Data Mining and Machine Learning

Statistical Modelling (1)

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

- Review basic statistical modelling
- Review the notions of probability distribution and probability density function (PDF)
- Gaussian PDF
- Multivariate Gaussian PDFs
- Parameter estimation for Gaussian PDFs

Discrete random variables

- Suppose that Y is a *random variable* which can take any value in a discrete set $X=\{x_1,x_2,\ldots,x_M\}$
- Suppose that $y_1,y_2,\ldots,y_T$ are samples of the random variable Y
- If c_m is the number of times that the $y_n = x_m$ then an estimate of the probability that y_n takes the value x_m is given by:

$$P(x_m) = P(y_n = x_m) \approx \frac{c_m}{N}$$

Continuous Random Variables

- In most practical applications the data are not restricted to a finite set of values – they can take any value in *N*-dimensional space
- Counting the number of occurrences of each value is no longer a viable way of estimating probabilities…
- …but generalisations of this approach are applicable to continuous variables – <u>non-parametric methods</u>

Continuous Random Variables

- An alternative is to use a parametric model
- Probabilities are defined by a small set of parameters

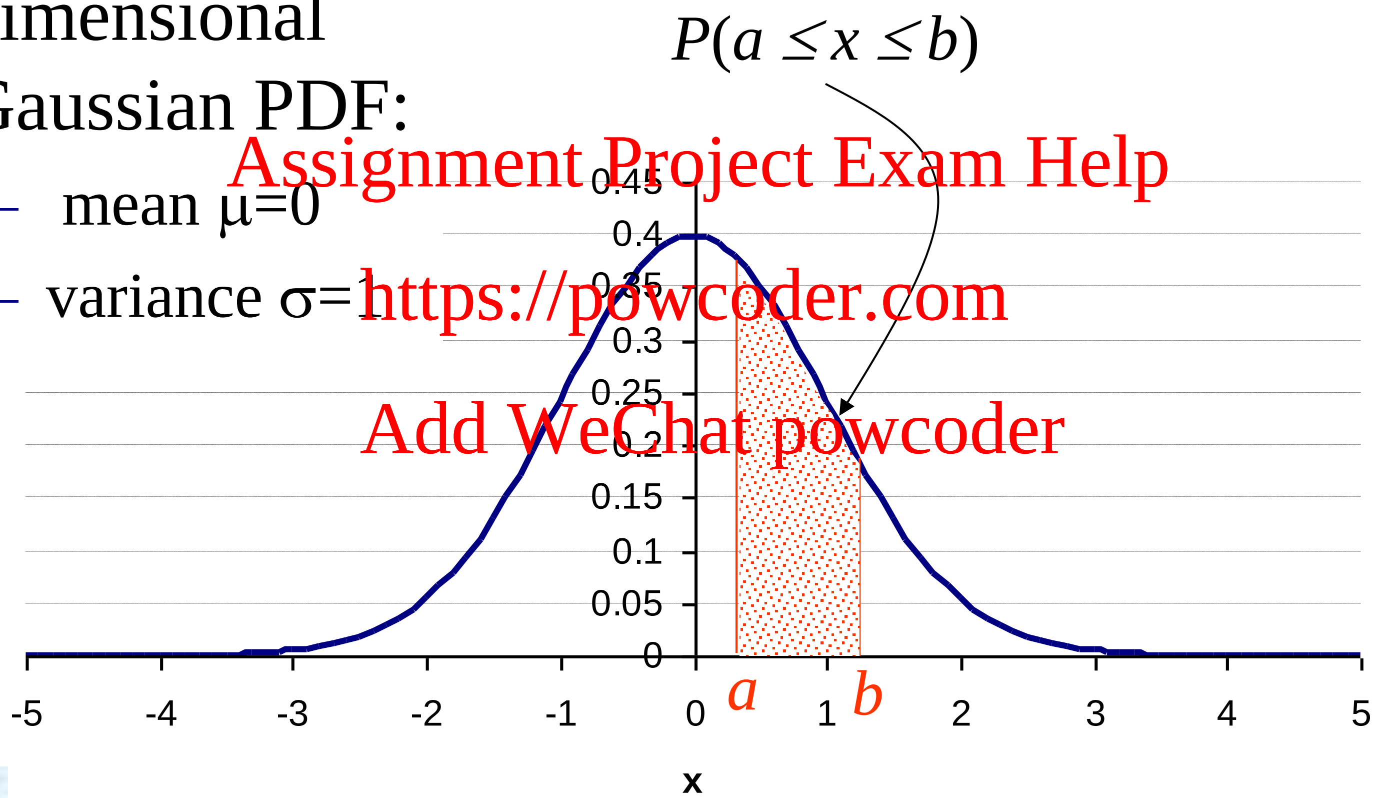
- Familiar example is a normal, or Gaussian model

