MLSM 2021 - Quiz 3
08/10/2021 - 20/10/2021 at 23:59

Questions based on Lecture 4 and 5

(1) (1.0 pt.)

Let x be chosen from the interval $[-1, +1]$, and the labels, y, from the set $\{0, 1\}$. We are given the following conditional probabilities for all x and y:

$$Pr(1|x) = \begin{cases} +x+1 & x \in [-1, 0] \\ -x+1 & x \in [0, +1] \end{cases}$$
$$Pr(0|x) = \begin{cases} -x & x \in [-1, 0] \\ +x & x \in [0, +1] \end{cases}$$

The question: what is the Bayes error of the Bayes classifier relating to this model if x is uniformly distributed on $[-1, +1]$?

Hint: You might solve the problem by drawing a plot representing the functions of the conditional probabilities.

(1) 0.75
(2) 0.5
(3) 0.3
(4) 0.25

(2) (1.0 pt.)

In applying the Perceptron algorithm on a data set, $\{(\mathbf{X}_i, y_i\}_{i=1}^m \; y_i = -1, +1, \forall i$, we might scale all input vectors $\{\mathbf{x}_i\}_{i=1}^m$ with a positive constant $\lambda > 0$. This issue arises when the input vectors are normalized in a learning problem. What is the effect of this scaling? Assume that the two classes are linearly separable, and no bias term is included into the predictor function.

Select that answer which is true if this kind of scaling is applied!

(1) The weigh vector $\mathbf{w}$ is the same when the algorithm is stopped.
(2) In the Novikoff's Theorem, the number of expected iterations t does not depend on the scaling. It is assumed that the margin γ is the largest value satisfying $y_i \mathbf{w}_* \mathbf{x}_i \geq \gamma$ for all $i = 1, \ldots, m$.
(3) In the Novikoff's Theorem, the number of expected iterations t increases when the λ increases.
(4) In the Novikoff's Theorem, the number of expected iterations t decreases when the λ increases.

(3) (2.0 pt.)

Let the stochastic gradient algorithm for Logistic Regression be applied to find the best classifier on the Breast Cancer dataset of the Sklearn package. That algorithm is presented in Lecture 5. In the learning 5-fold cross validation needs to be used on the data. To select the folds, the KFold method of the Sklearn should be used with the parameters shown in the program example below. The labels of the Breast Cancer dataset are of $\{0, 1\}$ which need to be converted into $\{-1, +1\}$. Normalize the rows of the input matrix to have the L_2 norm to be equal to 1. *Without that normalization overflow error can occur in the exponential function!*

The training examples are processed in the order of the original data file, within each fold. The learning parameters, number of iteration and the learning speed, need to be chosen as shown in the program example. The implementation might start with these lines:

```
import numpy as np
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import KFold

# load the data
X, y = load_breast_cancer(return_X_y=True)  ## X input, y output
## to convert the {0,1} output into {-1,+1}
y = 2*y - 1

## learning parameters
nitermax = 50  ## maximum iteration
eta = 0.1      ## learniung speed

nfold = 5         ## number of folds

## to split the data into 5-folds we need
cselection = KFold(n_splits=nfold, random_state=None, shuffle=False)
```

The task is to run the Logistic Regression algorithm on the 5-fold cross-validation, and compute the F1-score on the corresponding test sets. In selecting the training and the test sets you might follow the example code of the KFold function provided in the Sklearn.

The question: what is the average F1 score computed on the 5-folds? Round the numbers up to 2 decimals, and take the closest one.

Be aware, the Logistic Regression algorithm of Lecture 5 is not the same which is implemented in the Sklearn. use that version which is presented in the Lecture!

(1) 0.89
(2) 0.95
(3) 0.99
(4) 0.70

(4) (1.0 pt.)

In the previous question (Question 3), where the stochastic gradient algorithm is applied for the Logistic Regression, scale each of the input variables, columns, to have the maximum absolute value equal to 1. Repeat the same learning procedure of Question 3. In each fold also compute the maximum functional margin that the training can achieve.

The question: what is the average F1-score after scaling the input variables, columns, and what is the average functional margin? The averages are computed on all folds. Similarly to Question 3 round the numbers and find the closest case.

(1) $0.70, 29.30$
(2) $0.99, 45.10$
(3) $0.85, 20.90$
(4) $0.95, 23.20$