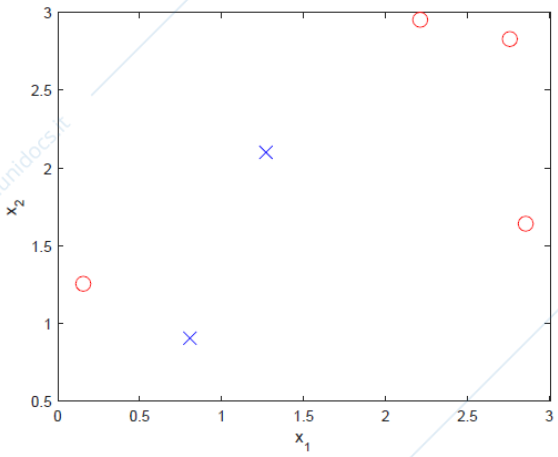
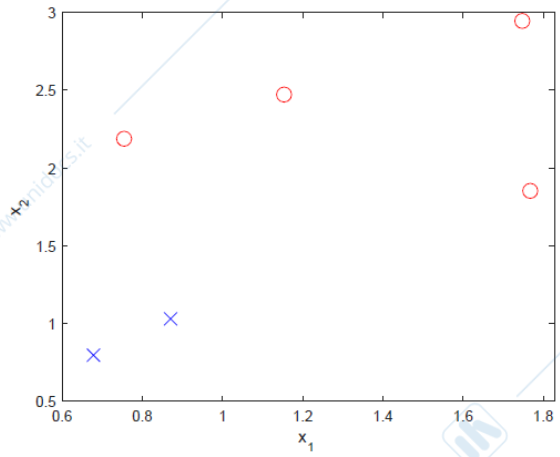
Linear Classification - Exercises

Exam: 04/07/2019

1. [Linear classification]

Consider two different datasets (see the left and right figures below) and the PLA algorithm to train a perceptron:





1.1 Describe the perceptron model.

1.2 Under which assumptions the PLA is guaranteed to converge? Provide a formal proof of convergence in such conditions.

1.3 Does the learning procedure terminate in the two cases illustrated in the figures?

1.4 If so, can we say how many steps we require to reach convergence?

[Solutions]

1.1 See Chapter 3 of the lecture notes.

1.2 The perceptron learning algorithm is guaranteed to converge in the case it exists a linear separation surface. In such a case, we are able to reduce the classification error to zero, otherwise the optimization procedure does not stop. See Chapter 3 of the lecture notes for the mathematical proof of convergence.

1.3 For the reasons discussed above, in the first case (left) we are sure the algorithm does not converge, while in the second case (right) the online procedure will eventually converge.

1.4 We do not have any assurance about the convergence rate, since it depends on the starting point for the parameter and on the ordering of the points we consider for training.

Linear Regression - Exercises

Exam: 04/02/2019

1. **[Linear regression]**
   Consider a simplified learning scenario. Assume that the input dimension is 1, the data consists of 2 points $\{x_1, x_2\}$ and the target function is $f(x) = x^2$. We want to compute the constant model $h(x) = H$ that best fits the given points.

1.1 Formulate and solve the learning problem via linear regression (i.e., write the cost function and the solution of the least squares formulas).

1.2 What is the average model $\bar{h}$ (with respect to all possible experiments) when the points $\{x_1, x_2\}$ are drawn independently from a normal distribution with mean 0 and variance 1?

[Solutions]

1.1 The in-sample error for linear regression (to be used as a cost function for learning) is

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N}(x_n^2 - h)^2 = \frac{1}{2}(x_1^2 - h)^2 + \frac{1}{2}(x_2^2 - h)^2$$

The least squares formula is given by setting the derivative of $E_{in}(h)$

$$\frac{dE_{in}(h)}{dh} = -(x_1^2 - h) - (x_2^2 - h) = 2h - x_1^2 - x_2^2$$

to zero. The optimal model turns out to be $h(x) = H = \frac{x_1^2 + x_2^2}{2}$ (reasonably, the best constant H fitting the data is the mean of the two measured points).