- A (scalar/univariate) Gaussian probability density function (PDF) is defined by two parameters – its mean μ and variance σ

- For a multivariate Gaussian PDF defined on a vector space, μ is the mean vector and σ is the covariance matrix

Gaussian PDF

- 'Standard' 1-dimensional Gaussian PDF:
  - mean $\mu=0$
  - variance $\sigma=1$



UNIVERSITY OF BIRMINGHAM

Gaussian PDF

- For a 1-dimensional Gaussian PDF *p* with mean μ and variance σ:

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

Constant to ensure area under curve is 1

Defines 'bell' shape

Standard Deviation

- Standard deviation is the square root of the variance
- For a Gaussian PDF:
  - 68% of the area under the curve lies within one standard deviation (s.d.) of the mean
  - 95% of the area under the curve lies within two s.ds of the mean
  - 99% of the area under the curve lies within three standard deviations of the mean

Standard Deviation

- In other words, if $s = \sqrt{\sigma}$ then:

$$P(\mu - s \le x \le \mu + s) = 0.68$$

$$P(\mu - 2s \le x \le \mu + 2s) = 0.95$$

$$P(\mu - 3s \le x \le \mu + 3s) = 0.99$$

Multivariate Gaussian PDFs

- A (univariate) Gaussian PDF assumes the random variable takes <u>scalar</u> values
- In the case where the random variable takes N dimensional <u>vector</u> values the corresponding PDF is called a <u>multivariate</u> Gaussian PDF and is given by:

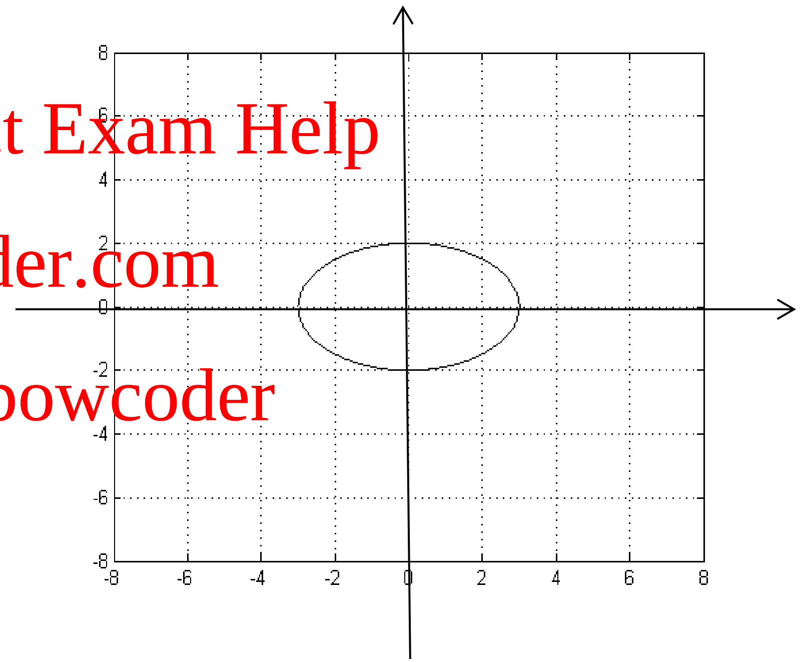
$$p(x) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(\frac{-1}{2} (m - x)^T \Sigma^{-1} (m - x) \right)$$

