FIN-407 Machine Learning in Finance

Exercise sheet 1

Assistant: Giuseppe Matera.

Faculty: Elise Gourier.

Question 1

Bayes rule

60% percent of all traders hired by a large financial firm are rated as performing satisfactorily or better in their first year review. Of these, 90% earned a first in financial econometrics. Of the traders who were rated as unsatisfactory, only 20% earned a first in financial econometrics.

(a) What is the probability that a trader is rated as satisfactory or better given they received a first in financial econometrics?

(b) What is the probability that a trader is rated as unsatisfactory given they received a first in financial econometrics?

Question 2

Bayes rule 2

Consider discrete random variables A, B, C that can take values a_1, ... , a_N ; b_1, ... , b_N ; and c_1, ... , c_N, respectively. Which of the following sums to 1?

1. $\sum_{i=1}^{N} P(A = a_i | B = b_2)$
2. $\sum_{i=1}^{N} P(A = a_2 | B = b_i)$

3. $\sum_{i=1}^{N} P(A = a_i | B = b_2, C = c_2)$

4. $\sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{i=1}^{N} P(A = a_i | B = b_j, C = c_k)$

5. $\sum_{j=1}^{N} \sum_{i=1}^{N} P(A = a_i, B = b_j | C = c_2)$

6. $\sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{i=1}^{N} P(A = a_i | B = b_j, C = c_k) P(B = b_j, C = c_k)$

Question 3

Union and intersect

A hedge fund company manages three distinct funds. In any given month, the probability that the return is positive is shown in the following table:

$$\mathbb{P}(r_{1,t} > 0) = 0.55, \quad \mathbb{P}(r_{1,t} > 0 \cup r_{2,t} > 0) = 0.82$$
$$\mathbb{P}(r_{2,t} > 0) = 0.60, \quad \mathbb{P}(r_{1,t} > 0 \cup r_{3,t} > 0) = 0.7525$$
$$\mathbb{P}(r_{3,t} > 0) = 0.45, \quad \mathbb{P}(r_{2,t} > 0 \cup r_{3,t} > 0) = 0.78$$
$$\mathbb{P}(r_{2,t} > 0 \cap r_{3,t} > 0 | r_{1,t} > 0) = 0.20$$

(a) Are the events of "positive returns" pairwise independent?

(b) Are the events of "positive returns" independent?

(c) What is the probability that funds 1 and 2 have positive returns, given that fund 3 has a positive return?

(d) What is the probability that at least one fund will have a positive return in any given month?

Question 4

Conditional probability

At a small high-frequency hedge fund, two competing algorithms produce trades. Algorithm 1 produces 80 trades per second and 5% lose money. Algorithm 2 produces 20 trades per second but only 1% lose money. Given the last trade lost money, what is the probability it was produced by algorithm 1?

Question 1

Bayes rule

60% percent of all traders hired by a large financial firm are rated as performing satisfactorily or better in their first year review. Of these, 90% earned a first in financial econometrics. Of the traders who were rated as unsatisfactory, only 20% earned a first in financial econometrics.

(a) What is the probability that a trader is rated as satisfactory or better given they received a first in financial econometrics?

(b) What is the probability that a trader is rated as unsatisfactory given they received a first in financial econometrics?

(a) $A = \{$ satisfactory or better $\}$ $B = \{$ first in financial econometrics $\}$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{90}{100}\cdot\frac{60}{100}}{\frac{90}{100}\cdot\frac{60}{100} + \frac{20}{100}\cdot\frac{40}{100}} = \frac{90\cdot 60}{90\cdot 60 + 20\cdot 40} = \frac{9\cdot 6}{9\cdot 6 + 2\cdot 4} = \frac{54}{62} = \frac{27}{31}$$

(b) $C = \{$ Unsatisfactory $\}$

$$\mathbb{P}(C|B) = \frac{\mathbb{P}(C \cap B)}{\mathbb{P}(B)} = \frac{\frac{20}{100}\cdot\frac{40}{100}}{\frac{90}{100}\cdot\frac{60}{100} + \frac{20}{100}\cdot\frac{40}{100}} = \frac{8}{62} = \frac{4}{31}$$

$$\left(\mathbb{P}(C|B) = 1 - \mathbb{P}(C^c|B) = 1 - \mathbb{P}(A|B)\right)$$

Question 2

Bayes rule 2

Consider discrete random variables A, B, C that can take values $a_1, \ldots, a_N$; $b_1, \ldots, b_N$; and $c_1, \ldots, c_N$, respectively. Which of the following sums to 1?

