Today’s Schedule

Real-valued distributions and Gaussians

1. Continuous random variables
2. The Gaussian distribution (one-dimensional)
3. Maximum likelihood estimation
4. 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

Lengths of male fish

Lengths of female fish

Real-valued distributions and Gaussians
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:

\[
P(a < X \leq b) = P(X \leq b) - P(X \leq a)
\]
\[
= F(b) - F(a)
\]
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.
The probability that the random variable lies in interval \((a, b)\) is given by:

\[
P(a < X \leq b) = F(b) - F(a)
\]

\[
= \int_{-\infty}^{b} p(x) \, dx - \int_{-\infty}^{a} p(x) \, dx
\]

\[
= \int_{a}^{b} p(x) \, dx
\]
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:

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

NB: \( \exp(f(x)) = e^{f(x)} \)

- The Gaussian is described by two parameters:
  - the **mean** \( \mu \) (location)
  - the **variance** \( \sigma^2 \) (dispersion)
Natural exponential function

\[ y = e^x = \exp(x) \]

\[ y = \exp(-x^2) \]
- 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 ($\mu = 0, \sigma^2 = 1$)
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 $\lambda$ is large
  - Other distributions including chi-squared and Students $T$
- The Wikipedia entry on the Gaussian distribution is good
Parameter estimation form data

- Estimate the mean and variance parameters of a Gaussian from data \( \{x_1, x_2, \ldots, x_N\} \)

- **Sample mean and sample variance (unbiased) estimates:**

  \[
  \hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n \\
  \hat{\sigma}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \hat{\mu})^2
  \]

- **Maximum likelihood estimates (MLE):**

  \[
  \hat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \\
  \hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu}_{ML})^2
  \]
Example: Gaussians

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

<table>
<thead>
<tr>
<th>Class</th>
<th>$S$</th>
<th>10</th>
<th>8</th>
<th>10</th>
<th>10</th>
<th>11</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>$T$</td>
<td>12</td>
<td>9</td>
<td>15</td>
<td>10</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

The mean and variance of each pdf are estimated with MLE.

$S$ : mean = 10; variance = 1

$T$ : mean = 12; variance = 4

\[
p(x|S) = \frac{1}{\sqrt{2\pi} \cdot 1} \exp\left(-\frac{(x-10)^2}{2 \cdot 1}\right)
\]

\[
p(x|T) = \frac{1}{\sqrt{2\pi} \cdot 4} \exp\left(-\frac{(x-12)^2}{2 \cdot 4}\right)
\]
Example: Gaussians (cont.)

Sketch the pdf for each class.

Sketch the pdf for each class. cf. the histograms
Parameter estimation as an optimisation problem

- Given an observation (training) set of \( N \) samples:
  \[ \mathcal{D} = \{ x_1, x_2, \ldots, x_N \} \]

- How can we estimate the mean \( \mu \) and variance \( \sigma^2 \) of the population?

- Define the problem as an optimisation problem

  **Maximum Likelihood (ML) estimation:**

  \[
  \max_{\mu, \sigma^2} p(\mathcal{D} | \mu, \sigma^2)
  \]

  NB: ML is just a one criterion for parameter estimation
ML estimation of a univariate Gaussian pdf

Assumption:
Samples $\mathcal{D} = \{x_n\}_{n=1}^N$ are drawn independently from the same distribution (i.i.d.)

Likelihood:

$$p(\mathcal{D} \mid \mu, \sigma^2) = p(x_1, \ldots, x_N \mid \mu, \sigma^2)$$
$$= p(x_1 \mid \mu, \sigma^2) \cdots p(x_N \mid \mu, \sigma^2) = \prod_{n=1}^N p(x_n \mid \mu, \sigma^2)$$
$$= L(\mu, \sigma^2 \mid \mathcal{D})$$

