Inf2b Learning and Data

Lecture 8: Real-valued distributions and Gaussians

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

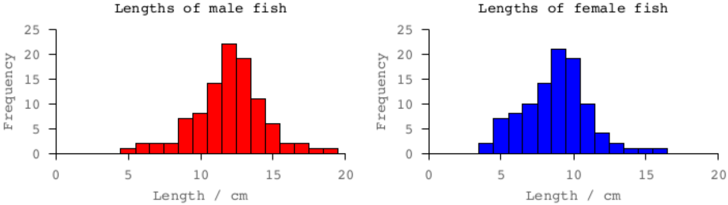
Jan-Mar 2014

Today's Schedule

1. Continuous random variables
2. The Gaussian distribution (one-dimensional)
3. The multidimensional Gaussian distribution

Discrete to continuous random variables

Fish example again:

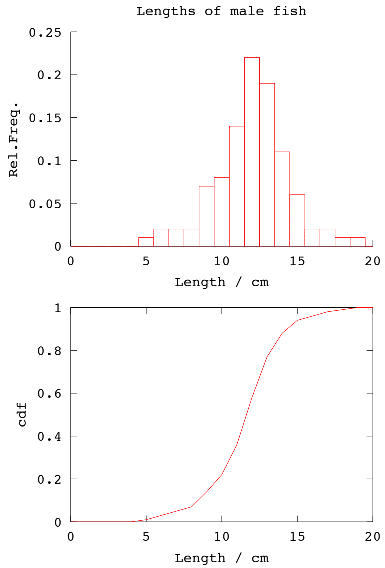


$$c^* = \arg\max_c P(c|x) = \arg\max_c \frac{P(x|c)P(c)}{P(x)} = \arg\max_c P(x|c)P(c)$$

- What if the number of bins $\to \infty$? (i.e. the width of bin $\to 0$)
- $P(X = x|C)$ will be almost 0 everywhere!
- We instead consider a cumulative distribution function (cdf) with a continuous random variable:

$$F(x) = P(X \leq x)$$

Cumulative distribution functions graphed



Cumulative distribution function properties

Cumulative distribution functions have the following properties:

1. $F(-\infty) = 0$;
2. $F(\infty) = 1$;
3. If $a \leq b$ then $F(a) \leq F(b)$.

To obtain the probability of falling in an interval we can do the following:

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a) \end{aligned}$$

Probability density function (pdf)

- The rate of change of the cdf gives us the probability density function (pdf) , $p(x)$:

$$p(x) = \frac{d}{dx} F(x) = F'(x)$$

$$F(x) = \int_{-\infty}^{x} p(x)\, dx$$

- $p(x)$ is **not** the probability that X has value x. But the pdf is proportional to the probability that X lies in a small interval $[x, x + dx]$.
- Notation: p for pdf, P for probability

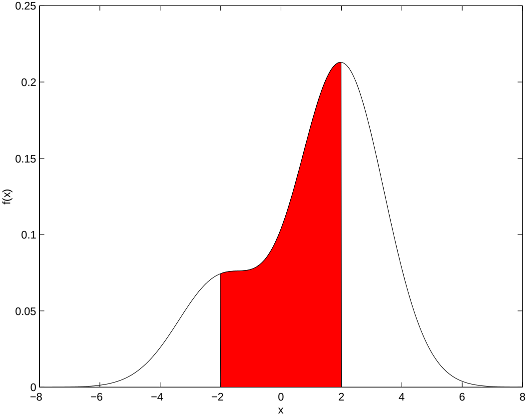
pdf and cdf

The probability that the random variable lies in interval (a, b) is given by:

$$
\begin{aligned}
P(a < X \leq b) &= F(b) - F(a) \\
&= \int_{-\infty}^{b} p(x)\, dx - \int_{-\infty}^{a} p(x)\, dx \\
&= \int_{a}^{b} p(x)\, dx
\end{aligned}
$$

pdf and cdf

The probability that the random variable lies in interval (a, b) is the area under the pdf between a and b:



The Gaussian distribution

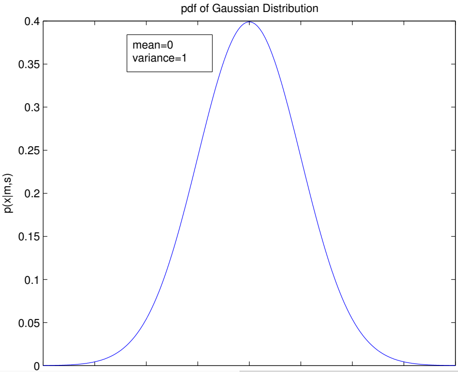
- The Gaussian (or Normal) distribution is the most common (and easily analysed) continuous distribution
- It is also a reasonable model in many situations (the famous bell curve)
- If a (scalar) variable has a Gaussian distribution, then it has a probability density function with this form:

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

- The Gaussian is described by two parameters:
  - the mean μ (location)
  - the variance σ^2 (dispersion)

Plot of Gaussian distribution

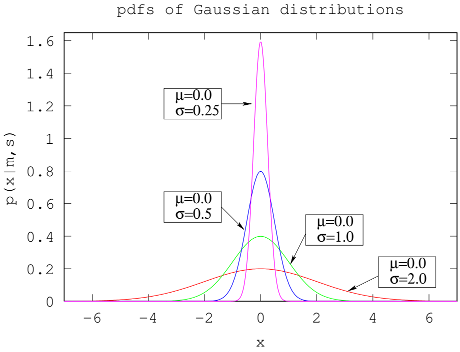
- Gaussians have the same shape, with the location controlled by the mean, and the spread controlled by the variance
- One-dimensional Gaussian with zero mean and unit variance
  $(\mu = 0, \sigma^2 = 1)$



Another plot of a Gaussian

Properties of the Gaussian distribution

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



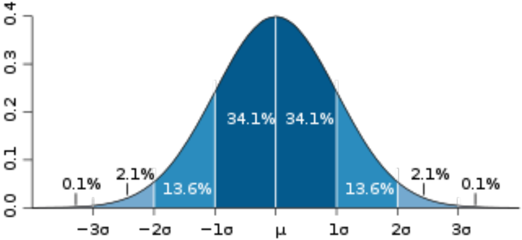
$$\int_{-\infty}^{\infty} N(x; \mu, \sigma^2) dx = 1$$

$$\lim_{\sigma \to 0} N(x; \mu, \sigma^2) = \delta(x - \mu)$$

(Dirac delta function)

Facts about the Gaussian distribution

- A Gaussian can be used to describe approximately any random variable that tends to cluster around the mean
- Concentration:
  - About 68% of values drawn from a normal distribution are within one SD away from the mean
  - About 95% are within two SDs
  - About 99.7% lie within three SDs of the mean



Central Limit Theorem

- Under certain conditions, the sum of a large number of random variables will have approximately normal distribution.
- Several other distributions are well approximated by the Normal distribution:
    - Binomial $B(n, p)$, when n is large and p is not too close to 1 or 0
    - Poisson $P_o(\lambda)$ when λ is large
    - Other distributions including chi-squared and Students T
- The Wikipedia entry on the Gaussian distribution is good

Parameter estimation

- Estimate mean and variance parameters of a Gaussian from data $x^{(1)}, x^{(2)}, \dots, x^{(n)}$
- Use sample mean and sample variance estimates:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} \qquad \text{(sample mean)}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu)^2 \qquad \text{(sample variance)},$$

Example: Gaussians

A pattern recognition problem has two classes, S and T. Some observations are available for each class:

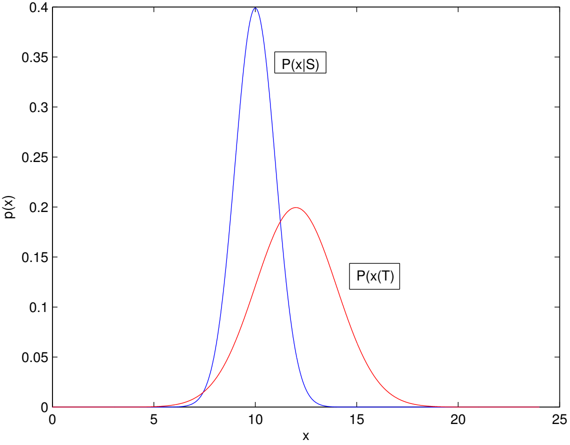
Class S	10	8	10	10	11	11
Class T	12	9	15	10	13	13

The mean and variance of each pdf are estimated by the sample mean and sample variance.

$$\begin{array}{lll} S: & \text{mean} = 10; & \text{variance} = 1 \\ T: & \text{mean} = 12; & \text{variance} = 4 \end{array}$$

Example: pdfs

Sketch the pdf for each class.