Visualising multivariate Gaussian PDFs

- It's easy to sketch a 1 dimensional Gaussian PDF, using the rules about the proportion of the area that lies within 1, 2 and 3 standard deviations of the mean and the value for $p(\mu)$.
- 2D Gaussian PDFs can be plotted using MATLAB's 3D plotting functions
- A simpler way to visualise a 2D Gaussian PDF is to plot the <u>1 standard deviation contour</u>. This is the set of points that lie 1 standard deviation from the mean

Example

- If $\Sigma = \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix}$, standard deviations in 'x' and 'y' directions are 3 and 2, respectively, and the 1 s.d. contour is an ellipse:



Example 2:

- Now suppose $\Sigma = \begin{bmatrix} 7.75 & 2.17 \\ 2.17 & 5.25 \end{bmatrix}$ and $m = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$

- Calculate the eigenvalue decomposition of Σ

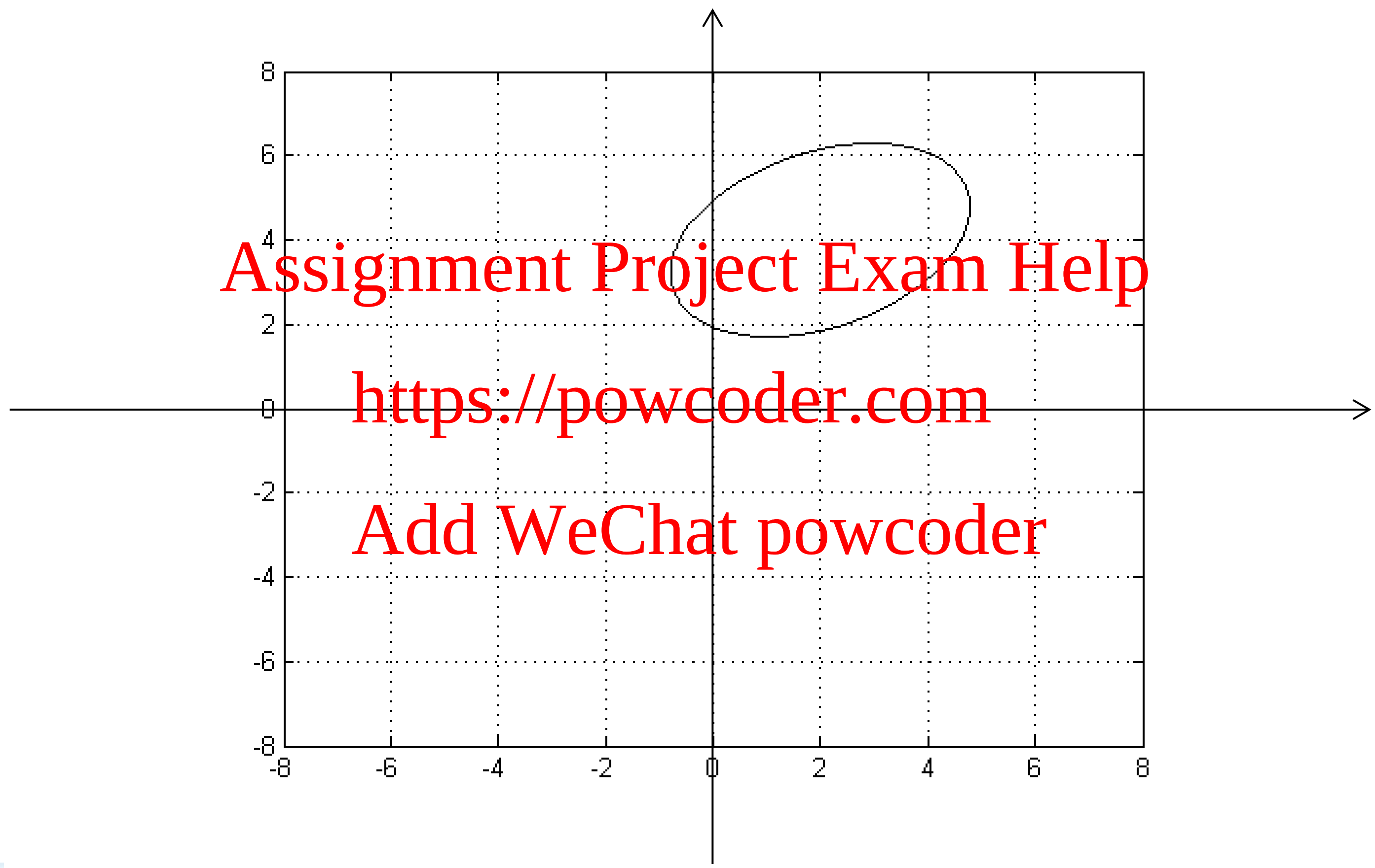
$$\Sigma = UDU^T = \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{-1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}$$