Optimisation problem:
Find such parameters $\mu$ and $\sigma^2$ that maximise the likelihood:

$$\max_{\mu, \sigma^2} L(\mu, \sigma^2 \mid \mathcal{D})$$
The log likelihood:

$$LL(\mu, \sigma^2 | D) = \ln L(\mu, \sigma^2 | D) = \ln \prod_{n=1}^{N} p(x_n | \mu, \sigma^2)$$

$$= \sum_{n=1}^{N} \ln p(x_n | \mu, \sigma^2)$$

$$= \sum_{n=1}^{N} \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( \frac{-(x_n - \mu)^2}{2\sigma^2} \right) \right)$$

$$= \frac{-N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}$$

NB: the natural log (ln) is assumed
ML estimation of a univariate Gaussian pdf (cont.)

\[
LL(\mu, \sigma^2 | D) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}
\]

\[
\frac{\partial LL(\mu, \sigma^2 | D)}{\partial \mu} = 2 \sum_{n=1}^{N} \frac{x_n - \mu}{2\sigma^2} = 0
\]

\[
\Rightarrow \hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n
\]

\[
\frac{\partial LL(\hat{\mu}, \sigma^2 | D)}{\partial \sigma^2} = -\frac{N}{2} \frac{1}{\sigma^2} + \sum_{n=1}^{N} \frac{(x_n - \hat{\mu})^2}{2(\sigma^2)^2} = 0
\]

\[
\Rightarrow \sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu})^2
\]
Examples of parameter estimation with MLE

True pdf (black) and estimated pdf (red) when $N=30$

$N = 30$

True pdf (black) and estimated pdf (red) when $N=1000$

$N = 1000$
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:

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

The pdf is parameterised by the mean vector $\mu = (\mu_1, \ldots, \mu_D)^T$ and the covariance matrix $\Sigma = (\sigma_{ij})$.

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

$$\mu = E[x]$$

The covariance matrix $\Sigma$ is the expectation of the deviation of $x$ from the mean:

$$\Sigma = E[(x - \mu)(x - \mu)^T]$$

$\Sigma$ is a $D \times D$ symmetric matrix: $\Sigma^T = \Sigma$

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

The sign of the covariance $\sigma_{ij}$ 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.
Covariance matrix (cont.)

\[ \Sigma = \begin{pmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \cdots & \cdots & \sigma_{1D} \\
\sigma_{21} & \sigma_{22} & \cdots & \cdots & \cdots & \sigma_{2D} \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \sigma_{ii} & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\sigma_{D1} & \sigma_{D2} & \cdots & \cdots & \cdots & \sigma_{DD}
\end{pmatrix} \]

- \[ \sigma^2_i = \sigma_{ii} \]
- \[ |\Sigma| = \det(\Sigma) : \text{determinant} \]
  - e.g. for \( D = 2 \),
  \[ |\Sigma| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = a \times d - b \times c \]
- See dimensionality reduction with PCA in Lecture Slides (3).
Parameter estimation

Maximum likelihood estimation (MLE):

\[ \mu = E[x] \]

\[ \hat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

\[ \Sigma = E[(x - \mu)(x - \mu)^T] \]

\[ \hat{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu}_{ML})(x_n - \hat{\mu}_{ML})^T \]
Correlation matrix

The covariance matrix is not scale-independent: Define the correlation matrix $R$ of correlation coefficients $\rho_{ij}$:

$$R = (\rho_{ij})$$

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

$$\rho_{ii} = 1$$

- Scale-independent (ie independent of the measurement units) and location-independent, ie:

  $$\rho(x_i, x_j) = \rho(ax_i + b, cx_j + d) \quad \text{for } a > 0, c > 0$$

- The correlation coefficient satisfies $-1 \leq \rho \leq 1$, and

  $$\rho(x, y) = +1 \quad \text{if } y = ax + b \quad a > 0$$

  $$\rho(x, y) = -1 \quad \text{if } y = ax + b \quad a < 0$$
Spherical Gaussian

\[ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]
2-D Gaussian with a diagonal covariance matrix

\[ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \quad R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]
2-D Gaussian with a full covariance matrix

\[ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \quad R = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix} \]
Example of parameter estimation of a 2D Gaussian

\[ \hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n, \quad \hat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu})(x_n - \hat{\mu})^T \]