1. $\sum_{i=1}^{N} P(A = a_i | B = b_2)$
2. $\sum_{i=1}^{N} P(A = a_2 | B = b_i)$
3. $\sum_{i=1}^{N} P(A = a_i | B = b_2, C = c_2)$
4. $\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} P(A = a_i | B = b_j, C = c_k)$
5. $\sum_{j=1}^{N}\sum_{i=1}^{N} P(A = a_i, B = b_j | C = c_2)$
6. $\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} P(A = a_i | B = b_j, C = c_k) P(B = b_j, C = c_k)$

(1) $\sum_{i=1}^{n} \mathbb{P}(A = a_i) = 1 \Rightarrow \sum_{i=1}^{n} \mathbb{P}(A = a_i | B = b_j) = 1 \quad \forall j = 1, \ldots, N.$ True

(2) Do a coin toss (B take values in $b_1 =$ heads, $b_2 =$ tails). If $B = b_1$ you throw a cubic dice, if $B = b_2$ you throw a dodecahedron. $A = 1$ if the number is a prime, $A = 0$ if it is not.

$$\sum_{i=1}^{2} \mathbb{P}(A = 1 | B = b_i) = \frac{3}{6} + \frac{5}{12} = \frac{6+5}{12} = \frac{11}{12}$$

(3) As 1. True.

(4) 3 Coin toss independent:

$$\sum_{i=1}^{2}\sum_{i=1}^{2}\sum_{i=1}^{2} \mathbb{P}(A = a_i | B = b_j, C = c_k) = 8 \cdot \frac{1}{2} = 4 \quad \text{False.}$$

(5) As 1 and 3. True.

(6) $\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} \mathbb{P}(A=a_i \mid B=b_j, C=c_k)\cdot \mathbb{P}(B=b_j, C=c_k) \overset{\text{Bayes}}{=}$

$= \sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} \mathbb{P}(A=a_i, B=b_j, C=c_k) = 1$

Question 3

Union and intersect

A hedge fund company manages three distinct funds. In any given month, the probability that the return is positive is shown in the following table:

$$\mathbb{P}(r_{1,t}>0)=0.55, \quad \mathbb{P}(r_{1,t}>0 \cup r_{2,t}>0)=0.82$$
$$\mathbb{P}(r_{2,t}>0)=0.60, \quad \mathbb{P}(r_{1,t}>0 \cup r_{3,t}>0)=0.7525$$
$$\mathbb{P}(r_{3,t}>0)=0.45, \quad \mathbb{P}(r_{2,t}>0 \cup r_{3,t}>0)=0.78$$
$$\mathbb{P}(r_{2,t}>0 \cap r_{3,t}>0 | r_{1,t}>0)=0.20$$

(a) Are the events of "positive returns" pairwise independent?

(b) Are the events of "positive returns" independent?

(c) What is the probability that funds 1 and 2 have positive returns, given that fund 3 has a positive return?

(d) What is the probability that at least one fund will have a positive return in any given month?

(a) $\mathbb{P}(r_{1,t}>0)\cdot\mathbb{P}(r_{2,t}>0) = 0,55\cdot 0,60 = 0,33$

$\mathbb{P}(r_{1,t}>0 \cup r_{2,t}>0) = \mathbb{P}(r_{1,t}>0)+\mathbb{P}(r_{2,t}>0)-\mathbb{P}(r_{1,t}>0 \cap r_{2,t}>0)$

$\Rightarrow \mathbb{P}(r_{1,t}>0 \cap r_{2,t}>0) = 0,55+0,60-0,82 = 0,33$. So $r_{1,t}$ $r_{2,t}$ are independent.

