Tutorial 1: Simple recommender system, clustering and classification

1. Consider the follow newspaper review ratings based on the 'Review of Reviews' in the Guardian (12 Jan 2009):

	Guardian	Times	Telegraph	Independent
Film				
Slumdog Millionaire	6	8	10	8
Sex Drive	4	4	4	4
The Reader	1	4	6	3
Che	6	6	6	4
Theatre				
Mary Goes First	6	6	6	3
Well	2	6	2	3
Three Women	4	8	9	6

Steve has seen the four films and rated them as follows:

Slumdog Millionaire	9
Sex Drive	2
The Reader	4
Che	5

Use the Euclidean distance for (a)–(c).

(a) Based on the reviews of the four films: Which pair of newspapers are the most similar? Which are the most dissimilar? Which is most similar to Steve?

(b) Using the similarity measure derived from the Euclidean distance introduced in the lectures, which theatre production should Steve go to see, based on the film reviews?

(c) Someone goes to the cinema to see *Che*, but it is sold out. Based on the critics' ratings which film should be recommended as most similar to *Che*?

(d) Write a Matlab/Python function to compute the similarity between two feature vectors using the Pearson correlation coefficient. Which two critics are most similar using the Pearson correlation coefficient?

Discuss the limitations and advantages of these simple recommender systems.

2. Pearson's correlation function for two sample sets, $X = \{x_n\}_1^N$ and $Y = \{y_n\}_1^N$, is given by

$$r_{xy} = \frac{1}{N-1} \sum_{n=1}^{N} \left(\frac{x_n - \bar{x}}{s_x} \right) \left(\frac{y_n - \bar{y}}{s_y} \right),$$

where $\bar{x}, \bar{y}$ are the sample means, and s_x, s_y are the sample standard deviations for X and Y respectively.

   (a) Prove that $-1 \leq r_{xy} \leq 1$.

   (b) Sketch some examples for $r_{xy} \approx 1, -1, 0$.

3. Consider the following 2-dimensional data set:

$$\mathbf{x}_1 = (1,1); \quad \mathbf{x}_2 = (4,4); \quad \mathbf{x}_3 = (5,1); \quad \mathbf{x}_4 = (7,1); \quad \mathbf{x}_5 = (7,4); \quad \mathbf{x}_6 = (7,10).$$

Cluster this data set into two clusters using k-means clustering, with the cluster centres initialised to $(1,1)$ and $(7,10)$.

4. Given a set of training samples and a distance measure, we can partition the input space into regions such that all points in a region will be assigned to one of the training samples, as we saw in Note 4 on nearest neighbour classification. This partitioning of a plane is referred to as the *Voronoi tessellation* or *Voronoi diagram.*

   (a) Consider the following three data points for training:

$$\mathbf{x}_1 = (0,0); \quad \mathbf{x}_2 = (0,4); \quad \mathbf{x}_3 = (2,2),$$

   which define three regions in the 2-dimensional input space. Compute the boundaries between the three pairs of regions and hence sketch the overall Voronoi diagram, assuming we use the Euclidean distance.

   (b) Consider additional training samples:

$$\mathbf{x}_4 = (3,0); \quad \mathbf{x}_5 = (3,4);$$

   Sketch the overall Voronoi diagram that is formed by all the training samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_5\}$.

   (c) Assume there are three classes, $C_1, C_2,$ and C_3, and we know that $\{\mathbf{x}_1, \mathbf{x}_2\}$ belong to C_1, $\{\mathbf{x}_3\}$ belongs to C_2, and $\{\mathbf{x}_4, \mathbf{x}_5\}$ belong to C_3. Identify the decision boundaries between the pairs of classes and decision regions in the diagram you sketched in (b).