\[ x : \begin{pmatrix} 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 \\ 2 \end{pmatrix}, \begin{pmatrix} 7 \\ 2 \end{pmatrix}, \begin{pmatrix} 7 \\ 3 \end{pmatrix} \]

\[ \mu = \frac{1}{4} \left\{ \begin{pmatrix} 5 \\ 1 \end{pmatrix} + \begin{pmatrix} 5 \\ 2 \end{pmatrix} + \begin{pmatrix} 7 \\ 2 \end{pmatrix} + \begin{pmatrix} 7 \\ 3 \end{pmatrix} \right\} = \begin{pmatrix} 6 \\ 2 \end{pmatrix} \]

\[ x_n - \mu : \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \]

\[ \Sigma = \frac{1}{4} \left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}[-1,-1] + \begin{pmatrix} -1 \\ 0 \end{pmatrix}[-1,0] + \begin{pmatrix} 1 \\ 0 \end{pmatrix}[1,0] + \begin{pmatrix} 1 \\ 1 \end{pmatrix}[1,1] \right\} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \]
Example \((cont.)\)

\[
\hat{\mu}_i = \frac{1}{N} \sum_{n=1}^{N} x_{ni}, \quad \hat{\sigma}_{ij} = \frac{1}{N} \sum_{n=1}^{N} (x_{ni} - \hat{\mu}_i)(x_{nj} - \hat{\mu}_j)
\]

\[
x : \begin{pmatrix} 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 \\ 2 \end{pmatrix}, \begin{pmatrix} 7 \\ 2 \end{pmatrix}, \begin{pmatrix} 7 \\ 3 \end{pmatrix}
\]

\[
\mu_1 = \frac{1}{4} (5 + 5 + 7 + 7) = 6
\]

\[
\mu_2 = \frac{1}{4} (1 + 2 + 2 + 3) = 2
\]

\[
x - \mu : \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}
\]

\[
\Sigma : \quad \sigma_{11} = \frac{1}{4} \left( (-1)^2 + (-1)^2 + 1^2 + 1^2 \right) = 1
\]
\[
\sigma_{12} = \frac{1}{4} \left( (-1) \cdot (-1) + (-1) \cdot 0 + 1 \cdot 0 + 1 \cdot 1 \right) = \frac{1}{2}
\]
\[
\sigma_{22} = \frac{1}{4} \left( (-1)^2 + 0^2 + 0^2 + 1^2 \right) = \frac{1}{2}
\]
Practical issues

Parameter estimation of multivariate Gaussian distribution can be difficult.

True pdf (black) and estimated pdf (red) when $N=30$

$N = 30$

True pdf (black) and estimated pdf (red) when $N=1000$

$N = 1000$
Exercise

- Try Q3, Q4, Q5 in Tutorial 6
- Try Q3 in Tutorial 8
- Try Q4 in Tutorial 8, and
  - Find $\Sigma_i^{-1}$ for $i = 1, 2$.
  - Find $|\Sigma_i|$ for $i = 1, 2$.
  - Find the correlation matrix for each class.
  - What the covariance matrix and pdf will be if the naive Bayes assumption is applied?
Additional to Q3 in Tutorial 8:
The sample variance \( \sigma_{ML}^2 \) is the maximum likelihood estimate for the variance parameter of a one-dimensional Gaussian. Consider the log likelihood of a set of \( N \) data points \( x_1, \ldots, x_N \) being generated by a Gaussian with the mean \( \mu \) and variance \( \sigma^2 \).

\[
L = \ln p(\{x_1, \ldots, x_N\} | \mu, \sigma^2) = -\frac{1}{2} \sum_{n=1}^{N} \left( \frac{(x_n - \mu)^2}{\sigma^2} + \ln \sigma^2 + \ln(2\pi) \right)
\]

Assuming that the mean \( \mu \) is known, show that that maximum likelihood estimate for the variance is indeed the sample variance.
## 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:
  \[
p(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
\]
- Estimate parameters (mean and covariance matrix) using maximum likelihood estimation
- Try Lab-6 (1st March)