The same for the others and we see they're pairwise independent.

(b) $\mathbb{P}(r_{1,t}>0)\cdot\mathbb{P}(r_{2,t}>0)\,\mathbb{P}(r_{3,t}>0) = 0,55\cdot 0,60\cdot 0,45 = 0,1485$

$\mathbb{P}(r_{1,t}>0 \cap r_{2,t}>0 \cap r_{3,t}>0) = \mathbb{P}(r_{2,t}>0 \cap r_{3,t}>0 \mid r_{1,t}>0)\cdot\mathbb{P}(r_{1,t}>0) = 0,20\cdot 0,55 = 0,11$

(Bayes)

(c) $\mathbb{P}(r_{1,t}>0, r_{2,t}>0 \mid r_{3,t}>0) = \frac{\mathbb{P}(r_{1,t}>0, r_{2,t}>0, r_{3,t}>0)}{\mathbb{P}(r_{3,t}>0)} = \frac{0,11}{0,45} = 0,24$

(d) $\mathbb{P}(r_{1,t}>0 \cup r_{2,t}>0 \cup r_{3,t}>0) =$

(exclusion and inclusion)

$= \mathbb{P}(r_{1,t}>0)+\mathbb{P}(r_{2,t}>0)+\mathbb{P}(r_{3,t}>0) - \mathbb{P}(r_{1,t}>0 \cap r_{2,t}>0) - \mathbb{P}(r_{1,t}>0, r_{3,t}>0) - \mathbb{P}(r_{2,t}>0, r_{3,t}>0)$

$+\mathbb{P}(r_{1,t}>0 \cap r_{2,t}>0 \cap r_{3,t}>0) = 0,8625$

Question 4

Conditional probability

At a small high-frequency hedge fund, two competing algorithms produce trades. Algorithm 1 produces 80 trades per second and 5% lose money. Algorithm 2 produces 20 trades per second but only 1% lose money. Given the last trade lost money, what is the probability it was produced by algorithm 1?

$A =$ "last trade lost money" $\quad B =$ "Trade produced by algorithm 1"

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\frac{5}{100} \cdot \frac{80}{100}}{\frac{5}{100} \cdot \frac{80}{100} + \frac{1}{100} \cdot \frac{20}{100}} = \frac{5 \cdot 8}{5 \cdot 8 + 2} = \frac{40}{42} = \frac{20}{21}$$

FIN-407 Machine Learning in Finance

Exercise sheet 2 - Supervised learning 1

Assistant: Giuseppe Matera.

Faculty: Elise Gourier.

Question 1

Statistical learning (*)

(a) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

2. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

3. We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

(b) You will now think of some real-life applications for statistical learning.

1. Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

2. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Question 2

Regression (**)

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\beta_0 = 50, \beta_1 = 20, \beta_2 = 0.07, \beta_3 = 35, \beta_4 = 0.01, \beta_5 = -10$.

(a) Which answer is correct, and why?

1. For a fixed value of IQ and GPA, males earn more on average than females.
2. For a fixed value of IQ and GPA, females earn more on average than males.
3. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
4. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Question 3

Bias-variance trade-off (**)

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would you expect one to be lower than the other, would you expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Same question as in (a) but with the test rather than the training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Same question as in (c) but with the test rather than the training RSS.

Question 4

R^2 (***)

Show that in the case of simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

Question 5

Cross-validation (*)

(a) Briefly describe the k-fold cross-validation, validation set and leave-one-out cross-validation approaches. What are the advantages and disadvantages of k-fold cross-validation relative to the other two approaches?

Question 1

Statistical learning (*)

(a) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

2. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

3. We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

(b) You will now think of some real-life applications for statistical learning.

1. Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

2. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

(a) 1. Regression problem (Regression prob : continuous values , Classification : classify , ex. True/false)

Inference (Inference: understand something from the data without make a prediction)

$n = 500$ $p = 3$ (n = number of data , p = features to find the target)

2. Classification;

Prediction;

$n = 20$ and $p = 13$

3. Regression;

Prediction;

$n = 52$ $p = 3$

(b) Classification: Colour of hair (↑ Inference) , Colour of a pixel (↑ prediction) , Illness (stage of a cancer) (↑ prediction)

Regression: stock price (↑ prediction) , average height (↑ inference) , salary (↑ inference)

(c) Flexible: big number of data , follows better the bias of data (not linearity), it could lead to overfitting

Less flexible: few number of data, it could be a problem when there is a lot of bias

Question 2

Regression (**)

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\beta_0 = 50, \beta_1 = 20, \beta_2 = 0.07, \beta_3 = 35, \beta_4 = 0.01, \beta_5 = -10$.

(a) Which answer is correct, and why?

1. For a fixed value of IQ and GPA, males earn more on average than females.
2. For a fixed value of IQ and GPA, females earn more on average than males.
3. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
4. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

(a) $Salary = 50 + 20 \cdot GPA + 0.07 \cdot IQ + 35 \cdot GENDER + 0.01 \cdot GPA \times IQ - 10 \cdot GPA \times GENDER$

$X = \mathbb{E}[Salary \mid IQ = K_2, GPA = K_1, MALES] = 50 + 20K_1 + 0.07K_2 + 0.01K_1K_2$

$Y = \mathbb{E}[Salary \mid IQ = K_2, GPA = K_1, Female] = 50 + 20K_1 + 0.07K_2 + 35 + 0.01K_1K_2 - 10K_1$

So $X - Y = 10K_1 - 35$ so if $K_1 > \frac{35}{10}$ the (c) is true.

(b) $\mathbb{E}[Salary \mid IQ = 110, GPA = 4.0, Female] = 50 + 20 \cdot 4 + 0.07 \cdot 110 + 35 + 0.01 \cdot 110 \cdot 4 - 40$

$= 137.1$

(c) False; we must examine the p-value of the regression coefficient if the interaction is statistically significant or not.

Question 3

Bias-variance trade-off (**)

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would you expect one to be lower than the other, would you expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Same question as in (a) but with the test rather than the training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Same question as in (c) but with the test rather than the training RSS.

(a) I expect the RSS for the cubic regression in the training set to be lower than the RSS for the linear regression. Indeed, in the training set you just want to fit better the data you have and so the cubic approximate better than the linear, in particular cubic implies linear setting $\beta_2 = \beta_3 = 0$

(b) On the other hand, in the test you have chosen the parameters $\beta_0, \beta_1, \beta_2, \beta_3$ using the training set and so since the real correlation between Y and X is linear, the cubic regression will overfit and so I expect RSS for linear regression to be lower than for cubic regression.

(c) The same as (a), when I consider the training set is always better a higher degree of the polynomial.

(d) For the reason already said, we can't say anything.

Question 4

R^2 (***)

Show that in the case of simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

$$RSS \underset{\text{def}}{=} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \underset{\text{linear regression}}{=} \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

$$TSS \underset{\text{def}}{=} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \left(\sum_{i=1}^{n} \left(y_i - \frac{\sum_{i=1}^{n} y_i}{n} \right)^2 \right)_{\bar{y}=0} = \sum_{i=1}^{n} y_i^2$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2}{\sum_{i=1}^{n} y_i^2}$$

$$\rho(X,Y)^2 = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2} \underset{\bar{x} = \bar{y} = 0}{=} \frac{\left(\sum_{i=1}^{n} x_i y_i \right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

$$\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = \left[\begin{pmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{pmatrix}\begin{pmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix}\right]^{-1}\begin{pmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \overset{0}{=} 0\\ \sum_{i=1}^n x_i \overset{0}{=} 0 & \sum_{i=1}^n x_i^2\end{pmatrix}^{-1}\begin{pmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix} =$$

