CSC345/M45:
Big Data & Machine Learning
(revision and exam)

Dr. Xianghua Xie

x.xie@swansea.ac.uk
http://csvision.swan.ac.uk
510 Faraday Tower

Exam

- 2 hours written examination
- Answer BOTH questions

- Question type
  - Bookwork, that is largely show what you remember from the lectures, e.g. algorithmic steps
  - Analytical;
    - you need to show your understanding of the presented problem and appropriate method to tackle the problem
    - be able to compare different types of similar methods
  - Not required to recite equations;
  - No multiple choices type questions

- Topic-by-topic
  - revision
  - Explain exam requirements
  - Mock questions

Basics (Introduction & Fundamentals)

- “Big data” concept
  - Slides 6 & 7
- Characteristics of “Big Data”
  - Slide 7

- Machine learning concept and types of machine learning
  - Slides 21, 22, & 23

- Supervised learning examples
- Unsupervised learning examples
  - Slide 25

Basics (Introduction & Fundamentals)

- Mock Exam Question

  - Describe “Supervised Learning” and provide two examples of supervised learning techniques.

Clustering: k-means

- Clustering concept & common clustering techniques
  - Slide 3

- Understand the impact of outliers on clustering results
  - Slide 4

- K-means concept & algorithmic steps
  - Slides 9, 10 & 11. No need to recite any equations
  - Be able to describe and illustrate the clustering process

Clustering: k-means

- Mock Exam Question
  - Explain how k-means calculate the quality of clustering. List the algorithmic steps of k-means. (You are not required to recite the equations)

Clustering: GMM

- Concept of GMM and its difference to K-means
  - Slides 1, 2 & 3

- Gaussian mixture parameters: be able to identify every GMM parameters and their role in GMM clustering
  - Slide 5

- GMM clustering steps
  - Slides 10 & 11

Clustering: GMM

- Mock exam question
  - Explain the difference between k-means and Gaussian mixture modelling in data clustering. List the algorithmic steps for GMM and its parameters.

Linear Regression

- Linear regression concept
  - Slides 2, 4 & 5 (no need to recite equations)
- What criteria does linear regression use? (no need to recite equation)
  - Slide 8
- Linear regression for nonlinear fitting
  - Slides 10, 12 & 13
  - No need to recite equations
- Quadratic regularisation
  - Slide 16
- Bias, variance, and expected loss
  - Slides 18 & 19

Linear Regression

- Mock exam questions
  - Explain how linear regression determine the “optimal” fitting and the importance of regularisation.
  - What are bias and variance? How to calculate expected loss in linear regression?

Dimensionality Reduction: PCA

- Why dimensionality reduction is useful?
  - Slides 3 & 4
- PCA concept
  - Slides 5, 17 -20
- Variance and Covariance
  - Slides 8 – 13
  - Understand what they measure (no need to recite equations)
  - Understand covariance matrix; be able to identify every element in the covariance matrix (Slide 13)
- Eigenvector and eigenvalue
  - Slides 14 – 16
  - Understand what they measure
- Understand the limitation of PCA (slides 25 and 26)

Dimensionality Reduction: PCA

- Mock exam question
  - The covariance matrices of two 2D clusters of points are given as:

$$\Sigma_1 = \begin{bmatrix} 5 & -2 \\ -2 & 6 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 5 & 0 \\ 0 & 6 \end{bmatrix}$$

