3 Classification

Exercise 3.1

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

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

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

Exercise 3.2

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 3.3

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

Do we have assurance about converge to the optimum?

Exercise 3.4

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.

Motivate your answer.

Exercise 3.5

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 ($y = 0$). You have trained your classifier and there are $N = 1000$ samples. The confusion matrix is:

	Actual Class: 1	Actual Class: 0
Predicted Class: 1	85	890
Predicted Class: 0	15	10

What is the classifier recall? What about the $F1$ score? What would you try to improve in such a system? Should we aim at solving a specific issue?

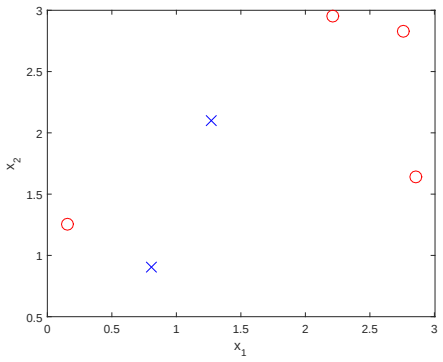
Exercise 3.6

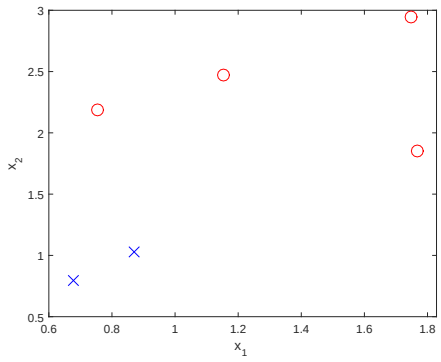
Which of the following is NOT a linear function in x:

1. $f(x) = a + b^2x$;
2. $\delta_k(x) = \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} + \log(\pi)$;
3. $logit(P(y = 1|x))$ where $P(y = 1|x)$ is a logistic regression;
4. $P(y = 1|x)$ from logistic regression;
5. $g(x) = \frac{x-1}{x+1}$;
6. $h(x) = \frac{x^2-1}{x+1}$.

Exercise 3.7

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 3.8

Starting from the formula of the softmax classifier:

$$y_k(x) = \frac{\exp(w_k^T x)}{\sum_j \exp(w_j^T x)},$$

derive the formula for the sigmoid logistic regression for the two classes problem.

Exercise 3.9

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.

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

Exercise 3.10

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

1. Naive Bayes;
2. Perceptron;
3. Logistic regression;
4. K-NN.

Motivate your answers.

Exercise 3.11

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 Naive 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

Answers

Answer of exercise 3.1

1. Quantitative variable: we are able to order heights and they usually are in a finite set (in $\{50, \dots, 230\}$ cm);
2. Quantitative variable: they are ordered values, all of them being integer numbers;
3. Quantitative variable: this variable takes values in the real numbers;
4. Qualitative variable: since we do not have an ordering over the colours (in general cases), we need to convert the set of the available colours into binary variables.

Given the set C ($|C| = p$) of all possible colours we would like to consider we create a new variable for each colour c. This variable for colour c is equal to one in the case a sample x_i has colour $c_i = c$ and zero otherwise. Clearly we need to create p new variables and for each sample, of which only one is equal to one.

Answer of exercise 3.2

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

$$P(t = 1|\mathbf{x}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$$

where $x_1 = 40$ and $x_2 = 3.5$

$$P(t = 1|\mathbf{x}) = \sigma(-6 + 0.05 \cdot 40 + 1 \cdot 3.5) = sigmoid(-0.5) = 0.3775$$

To have a $\alpha\%$ chance of having a bonus we need to invert the sigmoidal function, while in this case we know that we have 50% chance when the argument of the sigmoid is equal to zero, thus:

$$\begin{aligned} w_0 + w_1 \hat{x} + w_2 x_2 &= 0 \\ -6 + 0.05 \cdot \hat{x} + 3.5 &= 0 \\ \hat{x} = \frac{2.5}{0.05} &= 50 \end{aligned}$$

Since all the considered variables are positive definite, it makes only sense to consider values greater than -6 as predictions.

Answer of exercise 3.4

1. TRUE: the learning procedure is influenced by both the initial parameter we consider and the order we present the data to it.
2. FALSE: we are guaranteed that the error (loss) on the currently considered data does not increase.
3. FALSE: if the data are linearly separable, there is an infinite number of linear boundaries able to provide the same loss performance, which are all equivalent solutions for the perceptron.
4. FALSE: the parameter vector norm does not influence the result of the discrimination, thus the use of a generic parameter $\alpha > 0$ would work.

Answer of exercise 3.5

$$Rec = \frac{tp}{tp + fn} = \frac{85}{85 + 15} = 0.85$$
$$Pre = \frac{tp}{tp + fp} = \frac{85}{85 + 890} = 0.087$$
$$F1 = \frac{2 \cdot Rec \cdot Pre}{Pre + Rec} = 2 \cdot 0.85 \cdot 0.0870.85 + 0.087 \approx 0.16$$

In this case it is more harmful to put some non-spam mails in the spam folder, thus we would rather decrease the number of fp than decreasing the fn one.

Answer of exercise 3.6

1. Linear;
2. Linear;
3. Linear since $logit(P(y = 1|x)) = \mathbf{w}^T\mathbf{x}$;
4. Nonlinear;
5. Nonlinear;
6. Linear since can be simplified, thus is linear (if we exclude the value $x = -1$).

Answer of exercise 3.7

The perceptron learning algorithm is guaranteed to converge in the case it exists a linear separation surface. In this case, we are able to reduce the classification error to zero,

otherwise the optimization procedure does not stop. We do not have any assurance on the convergence rate, since it depends on the starting point for the parameter, and on the ordering of the points we consider for training.