$$= \begin{pmatrix}\frac{1}{n} & 0\\ 0 & \frac{1}{\sum_{i=1}^n x_i^2}\end{pmatrix}\begin{pmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix} = \begin{pmatrix}\frac{1}{n} & \cdots & \frac{1}{n}\\ \frac{x_1}{\sum_{i=1}^n x_i^2} & \cdots & \frac{x_n}{\sum_{i=1}^n x_i^2}\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix} =$$

$$= \begin{pmatrix}\left(\sum_{i=1}^n y_i\right)\frac{1}{n}\\ \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\end{pmatrix} = \begin{pmatrix}0\\ \beta_1\end{pmatrix}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n \left(y_i - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} x_i\right)^2}{\sum_{i=1}^n y_i^2} =$$

$$= \frac{\cancel{\sum_{i=1}^n y_i^2} - \cancel{\sum_{i=1}^n y_i^2} - \frac{\left(\sum_{i=1}^n x_i y_i\right)^2}{\left(\sum_{i=1}^n x_i^2\right)^2}\cancel{\sum_{i=1}^n x_i^2} + 2\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^2} =$$

$$= \frac{\left(\sum_{i=1}^n x_i y_i\right)^2}{\sum_{i=1}^n y_i^2 \sum_{i=1}^n x_i^2} = \rho(X,Y)^2$$

√**Question 5**

Cross-validation (*)

(a) Briefly describe the k-fold cross-validation, validation set and leave-one-out cross-validation approaches. What are the advantages and disadvantages of k-fold cross-validation relative to the other two approaches?

(a) • K-fold cross-validation: I divide the dataset in K subset; $D_1, \dots, D_K \subseteq$ dataset, i.e. $\bigcup_{i=1}^K D_i =$ dataset $D_i \cap D_j \neq \emptyset$ se $i \neq j$; $\forall j \in \{1, \dots, K\}$ I consider the test set $= \bigcup_{i \neq j} D_i$ and the test set D_j. I compute the error (that could be for example RSS) obtaining K RSS, $RSS_1, \dots, RSS_K$, and then I take the smaller.

- Validation set: we simply divide the data set in training set and test set.
- leave-one-out cross validation: K-fold validation where K = # dataset.

- Advantages of K-fold validation: all the data are used as training data and test data (vs validation set), cost of computation less then leave-one-out cross validation; more flexible than validation set.
- Disadvantages of K-fold validation: higher cost of computation than validation test, same variability as validation set when K is small.

FIN-407 Machine Learning in Finance

Exercise sheet 2 - Supervised learning (cont'd)

Assistant: Giuseppe Matera.

Faculty: Elise Gourier.

Question 1

Subset selection (*)

(a) Give two approaches to do subset selection.

(b) Are the coefficients unbiased in these methods?

Question 2

QR decomposition (***)

Consider, as in the lecture, a standard linear regression. The dependent variable Y is a random variable that satisfies:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

where $X_1, ..., X_p$ contain the explanatory variables and ϵ is the "error" term.

We are given training data: $(y_1, \boldsymbol{x}_1)$, $(y_2, \boldsymbol{x}_2)$,..., $(y_N, \boldsymbol{x}_N)$.

Assume that $p \leq N$, $\hat{\beta}$ can be obtained by minimizing the residual sum of squares:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 = \min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad , \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & ... & x_{Np} \end{bmatrix}$$

Theorem: QR decomposition. Any set of p linearly independent vectors in $\mathbb{R}^N$ can be decomposed into the product of a matrix $\boldsymbol{Q}$ of p orthonormal vectors in $\mathbb{R}^N$ and an upper-triangular matrix $R \in \mathbb{R}^{p \times p}$.

