MLSM 2022 - Quiz 1
06/09/2021 - 14/09/2021 at 23:59

Questions based on lecture 1

(1) (1 pt.) Which claim is true?
   (a) The test data for a machine learning model is usually assumed to be drawn from the same known distribution D that generated the training data
   (b) Consistent hypothesis correctly classifies all the training samples
   (c) Version space is the space of all the hypotheses of the hypothesis class

(2) (1 pt.) Which claim is true?
   (a) If a ROC curve shows a diagonal line ($x = y$), this means that the classifier performs perfectly
   (b) To obtain the ROC curve, the decisions of a binary classifier (classes -1 and 1) are analysed with various different thresholds θ, with decision function f as

$$f(x) = \begin{cases} -1 & \text{if } P(y = 1|x) < \theta \\ 1 & \text{otherwise} \end{cases}$$

   (c) A random classifier plotted in a ROC curve has AUC=1.

(3) (1 pt.) You are working in an engineering firm, building a machine learning system whose goal is to raise an alert for a faulty equipment based on some measurements. As the equipment is related to people's safety, is very important that no faulty equipment (=positive class) get past this model. However if some perfect products (=negative class) happen to be flagged as faulty, they can be tested further and confirmed to be in working order without too much cost.

   If the precision of the system is 99.99%, how would you assess the machine learning model? (Note: recall of the next exercise is not known.)
   (a) This is a good model and performs very well for the task.
   (b) You are unsure if the model is good or not: this metric does not give you enough information to make an informed assessment.
   (c) The model is not performing well.

(4) (1 pt.) Continuing with the setting of the previous question, if instead you obtain recall of 99.99% (without knowing the precision), how would you assess the model then? Assume that accuracy score is also high.
   (a) This is a good model and performs very well for the task.
   (b) You are unsure if the model is good or not: this metric does not tell you enough information to make an informed assessment.
   (c) The model is not performing well.

(5) (1 pt.) [*Programming exercise*] Consider the diabetes dataset loaded and split into training and test sets in the provided python code.

   Fit a linear regression model to the training data without considering bias (intercept). What is the root mean squared error on the test data?
   (a) 54.58
   (b) 166.14
   (c) 27601.29

```
import numpy as np
from sklearn.datasets import load_diabetes
from sklearn.metrics import mean_squared_error

# load the data
X, y = load_diabetes(return_X_y=True)
print(X.shape, y.shape)

# division into training and testing
np.random.seed(42)
order = np.random.permutation(len(y))
tst = np.sort(order[:221])
tr = np.sort(order[221:])

Xtr = X[tr, :]
Xtst = X[tst, :]
Ytr = y[tr]
Ytst = y[tst]

# assume that the test data is not known during the training stage
```