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

| | |
|---|---|
| C1 | |
| C2 | |
| C3 | |
| C4 | |
| C5 | |
| C6 | |
| C7 | |
| C8 | |

# Artificial Intelligence and Learning

**Duration: 120 minutes**

**Name and last name:**

**You must hand in your exam statement**

[2.4 points] **Question 1**. Answer the following questions. Each correct answer adds 0.4 points. Each incorrect answer subtracts 0.10 points.

*Question 1.1* Select the correct answer:

a) A class is better represented when there are more observations of that class.
b) The set of training observations should be independent of the test set. As a consequence, both sets must have a different number of features.
c) In supervised classification, an alternative to the imbalance or unbalance between classes is to change the label of some observations until all classes have the same number of instances or observations.
d) When doing the normalization of the features, the normalization parameters should be extracted only from the training set.

*Question 1.2*. Regarding data cleaning and pre-processing. Please indicate which of the following statements is correct:

a) Missing data is not a problem for machine learning methods.
b) Data must always be normalized.
c) If a record has outliers, you must always delete that record.
d) No answer is correct.

*Question 1.3*. Regarding the regularization, please indicate which of the following statements is correct:

a) Regularization attempts to control overfitting by modifying the cost function.
b) Regularization is a method that increases the overfitting of the model.
c) Regularization is a method that speeds up the execution of the model.
d) No answer is correct.

*Question 1.4* Indicate the **incorrect** answer:

a) In a linear regression model, the values of the coefficients are not related to the importance that the independent variables have on the dependent variable.
b) If $R^2 = r^2 = 1$ this indicates that there would be no error in the predictions made.
c) As for the simple linear regression, in the multiple linear regression also the method of least squares estimation is used to estimate the value of the coefficients $w_j$.
d) Considering the regression line shown in Figure1, and taking into account that we only have the following 3 pairs of points (X, Y): (1, 7.5), (2, 9.3) and (3, 10.7), the residuals obtained are, respectively, -0.5, -0.3 and 0.3.

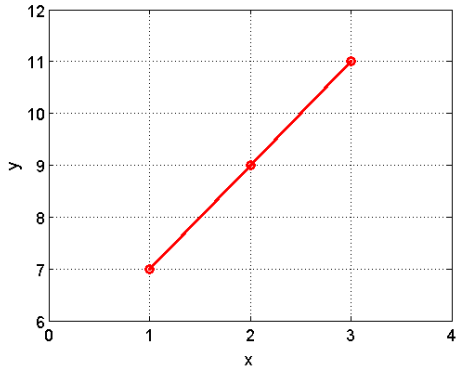


Figure 1. Linear regression.

*Question 1.5.* About decision tree-based schemes, please indicate the **incorrect** answer:

a) Classification trees are not very robust. Small changes in the training samples can cause large changes in the trees built and in their predictions.
b) Classification (decision) trees are supervised non-parametric, non-linear statistical learning methods.
c) Partitioning is usually used to prevent trees from being too deep and complex, as they tend to adjust to the data. These strategies are called pruning.
d) The construction of a classification tree is done recursively. A variable is chosen according to a measure of purity and the space is divided into two disjunct regions. Thus, two new branches are obtained in the tree.

*Question 1.6*. Which of the following feature differentiates a multilayer perception from a simple one?

a) Weights are initialized with random values.
b) It is a universal approximator of functions.
c) They have the ability to learn to recognize patterns.
d) No answer is correct.

[0.6 points] **Question 2**. Consider that you have a data set of 735 patients. In addition, you also have the results of the last encounters to the health system, which records values of: date, antibiotics intaken by the patient, blood samples tests and vital signs. If you wish to make use of all the available information, please indicate and justify:

a) The size of the feature vector that characterizes each patient, as well as the nature of these characteristics. Justify your answer.
b) Justify whether the following statement is true or false:

“First, the necessary tasks to prepare the data set should be carried out (data integration, data cleaning, data transformation and data reduction), and then the values of each variable should be normalized. “

[1 point] **Question 3.** Explain, in your own words the following statement: "garbage in garbage out", also known as GIGO.