In the first case (left) we are sure it does not converge, while in the second case (right) the online stochastic gradient descend will eventually converge.

Since the loss function for the logistic regression is convex, it has a single minimum point, which is reached by the gradient descent algorithm.

Answer of exercise 3.8

Since we are considering only two classes we have that summation is only over two parameter vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. If we consider class C_1 we may write:

$$\begin{aligned} y_1(x) &= \frac{\exp(\mathbf{w}_1^T x)}{\exp(\mathbf{w}_1^T x) + \exp(\mathbf{w}_2^T x)} \\ &= \frac{\frac{\exp(\mathbf{w}_1^T x)}{\exp(\mathbf{w}_1^T x)}}{\frac{\exp(\mathbf{w}_1^T x)+\exp(\mathbf{w}_2^T x)}{\exp(\mathbf{w}_1^T x)}} \\ &= \frac{1}{1 + \exp[(\mathbf{w}_2 - \mathbf{w}_1)^T x]}. \end{aligned}$$

Similarly for class C_2 we have the same formula with $(\mathbf{w}_1 - \mathbf{w}_2)^T$. Since we are considering a probability distribution it is not necessary to consider both $y_1(x)$ and $y_2(x)$. Indeed, $y_2(x) = 1 - y_1(x)$, which is why we just need to store a single parameter vector $\mathbf{w} = \mathbf{w}_2 - \mathbf{w}_1$ and the formula for the two class classifier has a single parameter vector:

$$y_1(x) = \frac{1}{1 + \exp(\mathbf{w}^T x)}.$$

Answer of exercise 3.9

1. PARAMETRIC: in the case we have a large dataset it is better to have a model which is able to capture the characteristics of the problem than by basing on the entire dataset to provide a prediction.

2. NON-PARAMETRIC: some of the algorithm we considered in machine learning requires computationally expensive training phases. If the entire system has to work on an embedded system, we should either perform the training phase on a different device or use non parametric methods, which does not require a learning phase.

3. PARAMETRIC: an easy way for introducing a-priori information on the dataset we have in a learning methods is to include them in the model. Since non-parametric only are based on data, it is not trivial to include prior knowledge in them.

4. PARAMETRIC/NON-PARAMETRIC: since we want a fast way of performing tasks a non-parametric method is probably a good idea, since it does not require to have a training phase. On the contrary, if we are able to provide an online method for training a parametric method, we could also consider parametric methodologies.

Answer of exercise 3.10

1. The Naive Bayes algorithm natively supports the existence of multiple classes. Indeed, it models the posterior probability of each class given the sample.

2. For K classes, we can consider K 1 vs. all models and give as label the one providing the maximum value for $\mathbf{w}^T\mathbf{x} + w_0$.

3. For K classes, we can consider K 1 vs. all logistic regressions and consider as a class the one providing the highest probability.

4. We might use majority voting to decide the class. We need to carefully choose the way we are breaking ties, since this might be crucial in the case of many classes.

Answer of exercise 3.11

1. The proper models in this case uses Bernoulli variables both as prior distributions and for the posteriors. More specifically we estimate the probabilities with the

MLE, which is the common empirical expected value:

$$
\begin{aligned}
P(C_1) &= \frac{5}{8} & P(C_2) &= \frac{3}{8} \\
P(x_1 = 0 \mid C_1) &= \frac{5}{5} & P(x_1 = 1 \mid C_1) &= \frac{0}{5} \\
P(x_2 = 0 \mid C_1) &= \frac{4}{5} & P(x_2 = 1 \mid C_1) &= \frac{1}{5} \\
P(x_3 = 0 \mid C_1) &= \frac{2}{5} & P(x_3 = 1 \mid C_1) &= \frac{3}{5} \\
P(x_4 = 0 \mid C_1) &= \frac{4}{5} & P(x_4 = 1 \mid C_1) &= \frac{1}{5} \\
P(x_5 = 0 \mid C_1) &= \frac{3}{5} & P(x_5 = 1 \mid C_1) &= \frac{2}{5} \\
P(x_1 = 0 \mid C_2) &= \frac{1}{3} & P(x_1 = 1 \mid C_2) &= \frac{2}{3} \\
P(x_2 = 0 \mid C_2) &= \frac{1}{3} & P(x_2 = 1 \mid C_2) &= \frac{2}{3} \\
P(x_3 = 0 \mid C_2) &= \frac{3}{3} & P(x_3 = 1 \mid C_2) &= \frac{0}{3} \\
P(x_4 = 0 \mid C_2) &= \frac{1}{3} & P(x_4 = 1 \mid C_2) &= \frac{2}{3} \\
P(x_5 = 0 \mid C_2) &= \frac{2}{3} & P(x_5 = 1 \mid C_2) &= \frac{1}{3}
\end{aligned}
$$

2. The probability is the product of the priors and the posteriors of each feature:

   - First sample: $P(C_1|x) \propto P(C_1)P(x_1 = 1 \mid C_1)P(x_2 = 1 \mid C_1)P(x_3 = 0 \mid C_1)P(x_4 = 1 \mid C_1)P(x_5 = 0 \mid C_1) = \frac{5}{8} \cdot 0 \cdots = 0$, therefore it belongs to C_2 for the trained NB model.
   - Second sample: $P(C_2|x) \propto P(C_2)P(x_1 = 0 \mid C_2)P(x_2 = 1 \mid C_2)P(x_3 = 1 \mid C_2)P(x_4 = 0 \mid C_2)P(x_5 = 1 \mid C_2) = \frac{3}{8} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot 0 \ldots = 0$, therefore it belongs to C_1 for the trained NB model.