3 Classification

Exercise 3.1

Which of the following is an example of *qualitative variable*?

1. Height
2. Age
3. Speed
4. Color

Provide a method to convert the qualitative ones into quantitative one, without introducing further structure over the data.

Exercise 3.2

Consider the following code lines in `Python`:

```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
dataset = pd.read_csv(url, names=names)

x = dataset[['sepal-length']].values
t = dataset['class'].values == 'Iris-setosa'
phi = zscore(np.hstack((x, x ** 2)))

lin_model = sm.OLS(t, x).fit()
qua_model = sm.OLS(t, phi).fit()

if qua_model.rsquared_adj > lin_model.rsquared_adj:
  y = lin_model.predict(x)
else:
  y = qua_model.predict(phi)
```

1. Describe the process and purpose of what is implemented in this snippet.
2. Tell if the method is sound or if it is necessary to modify the procedure to follow the classic ML guideline regarding this method.

Exercise 3.3

Suppose we collect data for a group of workers with variables hours spent working x_1, number of completed projects x_2 and receive a bonus t. We fit a logistic regression and produce estimated coefficients: $w_0 = -6$, $w_1 = 0.05$ and $w_2 = 1$.

Estimate the probability that a worker who worked for 40h and completed 3.5 projects gets an bonus.

How many hours would that worker need to spend working to have a 50% chance of getting an bonus?

Do you think that values of z in $\sigma(z)$ lower than -6 make sense in this problem? Why?

Exercise 3.4

Suppose you have trained a logistic regression classifier l on a dataset $Z = \{(\mathbf{x}_n, t_n)\}_n$ and the output corresponding to observation $\mathbf{x}_n$ is $\hat{y}_n$. Currently, you predict class 1 if $\hat{y}_n > \tau$, and predict 0 if $\hat{y}_n < \tau$, with $\tau = 0.5$. Suppose you increase the threshold to $\tau = 0.8$ getting a new classifier l_{new}. Which of the following are true? Check all that apply and provide a motivation.

1. The number of samples x_n from Z classified as positive instance will decrease if we use l_{new} instead of l.
2. The number of samples x_n from a test dataset Z_{test} classified as positive instance will decrease if we use l_{new} instead of l.
3. The classifier l_{new} is likely to have a higher accuracy.
4. The classification error over Z might decrease by using l_{new} instead of l.

* Exercise 3.5

Derive for logistic regression, the gradient descent update for a batch of K samples.

Do we have assurance about converge to the optimum?

Exercise 3.6

Tell if the following statement about the perceptron algorithm for classification are true or false.

1. Shuffling the initial data influences the perceptron optimization procedure;
2. We are guaranteed that, during the learning phase, the perceptron loss function

is decreasing over time;

3. There exists a unique solution to the minimization of the perceptron loss;
4. The choice of a proper learning rate α might speed up the learning process;
5. The solution of the Logistic regression and the one of the perceptron always coincide.

Motivate your answer.

Exercise 3.7

Consider a classification problem having more than two classes. Propose a method to deal with multiple classes in each one of the following methods:

1. K-Nearest Neighbors;
2. Näive Bayes;
3. Linear regression;
4. Logistic regression;
5. Perceptron.

Exercise 3.8

Tell if the following statements are true or false and motivate your answers.

1. The relationship between the input $\mathbf{x}$ and the estimated output class $y(\mathbf{x})$ induced by a generalized linear model used for classification is linear;
2. The solution of a classification problem using *discriminant function* provides a probability distribution of a generic input $\mathbf{x}$ to belong to a class C_k;
3. Both the Logistic regression and the perceptron use the same loss function to learn the boundary between the classes;
4. Both the Logistic regression and the perceptron use the same updating rule to learn the boundary between the classes.

Exercise 3.9

Given the following dataset:

$$\mathbf{x}_1 = (2,3,4)^\top,\ y_1 = 1 \qquad \mathbf{x}_2 = (0,1,2)^\top,\ y_2 = 0$$
$$\mathbf{x}_3 = (1,2,5)^\top,\ y_2 = 1 \qquad \mathbf{x}_4 = (1,4,3)^\top,\ y_4 = 0$$
$$\mathbf{x}_5 = (0,3,1)^\top,\ y_5 = 0 \qquad \mathbf{x}_6 = (1,2,2)^\top,\ y_6 = 0$$
$$\mathbf{x}_7 = (3,1,4)^\top,\ y_7 = 1 \qquad \mathbf{x}_8 = (4,2,5)^\top,\ y_8 = 1$$
$$\mathbf{x}_9 = (1,3,3)^\top,\ y_9 = 0 \qquad \mathbf{x}_{10} = (1,2,4)^\top,\ y_{10} = 1$$

- Classify the point $\mathbf{x}_{11} = (0,1,2)^\top$ according to a KNN classifier trained on the given dataset with $K = 3$;
- What happens if we use $K = 10$ instead? Do you think it is a good idea to choose such a parameter (hint: two pros or two cons);
- Suggest a technique to set the parameter K.