1.2 The average model $\bar{h}$ is, by definition,

$$\bar{h} = \mathbb{E}_x[h(x)] = \mathbb{E}_x\left[\frac{x_1^2 + x_2^2}{2}\right] = \frac{1}{2}\mathbb{E}_x[x_1^2] + \frac{1}{2}\mathbb{E}_x[x_2^2] = 1$$

Logistic Regression - Exercises

Exam: 04/02/2019

1. [Logistic regression]

You are working on a spam classification system using logistic regression. "Spam" is a positive class ($y = +1$) and "Not spam" is the negative class ($y = -1$).

1.1 Write the expression of the logistic regression model. How such a model can be used for classification?

1.2 You have trained your classifier with $N = 1000$ samples. The in-sample confusion matrix is:

	Actual Class: +1	*Actual Class: -1*
Predicted Class: +1	85	890
Predicted Class: -1	15	10

Compute the classifier precision, recall and F1-score.

1.3 Would you be satisfied by such results? Why? What would you do to try to improve the performance?

[Solutions]

1.1 See Chapter 6 of the lecture notes for the expression of the logistic regression model. Since the output of such a model represents the probability that a mail is classified as "Spam", logistic regression can be used for classification by simply setting a threshold ε on the output. If the output is greater than or equal to ε, the mail is classified as "Spam". Otherwise, it is classified as "Not spam".

1.2 Precision ("how many predicted spam emails correspond to real spam?"):

$$PPV = \frac{85}{85 + 890} \cong 0.087$$

Recall ("how many real spam emails are selected as spam?"):

$$TPR = \frac{85}{85 + 15} = 0.85$$

F1-Score (harmonic mean of PPV and TPR):

$$F1 = 2\frac{PPV \cdot TPR}{PPV + TPR} \cong 2\frac{0.07395}{0.937} \cong 0.158$$

1.3 F1 is much lower than 1 (there is a lot of false positive predictions, see also PPV), so we should not be satisfied by such a result. The reason is probably the fact that the number of spam emails in the training set is $85 + 15 = 100$, while the number of "Not spam" emails is $890 + 10 = 900$ (the information in the dataset is not well balanced). Possible solutions to such a problem could be, e.g., retraining the model using a subset of (better balanced) data or – if possible – including more spam emails in the dataset.

Exam: 13/06/2019

1. [Logistic regression]

Suppose we collect data for a group of workers with the following variables: hours spent working x_1 and number of completed projects x_2. The workers may receive a bonus or not (represented by the binary variable y) depending on their overall performance. We fit a logistic regression model using the sigmoidal function

$$\sigma(w_0 + w_1x_1 + w_2x_2) = \frac{e^{w_0+w_1x_1+w_2x_2}}{1 + e^{w_0+w_1x_1+w_2x_2}}$$

and produce the following estimated coefficients: $w_0 = -6,\ w_1 = 0.05,\ w_2 = 1.$

1.1 Estimate the probability that a worker who worked for $40h$ and completed 3.5 projects gets a bonus.
1.2 How many hours would that worker need to spend working to have a 50% chance of getting a bonus?
1.3 Do you think that values of z in the logistic function $\sigma(z)$ lower than -6 make sense in this problem? Why?