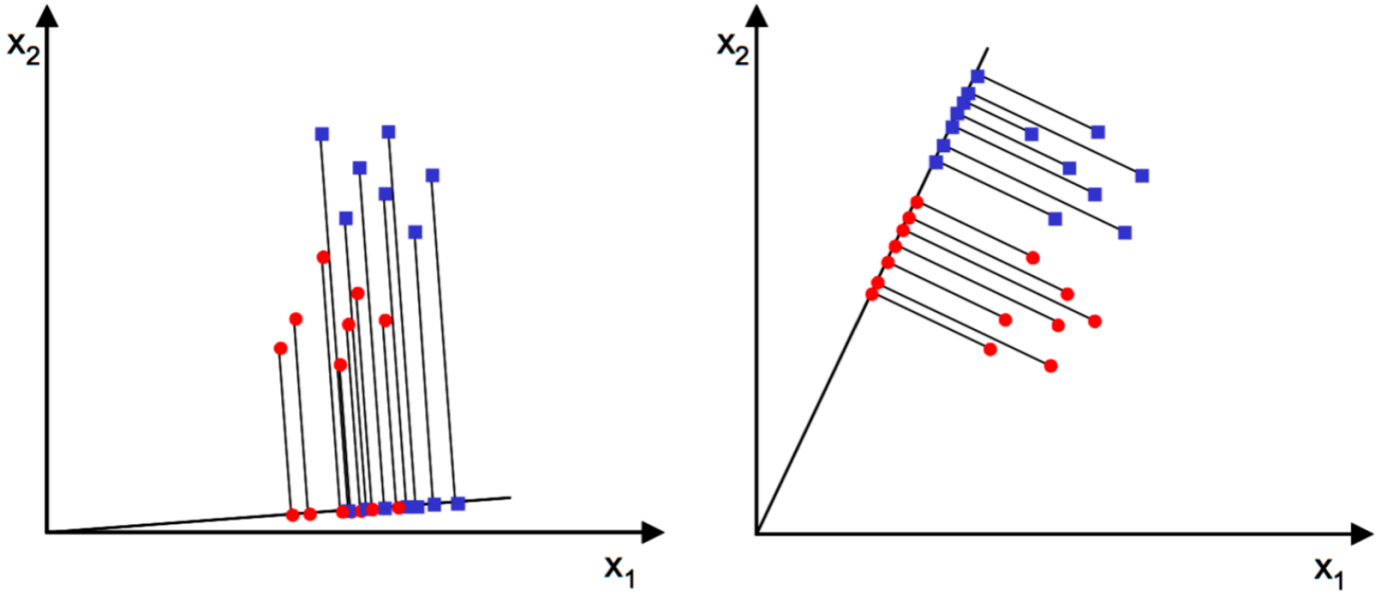
    - Identify the variances along each axis for both cases.
    - Assuming the data points are centred at the origin, draw diagrams to illustrate possible distributions of these clusters of points. Explain why you think the data points may take such shapes.
    - Which of the two is de-correlated? Why? How to de-correlated the other?

Dimensionality Reduction: LDA

- Concept of LDA and its advantages over PCA
  - Slides 2, 7 & 8; also slide 16
- Be able to describe how LDA can be performed
  - Slides10 & 11
  - No need to recite equations
- How LDA can be applied to multi-class problems (number of classes is larger than 2)
  - Slide 12
- LDA limitations
  - Slides 14 & 15

Dimensionality Reduction: LDA

- Mock exam question
  - A 2D dataset contains two classes of objects and their distributions are plotted below (red and blue denote class labels). Two possible projections to reduce the dimensionality of the data are also shown in the figure. Which projection should be considered better? Explain your answer.



- Slide 16 is also a mock exam question

Logistic Regression

- Concept of Logistic Regression
  - Slides 8 & 9
  - No need to recite equations
- Understand why linear regression is insufficient in solving discrete classification problems
  - Understand the toy example given on slide 7
- Understand how linear decision boundary is encoded in logistic regression
  - Slides 11 & 12
  - No need to recite equations
- Understand how logistic regression cope with nonlinear decision boundaries
  - Slide 13

Logistic Regression

- Understand how to evaluate decision boundary
  - Slides 20 – 23
  - No need to recite equations
- Understand how to deal with multi-class classification
  - Slides 27 & 28

Logistic Regression

- Mock Exam Question
  - Given the following data, how to apply logistic regression to predict outcome

Case Summaries[a]

	Cured?	Intervention	Number of Days with Problem before Treatment	Predicted probability	Predicted group
1	Not Cured	No Treatment	7	.42857	Not Cured
2	Not Cured	No Treatment	7	.42857	Not Cured
3	Not Cured	No Treatment	6	.42857	Not Cured
4	Cured	No Treatment	8	.42857	Not Cured
5	Cured	Intervention	7	.71930	Cured
6	Cured	No Treatment	6	.42857	Not Cured
7	Not Cured	Intervention	7	.71930	Cured
8	Cured	Intervention	7	.71930	Cured
9	Cured	No Treatment	8	.42857	Not Cured
10	Not Cured	No Treatment	7	.42857	Not Cured
11	Cured	Intervention	7	.71930	Cured
12	Cured	No Treatment	7	.42857	Not Cured
13	Cured	No Treatment	5	.42857	Not Cured
14	Not Cured	Intervention	9	.71930	Cured
15	Not Cured	No Treatment	6	.42857	Not Cured
Total N	15	15	15	15	15

a. Limited to first 15 cases.