Example 2 (continued)

- Note U is a rotation through 30^o
- Hence the one standard deviation contour is the same as in the previous example, but rotated through 30^o and translated by

$$m = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

Example 2 (continued)



Fitting a Gaussian PDF to Data

- Suppose $y = y_1, \ldots, y_t, \ldots, y_T$ is a set of T data values
- For a Gaussian PDF p with mean μ and variance σ, define:

$$p(y \mid \mu,\sigma) = \prod_{t=1}^{T} p(y_t \mid \mu,\sigma)$$

- How do we choose μ and σ to maximise $p(y|\mu, \sigma)$?

Fitting a Gaussian PDF to Data

Maximum Likelihood Estimation

- The ‘best fitting’ Gaussian maximises $p(y|\mu,\sigma)$
- Terminology:
  - $p(y|\mu,\sigma)$, as a function of y is the <u>probability (density)</u> of y
  - $p(y|\mu,\sigma)$, a function of μ,σ is the <u>likelihood</u> of μ,σ
- Maximising $p(y|\mu,\sigma)$ with respect to μ,σ is called <u>Maximum Likelihood (ML)</u> estimation of μ,σ

ML estimation of μ,σ

- Intuitively:
  – The maximum likelihood estimate of μ should be the average value of $y_1,\ldots,y_T$. (the <u>sample mean</u>)
  – The maximum likelihood estimate of σ should be the variance of $y_1,\ldots,y_T$. (the <u>sample variance</u>)
- This is true: $p(y|\,\mu,\,\sigma)$ is maximised by setting:

$$\mu = \frac{1}{T}\sum_{t=1}^{T} y_t, \quad \sigma = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \mu\right)^2$$

UNIVERSITY OF BIRMINGHAM

Proof

First note that maximising $p(y)$ is the same as maximising $\log(p(y))$

$$\log p(y \mid \mu, \sigma) = \log \prod_{t=1}^{T} p(y_t \mid \mu, \sigma) = \sum_{t=1}^{T} \log p(y_t \mid \mu, \sigma)$$

Also

$$\log p(y_t \mid \mu, \sigma) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(\mu - y_t)^2}{\sigma}$$

At a maximum:

$$0 = \frac{\partial}{\partial \mu}\log p(y \mid \mu, \sigma) = \sum_{t=1}^{T} \frac{\partial}{\partial \mu}\log p(y_t \mid \mu, \sigma) = \sum_{t=1}^{T} \frac{-2(\mu - y_t)(-1)}{\sigma}$$

So, $T\mu = \sum_{t=1}^{T} y_t, \mu = \frac{1}{T}\sum_{t=1}^{T} y_t$

Multi-modal distributions

- In practice the distributions of many naturally occurring phenomena do not follow the simple bell-shaped Gaussian curve
- For example, if the data arises from several difference sources, there may be several distinct peaks (e.g. distribution of heights of adults)
- These peaks are the modes of the distribution and the distribution is called multi-modal

Summary

- Reviewed basic statistical modelling, probability distribution, probability density function
- Gaussian PDFs
- Multivariate Gaussian PDFs
- Maximum likelihood (ML) parameter estimation

- In the next session we will introduce Gaussian mixture PDFs (GMMs) and ML parameter estimation for GMMs