(a) Decompose $\boldsymbol{X}$ into $\boldsymbol{QR}$ and show that this decomposition avoids calculating $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

(b) You regularize the regression using a ridge regularization. As shown in the lecture, the optimal $\boldsymbol{\beta}$ now satisfies

$$\hat{\boldsymbol{\beta}}^R = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \frac{\lambda}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} \right\}$$

and can be derived as

$$\hat{\boldsymbol{\beta}}^R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Apply the transformation

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}/\sigma \\ \sqrt{\boldsymbol{\Lambda}} \end{pmatrix}, \quad \tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y}/\sigma \\ 0_{p\times 1} \end{pmatrix}$$

where σ is the volatility of the error term and $\boldsymbol{\Lambda} = \frac{1}{\tau^2}\boldsymbol{I}$, $\lambda = \frac{\sigma^2}{2\tau^2}$. Show that the problem reduces to a standard regression.

Question 3

SVD decomposition (***)

Consider a linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

(a) Apply SVD to $\boldsymbol{X}$ and calculate the least square estimate $\hat{\boldsymbol{\beta}}^{LS}$ using this decomposition. Show what this is an attractive way of solving the regression.

(b) Assume that the number of predictors is large compared to the number of observations. Penalize the problem using ridge regression and calculate the ridge estimate $\hat{\boldsymbol{\beta}}^{ridge}$ using SVD on $\boldsymbol{X}$.

(c) Show that the ridge regression computes the coordinates of y with respect to the orthonormal basis U and shrinks these coordinates more when the corresponding singular value is smaller.

Question 4

Regression Tree versus Linear Regression

Figure 1 represents two two-dimensional problems $X_2 = f(X_1)$ (problem 1 is on row 1 and problem 2 on row 2).

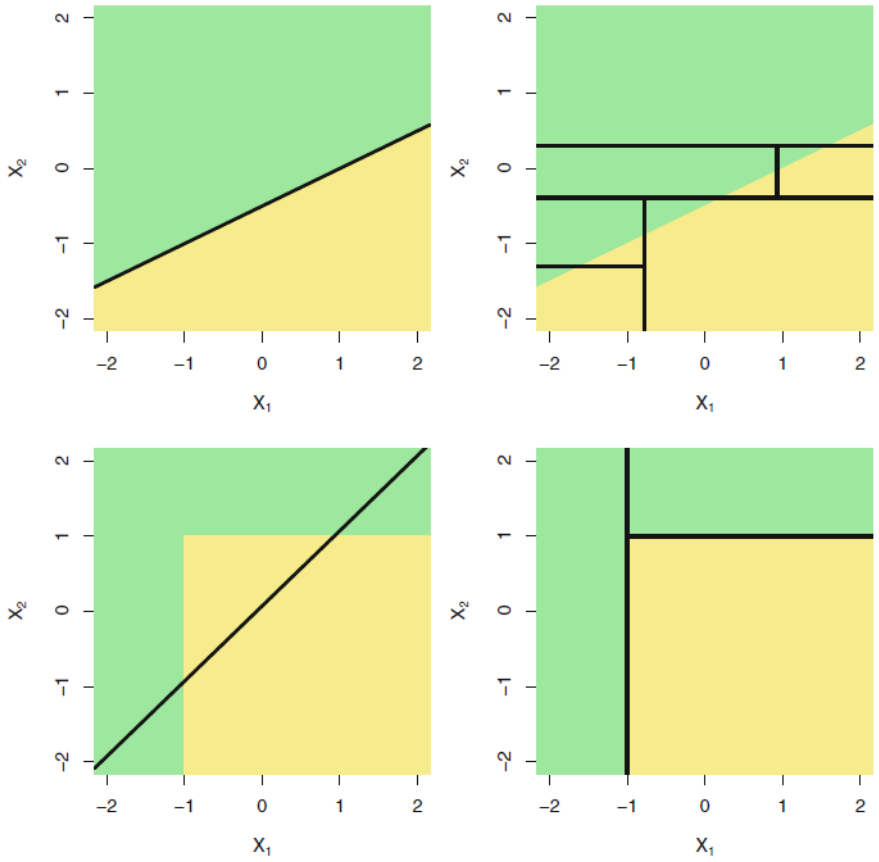


Figure 1: Left column: problem fit by linear regression. Right column: problem fit by regression tree.