Solution:

2.1 The logistic model provides as output the probability of getting a bonus, thus:

$$P(y = 1|x) = \sigma(w_0 + w_1x_1 + w_2x_2)$$

where $x_1 = 40$ and $x_2 = 3.5$. It follows that

$$P(y = 1|x) = \sigma(-6 + 0.05 \cdot 40 + 1 \cdot 3.5) = \sigma(-0.5) = \frac{e^{-0.5}}{1 + e^{-0.5}} = 0.3775\,.$$

2.2 We know that we have 50% chance of getting a bonus when the argument of the sigmoid is equal to zero (in fact, $\sigma(0) = 0.5$), thus we look for $\hat{x}_1$ such that $w_0 + w_1\hat{x}_1 + w_2x_2 = 0$.
By substituting the known parameters and variables,

$$-6 + 0.05 \cdot \hat{x}_1 + 3.5 = 0 \rightarrow \hat{x}_1 = 50h$$

2.3 Since all the considered variables as well as the parameters w_1 and w_2 are positive, it makes only sense to consider values of z greater than -6.

k-NN - Exercises

Exam: 11/01/2019

1. [k-NN]

A k-NN model is being used in a real-estate company to predict house prices.

1.1 Describe briefly the k-NN classifier and list the main differences with the perceptron.

1.2 Discuss how k should be selected depending on the size of the dataset.

1.3 A training set was used to generate a specific k-NN model with k $= 5$. The unseen observation below is presented to the model.

# of bedrooms	# of bathrooms	Size [m2]	Garage	House Price [€]
2	2	167	No	?

The k-NN model determines the observations in the table below from the training set to be the most similar. What would be the predicted house price value?

# of bedrooms	# of bathrooms	Size [m2]	Garage	House Price [€]
2	2	168	No	355000
2	2	142	Yes	352000
2	3	154	No	349000
3	2	202	No	356000
3	2	142	No	351000

Solutions

1.1 See Chapter 4 of the lecture notes for the theory of k-nearest neighboring.

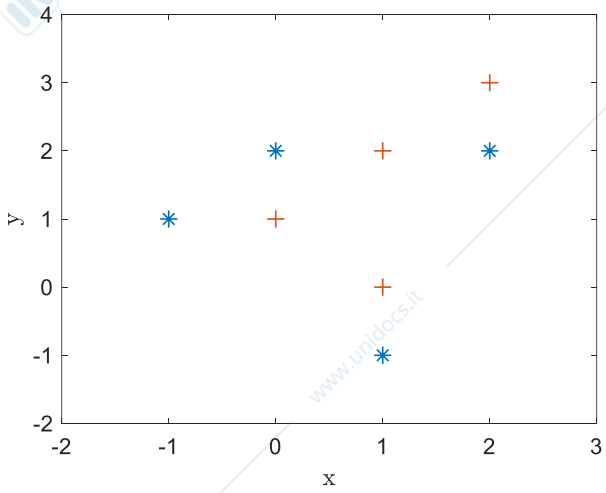
1.2 See again Chapter 4.

1.3 Since $k = 5$, according to the theory, the prediction will be the average of the 5 most similar (i.e., closest) points. In this case, given the table, the predicted house price will be the mean of the house prices, that is 352600 €.

Exam: 11/09/2019

2. [Non-parametric classification]

Suppose you are given a set of data where x and y are the 2 input variables and Class is the dependent variable, which can be either * or +. The scatter plot below shows the data in the regressor space.



2.1. You want to predict the class of a new data point with coordinates $x = 1$ and $y = 1$ via the K-NN method using $K = 7$ and the Euclidean distance. Which class does this data point belong to?

2.2. Does the previous answer change if you use 3-NN instead of 7-KNN?

2.3. Which of the following statement is likely to be true as $\frac{1}{K}$ moves above 1?

A. The prediction error will increase;

B. The prediction error will decrease;

C. It does not make sense to have $\frac{1}{K} > 1$;

D. The prediction error will converge to a constant value.

2.4. Describe the K-NN prediction algorithm by also specifying under which assumptions it can be used.

2.5. Describe one practical tool to assess the prediction performance of a classifier.

Solution

2.1 The new point will be classified as Class *, because there are 4 * and 3 + points in the nearest circle.

2.2 All the 3 nearest points are of Class +, so this point will be classified as +.

2.3 The K-NN classifier can only contemplate integer values for the parameter K, thus it does not make sense to decrease K below 1. The only true statement is the third one.

