathbf{w}_d$, each of which is N-dim. The d^{th} element of the projection, y_{nd}(where $\mathbf{y}_n = [y_{n1}, \cdots, y_{nD}]^T$), is computed as $y_{nd} = \mathbf{w}_d^T \mathbf{x}_n$
- The learning task is therefore to choose how many dims we want to project into (D) and then pick the projection vector, $\mathbf{w}_d$, for each
- PCA uses variance in the projected space as the criteria to choose $\mathbf{w}_d$, in particular $\mathbf{w}_1$ will be the projection that makes the variance in the y_{n1} as high as possible
- The second projection dim $\mathbf{w}_2$ is also chosen to maximize the variance, but they both should be orthogonal or $\mathbf{w}_1^T \mathbf{w}_2 = 0$
- In general $\mathbf{w}_i^T \mathbf{w}_j = 0, \forall j \neq i$

- PCA imposes the constraint that each $\mathbf{w}_i$ must have a unit length or $\mathbf{w}_i^T \mathbf{w}_i = 1$. It is only the dir that is important
- Assume each of the original dims have zero mean $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} x_n = 0$. This can be enforced by subtracting mean from the data $(\mathbf{x} - \bar{\mathbf{x}})$
- Start by finding a proj into $D = 1$ dim. In this case, the projection results in a scalar value, y_n, for each observation given by $y_n = \mathbf{w}^T \mathbf{x}_n$
- The variance, $\sigma_y^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \bar{y})^2$.
- Simplifying, $(\bar{\mathbf{x}} = 0) : \bar{y} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{w}^T \mathbf{x}_n$
  $= \mathbf{w}^T (\frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n) \quad \Longrightarrow \mathbf{w}^T \bar{\mathbf{x}} = 0$
- $\sigma_y^2 = \frac{1}{N} \sum_{n=1}^{N} y_n^2$

- Substituting y_n, $\sigma_y^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n)^2$
  $= \frac{1}{N}\sum_{n=1}^{N}\mathbf{w}^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{w} = \mathbf{w}^T(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T)\mathbf{w}$
- $\sigma_y^2 = \mathbf{w}^T\mathbf{C}\mathbf{w}$, where $\mathbf{C}$ is the sample covariance matrix defined as $C = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$, where $\bar{\mathbf{x}} = 0$.
- We didnot lose anything by transforming our data to zero mean. C would be the same whether we did this or not
- Our aim is to find the value of $\mathbf{w}$ that maximizes σ^2 and therefore also maximises $\mathbf{w}^T\mathbf{C}\mathbf{w}$

- We would keep increasing $\mathbf{w}^T C\mathbf{w}$ by increasing the values of the elements in $\mathbf{w}$ and this is why length of $\mathbf{w}$ is constrained to $\mathbf{w}^T\mathbf{w} = 1$.
- We incorporate this constraint into our optimization using Lagrangian term. We wish to find the $\mathbf{w}$ that maximises $L = \mathbf{w}^T\mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1)$
- Taking partial der w.r.t $\mathbf{w}$, equating to zero
  $\frac{\partial L}{\partial \mathbf{w}} = \mathbf{C}\mathbf{w} - \lambda\mathbf{w} = 0$
  $\mathbf{C}\mathbf{w} = \lambda\mathbf{w}$, this is a very common form, known as eigenvector/eigenvalue equation

- Comparing this with the $\Sigma v = \lambda v$, we can see that the projection $\mathbf{w}$ that maximizes the variance is one of the eigenvectors of the covariance matric $\mathbf{C}$. There will be M of these. How do we know which one corresponds to the heighest variance?
- We have $\sigma_y^2 = \mathbf{w}^T \mathbf{C} \mathbf{w}$, using $\mathbf{w}^T \mathbf{w} = 1$, multiplying l.h.s by this we get
  $\sigma^2 \mathbf{w}^T \mathbf{w} = \mathbf{w}^T \mathbf{C} \mathbf{w}$, removing $\mathbf{w}^T$ from each side leaves us
  $\sigma^2 \mathbf{w} = \mathbf{C} \mathbf{w}$
- Given an eigenvector/eigenvalue pair $(\lambda, \mathbf{w}), \lambda$ corresponds to the variance of the data in the projected space defined by $\mathbf{w}$
- If we find M eigenvector/eigenvalues pairs of the covariance matrix $\mathbf{C}$, the pair with the highest eigenvalue corresponds to the projection with maximal variance, $\mathbf{w}_1$. The second highest eigenvalue corresponds to $\mathbf{w}_2$, the third to $\mathbf{w}_3$ and so on

References

- Pattern Recognition and Classification, Geoff Dougherty, Springer.
- Pattern Recognition and Image Analysis, Earl Gose, Richard Johnsonbaugh, Steve Jost, PHI.
- Machine Learning, Tom M. Mitchell, Mc Graw Hill.
- Pattern Classification, Duda, Hart, Stork, Wiley.
- Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp- 4-37, January 2000.