COMS21202: Symbols, Patterns and Signals
Review - Part 1

Dima Damen
Dima.Damen@bristol.ac.uk

Bristol University, Department of Computer Science
Bristol BS8 1UB, UK

April 28, 2016

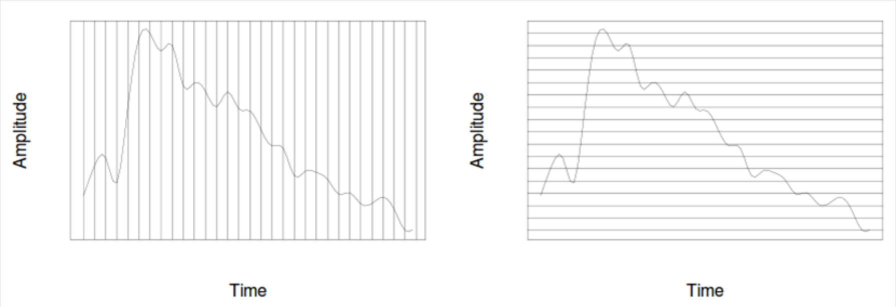
What is Data?

Data Acquisition - Analogue to Digital Conversion

Analogue to Digital conversion involves

1. Sampling
2. Quantisation

e.g. Audio Signal - 1D



Distance

- Distance is measure of separation between data.
- Can be defined between single-dimensional data, multi-dimensional data or data sequences.
- Distance is important as it:
  - enables data to be ordered
  - allows numeric calculations
  - enables calculating similarity and dissimilarity
- Without defining a distance measure, almost all statistical and machine learning algorithms will not be able to function.

Distance

A valid distance measure $D(a, b)$ between two components a and b has properties

- non-negative: $D(a, b) > 0$
- reflexive: $D(a, b) = 0 \iff a = b$
- symmetric: $D(a, b) = D(b, a)$
- satisfies triangular inequality: $D(a, b) + D(b, c) \geq D(a, c)$

Covariance Matrix

In three dimensions,

$$\Sigma = \frac{1}{N-1} \sum_i \begin{bmatrix} (v_{i1}-\mu_1)^2 & (v_{i1}-\mu_1)(v_{i2}-\mu_2) & (v_{i1}-\mu_1)(v_{i3}-\mu_3) \\ (v_{i1}-\mu_1)(v_{i2}-\mu_2) & (v_{i2}-\mu_2)^2 & (v_{i2}-\mu_2)(v_{i3}-\mu_3) \\ (v_{i1}-\mu_1)(v_{i3}-\mu_3) & (v_{i2}-\mu_2)(v_{i3}-\mu_3) & (v_{i3}-\mu_3)^2 \end{bmatrix}$$

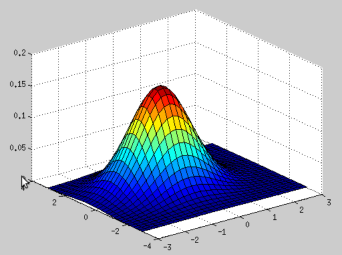
Covariance matrix is always

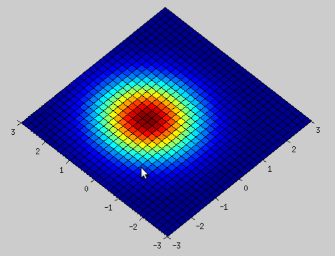
- square and symmetric
- variances on the diagonal
- covariance between each pair of dimensions is included in non-diagonal elements

Normal Distribution - Multi-dimensional

For multi-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ in M dimensions, the probability density function (pdf) can be calculated as

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})} \qquad (1)$$

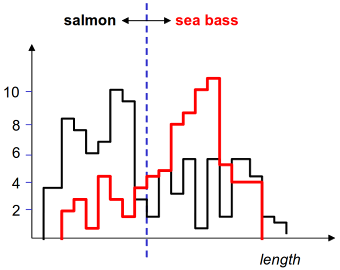




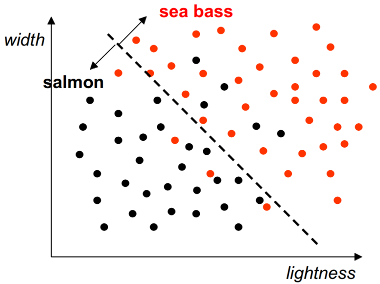
WARNING: Σ is the capital letter of σ, not the summation sign!