SVM

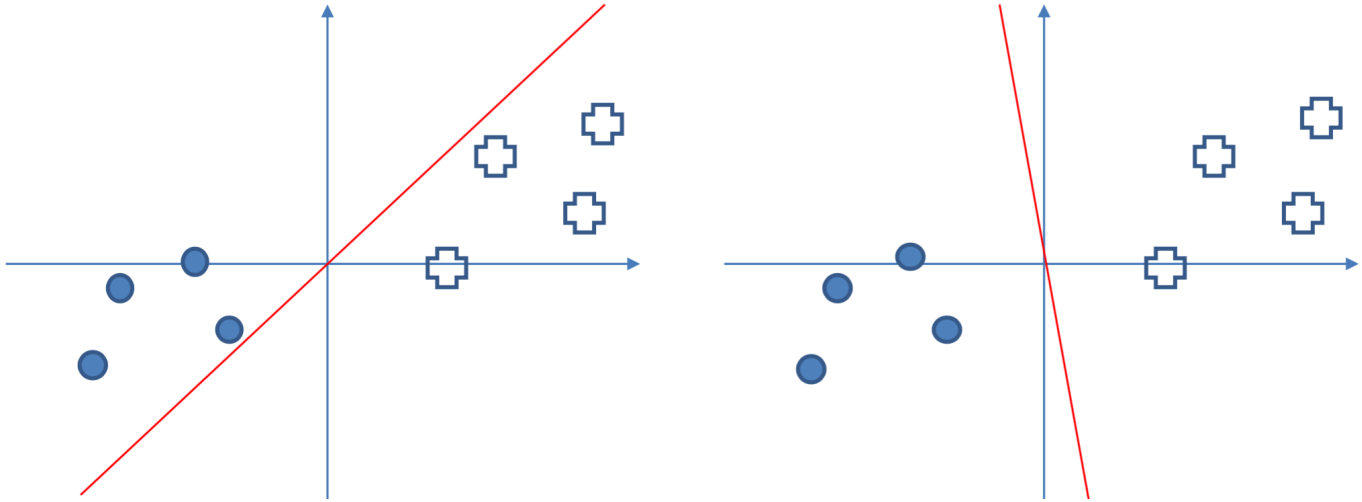
- Understand the difference between clustering and classification
  - Slides 2 & 3
- Understand why SVM is considered to be a large margin classifier
  - Slides 12, 14
- Understand how SVM cope with nonlinear decision boundaries
  - Slides16 & 17
  - No need to recite equations

SVM

- Concept of cross validation
  - Slides 19 & 20
- Confusion matrix
  - Slide 21
- ROC curve
  - Slides 22 – 25
  - Understand how to evaluate using ROC
  - No need to recite equations
- Understand the trade-off between false positive and false negative

SVM

- Mock exam question
  - According to SVM, which one of the two decision boundaries (red lines) is considered to be better? Why?



  - For a given class distribution and prediction distribution, be able to write down the confusion matrix. E.g. 200 positives and 400 negatives. 150 correct identified as positives and 200 correctly identified as negatives.

NN

- NN concept
  - Slides 7, 10, 11 and 13
- Understand the role logistic regression in NN
- Understand how to deal with multiple classes
  - Slide 18
- Understand why random initialisation is necessary
  - Slide 30
- Understand what is forward propagation
  - Slides 19
  - No need to recite equations
- Understand what is backward propagation
  - Slides 20, 21
  - No need to recite equations

NN

- Be able to provide examples of NN to approximate logical operators
  - Slides 34, 36, 37, 39, and 41

- Understand various forms of NN architecture
  - Slide 43

NN

- Mock exam question
  - Give an example of weight values, ω, so that this perceptron approximates the logical (NOT x1) AND (NOT x2) operation. Show how your weights give the desired effect.