May 2014 **Question 3:**

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	0.346	0.124	0.124
PC 2	0.299	0.092	0.216
PC 3	0.293	0.089	0.305
PC 4	0.285	0.084	0.389
PC 5	0.272	0.077	0.466
PC 6	0.271	0.076	0.542
PC 7	0.268	0.074	0.616
PC 8	0.260	0.070	0.686
PC 9	0.244	0.062	0.747
PC 10	0.239	0.059	0.807
PC 11	0.237	0.058	0.865
PC 12	0.228	0.054	0.919
PC 13	0.204	0.043	0.962
PC 14	0.193	0.038	1.000

Figure 1. PCA model

a) Give an overview of the effect of applying *Principal Component Analysis* (PCA) to a dataset of document vectors with 200 terms.

In your answer, explain how to interpret the *Proportion of Variance* and *Cumulative Variance* of the PCA model in Figure 2 above.

(*11 marks*)

b) With the aid of a diagram, give an overview of how a *Support Vector Machine* classifies a dataset of document vectors. Assume it's a binary classification task.

(*8 marks*)

c) Compare a *Support Vector Machine* with *one* other classification algorithm you are familiar with in terms of: accuracy; how well they handle large dimensionality; and ability to detect non-linear class boundaries.

(6 marks)

Total: 25 Marks

Repeat 2014: **Question 3:**

a) Explain the difference between a *symmetric* and *asymmetric* variable. What is the significance of this distinction when clustering document vectors? Illustrate your answer with reference to the two document vectors in Table 3 below.

Table 3: Document Vectors

	Course	Golf Club	Student	Work
Document A	0	1	0	0
Document B	1	0	0	0

(*8 marks*)

b) Using the five document vectors and cosine similarity measures given in Table 4 below, explain and illustrate how *agglomerative hierarchical clustering* would cluster these documents. Assume you are using single linkage. Include the resulting *dendogram* in your answer.

Table 4: Five document vectors and their cosine similarities

Document vectors for 5 documents, and 4 terms:

	author	body	crime	novel
Doc1	1	0	1	1
Doc2	0	1	1	0
Doc3	2	1	4	1
Doc4	0	3	5	0
Doc5	4	0	0	4

Cosine similarity between documents pairs:

Documents:	Similarity
Sim(Doc2, Doc4) =	1.0
Sim(Doc1, Doc3) =	0.9
Sim(Doc3, Doc4) =	0.8
Sim(Doc1, Doc5) =	0.8
Sim(Doc2, Doc3) =	0.8
Sim(Doc1, Doc4) =	0.5
Sim(Doc3, Doc5) =	0.5
Sim(Doc1, Doc2) =	0.4
Sim(Doc2, Doc5) =	0.0
Sim(Doc4, Doc5) =	0.0

(*10 marks*)

c) Discuss both *subjective* and *objective* evaluation of the performance of a clustering algorithm. Your answer should explain the difference between a subjective and objective evaluation, and the benefits of each approach.

(7 marks)

Total: 25 Marks

Question 4, Summer 2013.

a) Discuss the difference between *symmetric* and *asymmetric* variables, and the relevance of this when choosing the most appropriate similarity measures for clustering text documents. Illustrate your answer with reference to the three document vectors given below:

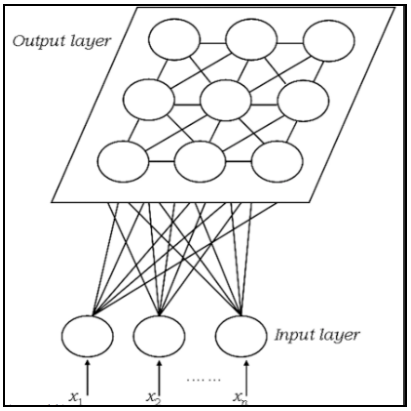
	Design	**Police**	**Control**	**Salary**
Document 1	1	1	1	0
Document 2	0	1	0	0
Document 3	0	0	0	1

(*8 marks*)

b) Discuss both *subjective* and *objective* evaluation of the performance of a clustering algorithm. Your answer should explain the benefits of both approaches, and explain one objective performance measure that can be used.