Model Parameters

- Models are defined in terms of parameters (one or more)
- These may be empirically obtained e.g. by trial and error
- or from training data by tuning or training the model



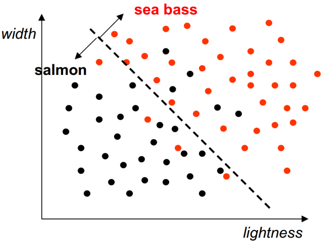
one parameter needed $x = t$



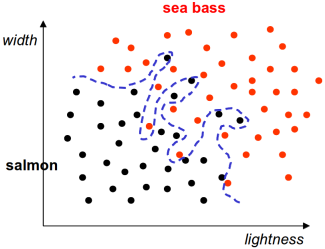
two parameters needed
$y = mx + c$

Generalisation vs. Overfitting

- **Simpler models** often give good performance and can be more **general**
- **highly complex models** over-fit the training data



two parameters needed
$y = mx + c$



A large number of parameters needs to be tuned

Another Fish Problem

Data: a set of data points $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ where x_i is the length of fish i and y_i is the weight of fish i.

Task: build a model that can predict the weight of a fish from its length

Model Type: assume there exists a polynomial relationship between length and weight

Model Complexity: assume the relationship is linear
$weight = a + b * length$

$$y_i = a + bx_i \tag{2}$$

Model Parameters: model has two parameters a and b which should be estimated.

- a is the y-intercept
- b is the slope of the line

General Least Squares - matrix form

- Matrix formulation also allows least squares method to be extended to **polynomial fitting**
- For a polynomial of degree $p+1$

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_p\ x_i^p$$

General Least Squares - matrix form

- Solved in the same manner

$$\mathbf{y}_{(N\times 1)} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{X}_{(N\times(p+1))} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^p \end{bmatrix}, \mathbf{a}_{((p+1)\times 1)} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}$$

$$\mathbf{a}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y}$$

where $(\mathbf{X}^T\mathbf{X})$ is a $(p+1)\times(p+1)$ square matrix

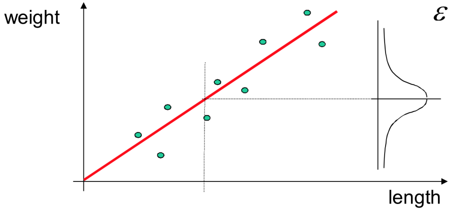
Back to Fish - Continuous

$$weight = a \times length + \epsilon$$

This is a model with one parameter, apart from the uncertainty

We can assume, for example, that ϵ is $\mathcal{N}(0, \sigma^2)$

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}}$$



Maximum Likelihood Estimation - General

- Maximum Likelihood Estimation (MLE) is a common method for solving such problems

$$\begin{aligned}\theta_{MLE} &= arg\ max_\theta\ p(D|\theta)\\ &= arg\ max_\theta \ln p(D|\theta)\\ &= arg\ min_\theta\ -\ln p(D|\theta)\end{aligned}$$

MLE Recipe

1. Determine θ, D and expression for likelihood $p(D|\theta)$
2. Take the natural logarithm of the likelihood
3. Take the derivative of $\ln p(D|\theta)$ w.r.t. θ. If θ is a multi-dimensional vector, take partial derivatives
4. Set derivative(s) to 0 and solve for θ

Probabilistic Model - Ex2

Example

Given a coin, you were assigned the task of figuring out whether the coin will land on its head or tails. You were asked to build a probabilistic model (i.e. with confidence)

- Use binomial distribution for likelihood

$$\theta_{ML} = \frac{D}{N}$$

  where D is the number of success (i.e. heads)
- Use Gaussian distribution for likelihood

$$\theta_{ML} = \frac{1}{N} \sum_{i=1}^{N} d_i$$

  where $d_i = 1$ if success (i.e. heads) or $d_i = 0$ if failure (i.e. tails)
- same answer, different view

Probabilistic Model - Likelihood and Prior

- MLE ignores any prior knowledge we may have about θ
- If we have prior knowledge about values that θ is likely to have, then we can built this into MLE

$$\theta_{ML} = arg\, max_{\theta}\, p(D|\theta)\, p(\theta)$$

- This is known as **Maximum a Posteriori (MAP)** estimation

Test - Q1

1. When calculating the Hamming distance D_H and the Edit distance D_E given two words 'bridge' and 'burger', you found that

   $\boxed{\text{(a)}}$ $D_H(\text{'bridge'}, \text{'burger'}) = 5$, $D_E(\text{'bridge'}, \text{'burger'}) = 4$

   (b) $D_H(\text{'bridge'}, \text{'burger'}) = 5$, $D_E(\text{'bridge'}, \text{'burger'}) = 5$

   (c) $D_H(\text{'bridge'}, \text{'burger'}) = 4$, $D_E(\text{'bridge'}, \text{'burger'}) = 4$

   (d) $D_H(\text{'bridge'}, \text{'burger'}) = 4$ but D_E cannot be calculated over words of the same length.

