Escuela Técnica Superior de Ingeniería de Telecomunicación
Dpto. de Teoría de la Señal y Comunicaciones, Sistemas Telemáticos y Computación

Artificial Intelligence and Learning

(26/01/2021)

Name and last name: **Duration: 90 minutes**

[4.0 points] **Question 1**. **Multiple answers**. Mark ALL correct choices: there may be more than one correct choice, but there is always at least one correct choice.

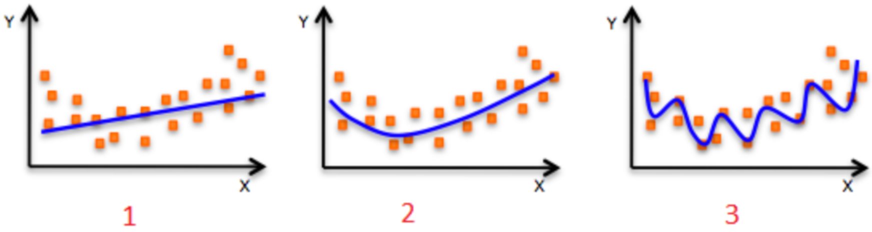
Question 1.1 Which of the following are true for the k-nearest neighbour (k-NN) algorithm?

□ k-NN can be used for both classification and regression.
□ As k increases, the bias usually increases
□ The decision boundary looks smoother with smaller values of k.
□ As k increases, the variance usually increases.

Question 1.2 What strategies can help reduce overfitting in decision trees?

□ Enforce a unique random state
□ Enforce a minimum number of samples in leaf
□ Make sure each leaf node is one pure class nodes
□ Enforce a maximum depth for the tree

Question 1.3 The following visualization shows the fit of three different models (in blue line) on same training data. What can you conclude from these visualizations?



□ The training error in first model is higher when compared to second and third model
□ The best model for this regression problem is the last (third) model, because it has minimum training error
□ The second model is more robust than first and third because it will perform better on unseen data
□ The third model is overfitting data as compared to first and second model
□ All models will perform same because they have not seen the test data.

Question 1.4 Which of the following are metrics suitable for measuring regression performance?

- ☐ Correlation coefficient
- ☐ Root mean squared error
- ☐ Mean absolute error
- ☐ Accuracy
- ☐ Recall
- ☐ Coefficient of determination

Question 1.5 Why is PCA sometimes used as a pre-processing step before regression?

- ☐ To reduce overfitting by removing poorly predictive dimensions.
- ☐ To expose information missing from the input data.
- ☐ To make computation faster by reducing the dimensionality of the data.
- ☐ For inference and scientific discovery, we prefer features that are not axis-aligned.

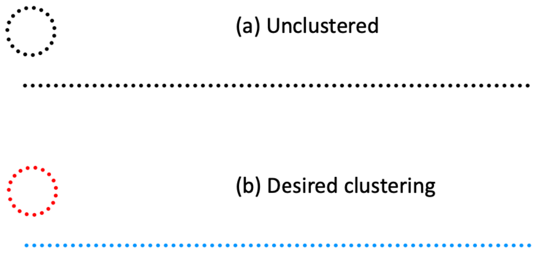
Question 1.6 What purpose(s) may be pursued when selecting features?

- ☐ Reducing the number of missing values
- ☐ Removing useless features to save computing time and data storage
- ☐ Improving prediction performance (there is less risk of overfitting if we start from a lower dimensional space)
- ☐ Creating new features from the raw data

Question 1.7 What is the difference between filters, wrappers, and embedded methods?

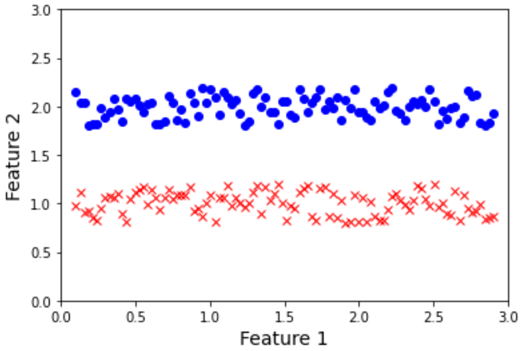
- ☐ Embedded methods select features independently of the performances of the learning machine, using a "relevance" criterion.
- ☐ Wrappers and filters methods both select feature subsets using a learning machine.
- ☐ Wrappers explore the space of all possible feature subsets using a search method; the learning machine is used to assess the subsets
- ☐ Embedded methods perform feature selection in the process of learning. They return a feature subset and a trained learning machine

Question 1.8 Which of the following methods will cluster the data in panel (a) of the figure below into the two clusters (red circle and blue horizontal line) shown in panel (b)? Every dot in the circle and the line is a data point. In all the options that involve hierarchical clustering, the algorithm is run until we obtain two clusters.



☐ Hierarchical agglomerative clustering with Euclidean distance and complete linkage
☐ Hierarchical agglomerative clustering with Euclidean distance and single linkage
☐ Hierarchical agglomerative clustering with Euclidean distance and centroid linkage
☐ k-means clustering with k = 2.

[1.0 points] **Question 2.** Let's try to identify the most important features in a simple dataset in 2D.



A) Assume that you have a linear classification model that can be fitted on Feature 1, Feature 2, or both.
   a. Describe the training error of this linear classifier that can see only the first feature of the data.
   b. Describe the training error of this linear classifier that can see only the second feature.

B) Construct a toy dataset in 2D (two features) where a variable is useless by itself, but potentially useful alongside the second one.

[1.5 points] **Question 3**. PCA. Given 3 data points in 2-d space: (1, 1), (2, 2) and (3, 3),

A) What is the first principal component?
B) If we want to project the original data points into 1-d space by principal component you choose, what is the variance of the projected data?
C) For the projected data in (B), now if we represent them in the original 2-d space, what is the reconstruction error?

[2.0 points] **Question 4.** In this problem, you will perform K-means clustering manually, using Euclidean distance, with *K = 2*, on a small example with *n = 6* observations and *p* = 2 features (x_1 and x_2). The observations are as follows.

x_1	x_2	Obs.
1	4	1
1	3	2
0	4	3
5	1	4
6	2	5
4	0	6

A) Plot the observations.
B) Assume that cluster centroids were initialized randomly at (1,1) and (3,4), respectively. Assign a cluster label to each observation.

C) Compute the new centroid for each cluster.

D) Assign each observation to the centroid to which it is closest. Report the cluster labels for each observation.

E) Repeat (C) and (D) until the answers obtained stop changing.

[1.5 points] **Question 5**. Suppose that we have four observations, for which we compute a Euclidean distance matrix, given by

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

For instance, the (Euclidean) distance between the first and second observations is 0.3, and the distance between the second and fourth observations is 0.8.

A) On the basis of this distance matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage[1]. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

B) Repeat (A), this time using single linkage[2] clustering.

C) Suppose that we cut the dendrogram obtained in (A) such that two clusters result. Which observations are in each cluster?

D) Suppose that we cut the dendrogram obtained in (B) such that two clusters result. Which observations are in each cluster?

[1] Complete linkage: the similarity of two clusters is the similarity of their most dissimilar members

[2] Single linkage: the similarity of two clusters is the similarity of their most similar members