Summary of one-dimensional Gaussians

Gaussians

- Continuous random variable: cumulative distribution function (cdf) and probability density function (pdf)
- Gaussian pdf (one dimension):

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

- Estimate parameters (mean and variance) using maximum likelihood estimation (See Tutorial 8)

The multidimensional Gaussian distribution

- The d-dimensional vector $\mathbf{x} = (x_1, \ldots, x_d)^T$ is multivariate Gaussian if it has a probability density function of the following form:

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

  The pdf is parameterised by the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.
- The 1-dimensional Gaussian is a special case of this pdf
- The argument to the exponential $\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is referred to as a*quadratic form*.

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

- The covariance matrix $\vec{\Sigma}$ is the expectation of the deviation of $\mathbf{x}$ from the mean:

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

- $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric matrix:

$$\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = E[(x_j - \mu_j)(x_i - \mu_i)] = \sigma_{ji}\,.$$

- The sign of the covariance helps to determine the relationship between two components:
    - If x_j is large when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be positive;
    - If x_j is small when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be negative.

Parameter estimation

Maximum likelihood estimation (MLE):

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}^{(i)}$$

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^T$$

Correlation matrix

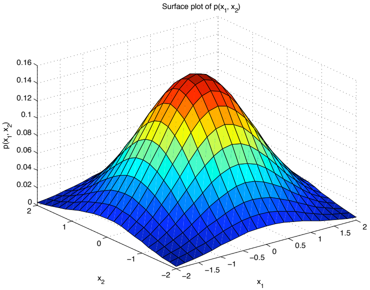
The covariance matrix is not scale-independent: Define the correlation coefficient:

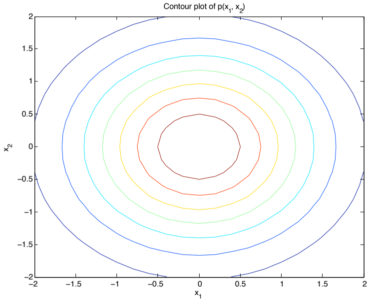
$$\rho(x_i, x_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

- Scale-independent (ie independent of the measurement units) and location-independent, ie:
  $$\rho(x_i, x_j) = \rho(ax_i + b, cx_j + d)$$
- The correlation coefficient satisfies $-1 \leq \rho \leq 1$, and

$$\begin{aligned} \rho(x, y) &= +1 \qquad \text{if } y = ax + b \quad a > 0 \\ \rho(x, y) &= -1 \qquad \text{if } y = ax + b \quad a < 0 \end{aligned}$$

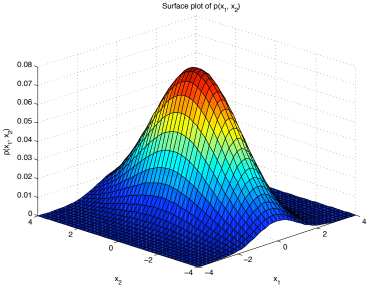
Spherical Gaussian

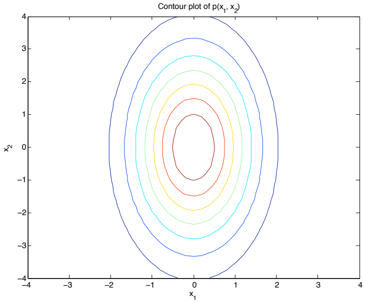




$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \rho_{12} = 0$$

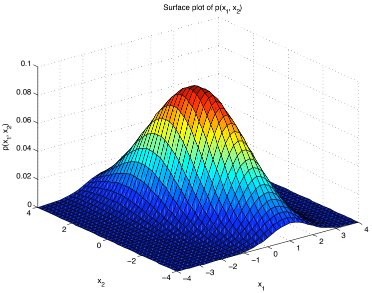
2-dimensional Gaussian with a diagonal covariance matrix

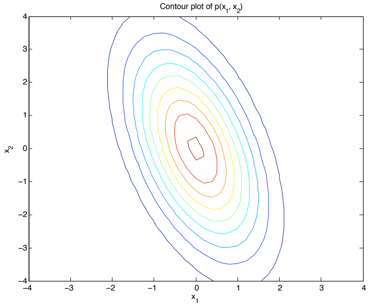




$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \qquad \rho_{12} = 0$$

2-dimensional Gaussian with a full covariance matrix





$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \qquad \rho_{12} = -0.5$$

Practical issues

Parameter estimation of multivariate Gaussian distribution can be difficult

Summary

Gaussians

- Continuous random variable: cumulative distribution function and probability density function
- Univariate Gaussian pdf:

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

- Multivariate Gaussian pdf:

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

- Estimate parameters (mean and covariance matrix) using maximum likelihood estimation