(8 marks)

c) With reference to the diagram below, give an overview of how a **Kohonen network** can be used to as a self organising map to arrange documents by topic.



(*9 marks*)

[Total: 25 marks]

Question 4, Repeat 2013.

Document vectors for 5 documents, and 4 terms:

	design	engineer	manufacture	cost
Doc1	1	0	1	3
Doc2	0	2	0	0
Doc3	1	1	2	4
Doc4	0	2	0	3
Doc5	0	0	3	0

Cosine similarity between every pair of documents

Documents:	**Similarity**
Sim(Doc1, Doc3) =	1.0
Sim(Doc3, Doc4) =	0.8
Sim(Doc1, Doc4) =	0.7
Sim(Doc2, Doc4) =	0.6
Sim(Doc3, Doc5) =	0.4
Sim(Doc1, Doc5) =	0.3
Sim(Doc2, Doc3) =	0.2
Sim(Doc1, Doc2) =	0.0
Sim(Doc2, Doc5) =	0.0
Sim(Doc4, Doc5) =	0.0

a) Using the similarity measures given above between the five documents, explain and illustrate how agglomerative hierarchical clustering would cluster these documents. Assume you are using single linkage. Include the resulting dendogram in your answer.

(*12 marks)*

b) Explain the difference between *single linkage* and *complete linkage*. How does the selection of merge points impact on the clusters formed?

(8 marks)

c) What are the advantages of using hierarchical clustering? Are there any disadvantages to this clustering approach?

(5 marks)

[Total: 25 marks]

Question 2, May 2012.

a) Explain how a Genetic Algorithm can be adapted to cluster document vectors. Your answer should cover how clusters are represented, a suitable fitness measure, how mutation is implemented and how crossover is implemented. Use the two solutions given below to illustrate your answer.

Solution X

doc10	doc7	doc9	doc4	doc1	doc5	doc6	doc2	doc8	doc3
0.5	0	0.2	0.1	0	0	0.3	0.1	0	0.1

Solution Y

doc3	doc9	doc4	doc7	doc1	doc5	doc6	doc10	doc8	doc2
0	0.1	0.3	0	0.2	0	0	0.5	0.1	0.6

Note: the number under each document is the cosine similarity measure between that document and the document to its right.

(20 marks)

b) Explain why asymmetric distance measures should be used when clustering document vectors.

(5 marks)

[Total: 25 marks]

Question 4, August 2012

Document vectors for 5 documents, and 4 terms:

	game	win	crisis	debt
Doc1	1	0	1	0
Doc2	3	2	0	0
Doc3	0	1	2	4
Doc4	0	2	3	3
Doc5	4	0	3	0

Cosine similarity between every pair of documents

Documents:	**Similarity**
Sim(Doc1, Doc5) =	1.0
Sim(Doc3, Doc4) =	0.9
Sim(Doc2, Doc5) =	0.7
Sim(Doc1, Doc2) =	0.6
Sim(Doc1, Doc4) =	0.5
Sim(Doc4, Doc5) =	0.4
Sim(Doc1, Doc3) =	0.3
Sim(Doc3, Doc5) =	0.3
Sim(Doc2, Doc4) =	0.2
Sim(Doc2, Doc3) =	0.1

d) Using the similarity measures given above between the five documents, explain and illustrate how agglomerative hierarchical clustering would cluster these documents. Assume you are using single linkage. Include the resulting dendogram in your answer.

(*12 marks)*

e) Explain the difference between single linkage and complete linkage. How does the selection of merge points impact on the clusters formed?

(8 marks)

f) What are the advantages of using hierarchical clustering? Are there any disadvantages to this clustering approach?

(5 marks)

[Total: 25 marks]

May 2011, Question 4.

*a) A **Euclidean** distance measure, and a **Cosine** similarity measure, are two alternative approaches to measure the level of similarity between rows of data. Which is better suited to document vectors? Explain your answer.*

(6 marks)

b) Explain with the aid of a diagram how a neural network can be used to organise documents, represented as document vectors, into a grid structure (i.e. a Self Organising Map).

(11 marks)

c) Discuss both subjective and objective evaluation of the performance of a clustering algorithm. Your answer should explain the benefits of both approaches, and give an example of one performance measure that can be used.

(8 marks)

[Total: 25 marks]