Test - Q2

2. For a sample of size N, and considering the model:

$$y = a_0 + a_1 x + a_2 xy + a_3 x.^3$$

The size of the matrices $\mathbf{y}$, $\mathbf{X}$, $\mathbf{a}$ used in the matrix form of the least squares method would be

(a) $\mathbf{y}_{N\times 1}$, $\mathbf{X}_{N\times 4}$, $\mathbf{a}_{1\times 4}$

(b) $\mathbf{y}_{N\times 4}$, $\mathbf{X}_{N\times 4}$, $\mathbf{a}_{4\times 1}$

$\boxed{\text{(c)}}$ $\mathbf{y}_{N\times 1}$, $\mathbf{X}_{N\times 4}$, $\mathbf{a}_{4\times 1}$

(d) $\mathbf{y}_{N\times 1}$, $\mathbf{X}_{N\times 3}$, $\mathbf{a}_{4\times 1}$

(e) Least squares cannot be used to solve for this polynomial due to the presence of the term $a_2 xy$

Test - Q3

3. For $x = (10, 2)$ and $y = (6, 5)$ which of the following is a correct Minkowski distance?

   (a) For $p = 1$, D(x,y) = 1

   (b) For $p = 2$, D(x,y) = 4

   (c) For $p = 3$, D(x,y) = 9.5

   $\boxed{\text{(d)}}$ For $p = \inf$, D(x,y) = 4

Test - Q4

4. Which of the following pairs of a model and its parameters are incorrect

   $\boxed{\text{(a)}}$ A normal distribution has a single parameter μ

   (b) A uniform distribution has two parameters representing the range $[a, b]$

   (c) A linear function $y = mx + c$ has two parameters representing the slope and the y-intercept

   (d) A binomial distribution has one parameter representing the probability of a success α

Test - Q5

5. For a one dimensional numeric data, given a probabilistic model with a single parameter b, $var(b_{ML})$ was calculated to be $var(b_{ML}) = \sigma^2 \sum_i x_i$. Based on this finding you advise the data collection team to:

   (a) Collect samples with large values of x_i if possible.

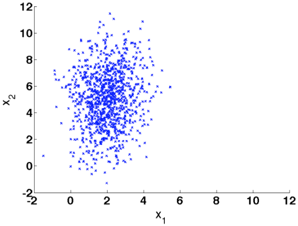
   $\boxed{\text{(b)}}$ Collect samples with small values of x_i if possible.

   (c) Model parameter estimation does not depend on the sample collected, so no change in data collection is needed.

   (d) Collect samples that achieve a uniform distribution of x_i over its range

Test - Q6

6. For the data sample of 1000 points shown here



which of the following is a reasonable estimate of the model parameters

$\boxed{\text{(a)}}$ $\mu = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 5 \end{bmatrix}$

(b) $\mu = \begin{bmatrix} 2 \\ 8 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 3 & 3 \\ 3 & 1 \end{bmatrix}$

(c) $\mu = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & -3 \\ -3 & 4 \end{bmatrix}$

(d) $\mu = \begin{bmatrix} 2.5 \\ 5 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Test - Q7

7. When discussing the concepts of generalisation versus overfitting, which of the following statements is NOT correct:

   (a) An overfitted model achieves better results when tested on the 'training' data

   (b) A general model achieves better results on 'future' data

   $\boxed{\text{(c)}}$ A general model is more complex than an overfitted model

   (d) An overfitted model has a higher number of parameters to optimise when compared to a general model

Test - Q8

8. For a one-dimensional numeric data D, given a representation of $p(D|\theta)$ for a probabilistic model, **MLE** estimates the model parameter $\hat{\theta} = \arg\max_\theta p(D|\theta)$. Which of the following is incorrect

   (a) $\hat{\theta} = \arg\max_\theta lnp(D|\theta)$ where ln is the natural logarithm function

   (b) $\hat{\theta} = \arg\max_\theta p(D|\theta) + c$ where c is a constant

   (c) $\hat{\theta} = \arg\min_\theta bp(D|\theta)$ where $b < 0$ is a constant

   $\boxed{\text{(d)}}$ $\hat{\theta} = \arg\max_\theta p(D + c|\theta)$ where $c > 0$ is a constant

Test - Q9

9. The assumption that a sample is **i.i.d** implies that

   (a) The data has been sampled by an expert who has studied the full population.

   $\boxed{\text{(b)}}$ The observations are believed to be independent.

   (c) The sample is large enough to estimate the model parameters.

   (d) The sample is multi-dimensional.

Test - Q10

10. Which of these files has the largest size if stored, raw/uncompressed?

    (a) A one minute phone call with your friend. Recall that speech is sampled at 8Khz and quantised at 8bps.

    (b) 10 seconds of an Audio CD. Recall that Audio CD contains stereo data sampled at 44KHz and quantised at 16bps.

    $\boxed{\text{(c)}}$ A colour photo on a 16Mega Pixels camera. Recall that each colour channel is quantised at 8bps.

    (d) A 0.5 second colour video recorded without audio using 1Mega Pixels camera. Note that videos are recorded at 30 frames per second. Recall that each colour channel is quantised at 8bps.

Note

- Use this lecture for revision NOT for studying!