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?

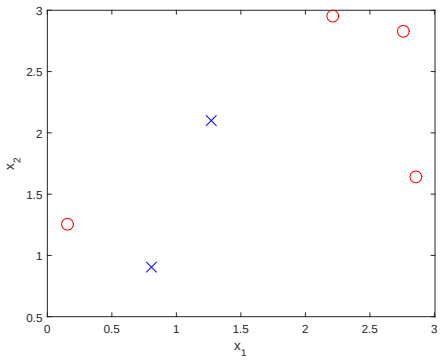
Exercise 3.6

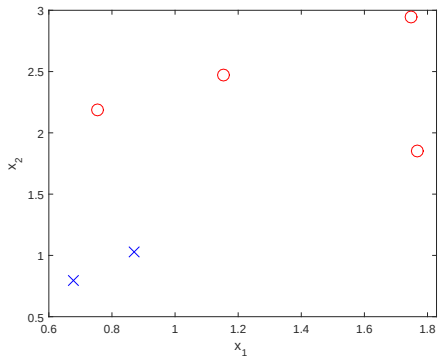
Which of the following is NOT a linear function in x:

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