Considering the description of the type of variables shown in Table 1, indicate and reasonably justify (but briefly and concisely) three pre-processing techniques that you would consider appropriate to apply.

| Age | Birthday | City | Postal code | Main Diagnosis | Salary (€) |
|---|---|---|---|---|---|
| 35 | | Toledo | | 250 | 32 000 |
| 23 | 02/06/1987 | Murcia | | V27 | 30 000 |
| 18 | 12/12/1992 | Madrid | 28900 | 401 | 28 000 |
| 42 | | Cuenca | | 496 | |
| 29 | 23/05/1990 | Madrid | 28914 | 384 | 36 000 |
| 23 | 17/08/1987 | | | 270 | 25 000 |
| 37 | 28/03/1972 | Sevilla | | V27 | 50 000 |
| 54 | 05/05/1954 | Teruel | | 401 | |
| 21 | 30/01/1989 | | | 250 | 48 000 |
| 40 | 13/02/1970 | | | 296 | 66 000 |
| 90 | | Murcia | | 382 | |

Table 1. Features considered in a data set.

[1 point] **Question 4.** Briefly explain what the Training, Validation and Test sets are used for. Explain, in your own words, how to compute a 4-fold cross validation strategy. You can use a scheme to support your answer.

[2 points] **Question. 5** For each code shown in Figure 2:

a) Explain briefly the task to be learned (classification of regression).
b) The learning technique (model) used to achieve the proposed task.
c) Explain if the proposed code should be enough for achieving generalization purposes. If not, you should explain what is missing or what should be included.

## Code 1

```
### Code 1
# Load data
data = pd.read_csv("pima_indian_diabetes.csv")
data.head(10)
x = data[["Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]
y = data["Outcome"]

# Learning process
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
scaler = StandardScaler()
scaler.fit(X_train)
X_train_norm = scaler.transform(X_train)
X_test_norm = scaler.transform(X_test)
regr = linear_model.LinearRegression()
regr.fit(X_train, y_train)
y_pred=regr.predict(X_test)
r2_score(y_test, y_pred)
```

## Code 2

```
### Code 2
# Load data
data = pd.read_csv("pima_indian_diabetes.csv")
data.head(10)
x = data[["Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]
y = data["Outcome"]

# Learning process
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
scaler = StandardScaler()
scaler.fit(X)
X_norm=scaler.transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred=regr.predict(X_test)
r2_score(y_test, y_pred)
```

## Code 3

```
### Code 3
# Load data
data = pd.read_csv("pima_indian_diabetes.csv")
data.head(10)
x = data[["Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]
y = data["Outcome"]

# Learning process
import numpy as np
from sklearn.linear_model import LinearRegression
reg = LinearRegression().fit(X, y)
y_pred=reg.predict(X)
reg.score(X, y)
r2_score(y, y_pred)
```

## Code 4

```
### Code 4

# Load data
data = pd.read_csv("pima_indian_diabetes.csv")
data.head(10)
x = data[["Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]
y = data["Outcome"]

# Learning process
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline
from sklearn.linear_model import LinearRegression
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
scaler = StandardScaler()
scaler.fit(X_train)
X_train_norm = scaler.transform(X_train)
X_test_norm = scaler.transform(X_test)
degree=3
polyreg=make_pipeline(PolynomialFeatures(degree),LinearRegression())
polyreg.fit(X_train,y_train)
y_pred=regr.predict(X_test)
r2_score(y_test, y_pred)
```

[0.75 points] **Question 6**. Justify whether the following statement is true or false: "The 3-NN rating per vote requires more computational cost (more calculations) in the test stage than in the training stage.

[1.25 points] **Question 7.** Consider the set of two-dimensional observations ($x_1$, $x_2$) shown in Table 2, where the class to which each observation is associated also appears. Explain and determine the process to assign the class for the observation (1,0) using a 5-NN classifier and considering Euclidean distance.
Indicate the coordinates of the training observation that is not part of the 5-NN.

| $x_1$ | $x_2$ | Class |
|---|---|---|
| 0 | 1 | A |
| -1 | 1 | B |
| 2 | 1 | A |
| 0 | 2 | A |
| 1 | -1 | B |
| 0 | -1 | B |
| 0 | 0 | A |

Table 2. Set of training samples and class to which each one belongs.

[1 point] **Question 8**. Regarding decision trees:

A) Indicate and justify which parameters should be considered to obtain a more adequate sub-tree. Justify which procedure should be followed to obtain a less complex tree. Identify and justify clearly the steps followed and the result obtained.
B) Justify if results change after considering a normalization process.