2.4 See Chapter 4 of the lecture notes.

2.5 See Chapter 4 of the lecture notes.

K-Means - Exercises

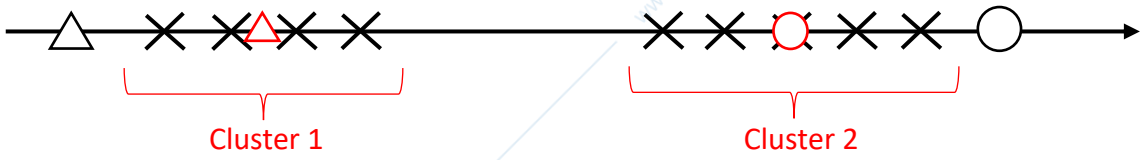
Exam: 11/01/2019

1. [Clustering]

1.1 Define the clustering problem. Is it a supervised or an unsupervised learning problem? Why?

1.2 Introduce the cost function for the K-means clustering problem and describe the Lloyd iterative algorithm.

1.3 Mark approximately in the graph below the solution (clusters and centers) found by Lloyd algorithm for the 2-cluster (K = 2) problem, when the data ($x_i \in \mathbb{R}$) are the crosses in the figure below and the algorithm is initialized with center values indicated with the triangle (cluster 1) and the circle (cluster 2).



Solutions

1.1 See the introductory slides for the definitions of supervised/unsupervised learning and Chapter 12 of the lecture notes for the theory of K-means clustering.

1.2 See Chapter 12 of the lecture notes for the theory of K-means clustering and the Lloyd algorithm.

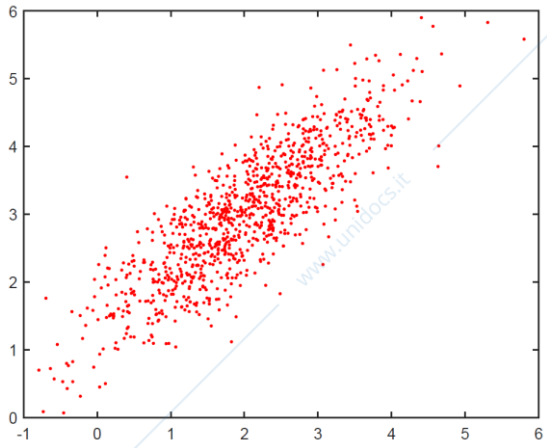
1.3 The Lloyd algorithm will find the two clusters underbraced in red in the figure. The centroids are moved to the centers (see the small triangle and circle in red) in only one iteration, then the algorithm stops.

Data Pre-Processing - Exercises

Exam: 13/06/2019

1. [Data pre-processing]

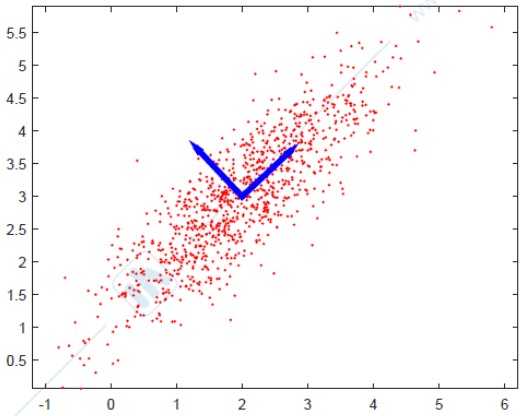
Considering the following dataset.



1.1. Draw the direction of the principal components and say whether the dimensionality of the data could be reduced.
1.2. Apart from PCA, how can data be preprocessed in order to be best prepared for learning?

Solution

1.1 The computed principal components are:



The variance of the data along the first dimension is larger than that along the second, thus a reduced (1-D) feature space could be considered, depending on the specific problem.

1.2 See Chapter 11 of the lecture notes about data pre-processing.