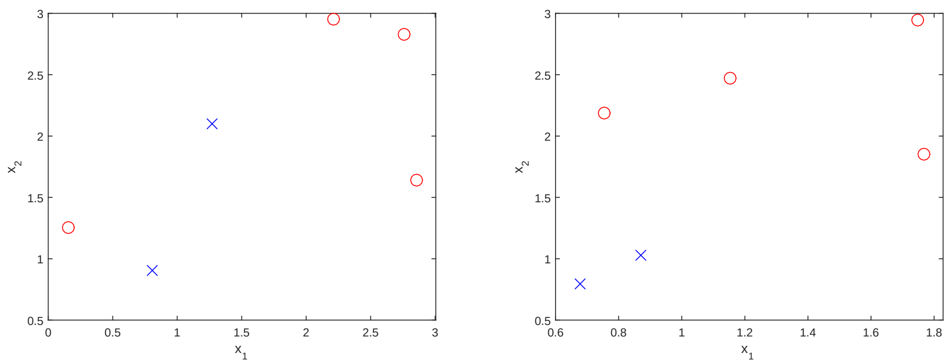
Exercise 3.10

Consider a binary classifier trained on a dataset made of $N = 100$ samples.

1. Suppose that the Precision = 0.25 and the F1 = 0.4, compute the Recall.
2. Knowing, in addition, that the Accuracy = 0.85, compute the full confusion matrix.
3. In which circumstances the Accuracy is not a reliable index to assess the quality of the trained model?

Exercise 3.11

Consider the following datasets:



and consider the online stochastic gradient descend algorithm to train a perceptron.

Does the learning procedure terminates? If so, how many steps we require to reach convergence? Provide motivations for your answers.

What about the Logistic regression?

Exercise 3.12

Starting from the formula of the softmax classifier for k classes:

$$y_k(\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\sum_{j=1}^K \exp(\mathbf{w}_j^T \mathbf{x})}$$

derive the formula for the sigmoid logistic regression parameter $\mathbf{w}$ for the two-class problem.

Assume that the estimated parameter is $\mathbf{w} = (4, 2, 1)^\top$ and the input vector is of the form $\mathbf{x} = (x_1, x_2, 1)^\top$. Draw the boundary of the logistic regression in the input space and in the parameter space.

Exercise 3.13

Consider one at a time the following characteristics for an ML problem:

1. Large dataset (big data scenario);
2. Embedded system;
3. Prior information on data distribution;
4. Learning in a Real-time scenario;
5. Reduced computational capabilities.

Provide motivations for the use of either a parametric or non-parametric method in the above situations.

Exercise 3.14

Consider the following dataset to implement a spam filter function:

"pills"	"fee"	"kittens"	Url Presence	"PoliMi"	spam
0	1	0	0	1	0
0	0	1	1	0	0
0	0	1	0	0	0
0	0	1	0	1	0
0	0	0	0	0	0
1	1	0	0	1	1
0	1	0	1	0	1
1	0	0	1	0	1

where we enumerate the presence of specific word or of an URL in 8 different e-mails and the corresponding inclusion in the spam or non-spam class.

1. Estimate a Naïve Bayes classifier, choosing the proper distributions for the classes priors and the feature posteriors.
2. Predict the probability of the following samples to belong to the spam and no-spam classes.

"pills"	"fee"	"kittens"	Url Presence	"PoliMi"
1	1	0	1	0
0	1	1	0	1

Exercise 3.15

Consider the following snippet of code:

```
1 X = zscore(dataset[['sepal-length', 'sepal-width']].values)
2 t = dataset['class'].values == 'Iris-setosa'
3 X, t = shuffle(X, t, random_state=0)
4 
5 w = np.ones(3)
6 for i, (x_i, t_i) in enumerate(zip(X, t)):
7   ext_x = np.concatenate([np.ones(1), x_i.flatten()])
8   if np.sign(w.dot(ext_x)) != t_i:
9       w = w + ext_x * t_i
```

1. Describe the procedure presented above. What is the purpose of such a procedure? Which problem it solves?
2. Tell if the procedure above is correct and, in the case it is not, propose a modification to fix it.
3. Do you think that Line 3 is fundamental for this procedure? Can you describe an example where it can be removed and motivate why it is not useful in such a case?

Exercise 3.16

Consider a binary perceptron classifier defined by parameters $w = [2, 1, 1]^\top$ with features vector $\phi([a, b]) = [a, b, ab]^\top$). Answer to the following questions related to the perceptron algorithm. Provide full calculations and clear motivations.

1. Given a new data point $(x, t) = ([a, b], t)$, explain the procedure you would follow to decide whether the classifier should be retrained.
2. Consider the data point $(x_1, t_1) = ([1, 2], +1)$. Update the classifier with the perceptron algorithm ($\alpha = 1$).
3. Consider the data point $(x_2, t_2) = ([-1, -2], +1)$. Update the classifier with the perceptron algorithm ($\alpha = 1$).
4. After the previous updates to the classifier, can we say the retrain procedure is completed?