(a) Comment on the performance of linear regression versus regression tree. When should you use one or the other?

Question 5

Boosting as gradient descent in prediction space

(a) The first step of a Gradient Boosting Machine (GBM) is to find a base prediction. Show that it makes sense, when the MSE is the loss function, to use the mean of the N observations as base prediction.

(b) Each step, a new tree is estimated using the residuals of the last step. Show that these residuals can be computed as follows:

$$r_{im} = - \left[\frac{\partial L(y_i, \hat{f}(x_i)}{\partial \hat{f}(x_i)} \right]_{\hat{f}(x_i)=\hat{f}_{m-1}(x_i)}$$

where m denotes the index of the iteration and $\hat{f}_{m-1}(x_i)$ denotes the prediction made for observation i at the previous iteration $m-1$.

(c) If the MAE is used as loss function, how do the above results change?

Question 1

Subset selection (*)

(a) Give two approaches to do subset selection.

(b) Are the coefficients unbiased in these methods?

(a) For example:

- Best subset selection: we have p features and fore every $1 \leq j \leq p$ we compute the RSS for every $\binom{p}{j}$ groups that has j features. Then, we chose the group that has the minimum RSS and it will be the "best" group with j features.

- Forward stepwise selection: we start with $j=1$ and as in Best subset selection we choose the best feature. Than we fix this feature, F_1, and we construct a group with F_1 and another feature F_2 that is the "best" group of two where I have F_1. So, at step j we have fix $\{F_1, \dots, F_{j-1}\}$ and we want to add F_j s.t. $\{F_1, \dots, F_{j-1}, F_j\}$ is the best. Remark: we don't have the best j-group but the computation is much lower then "Best solution selection".

- Backward stepwise selection: same as forward but I start with all features and the I remove the feature such $\{F_1, \dots, F_{p-1}\}$ is the best.

(b) There could be bias if the model chosen does not include the true predictors of the model. In Lasso, for example, the coefficients are biased (is not always bad since there is bias-variance trade off)

Question 2

QR decomposition (***)

Consider, as in the lecture, a standard linear regression. The dependent variable Y is a random variable that satisfies:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

where $X_1, ..., X_p$ contain the explanatory variables and ϵ is the "error" term.

We are given training data: $(y_1, \boldsymbol{x}_1)$, $(y_2, \boldsymbol{x}_2)$,..., $(y_N, \boldsymbol{x}_N)$.

Assume that $p \leq N$, $\hat{\beta}$ can be obtained by minimizing the residual sum of squares:

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 = \min_{\beta} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2$$

where

Theorem: QR decomposition. Any set of p linearly independent vectors in $\mathbb{R}^N$ can be decomposed into the product of a matrix $\boldsymbol{Q}$ of p orthonormal vectors in $\mathbb{R}^N$ and an upper-triangular matrix $R \in \mathbb{R}^{p \times p}$.

(a) Decompose $\boldsymbol{X}$ into $\boldsymbol{QR}$ and show that this decomposition avoids calculating $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

(b) You regularize the regression using a ridge regularization. As shown in the lecture, the optimal $\boldsymbol{\beta}$ now satisfies

$$\hat{\boldsymbol{\beta}}^R = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \frac{\lambda}{2} \boldsymbol{\beta}^T \boldsymbol{\beta} \right\}$$

and can be derived as

$$\hat{\boldsymbol{\beta}}^R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Apply the transformation

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}/\sigma \\ \sqrt{\boldsymbol{\Lambda}} \end{pmatrix}, \quad \tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y}/\sigma \\ 0_{p \times 1} \end{pmatrix}$$

where σ is the volatility of the error term and $\boldsymbol{\Lambda} = \frac{1}{\tau^2}\boldsymbol{I}$, $\lambda = \frac{\sigma^2}{2\tau^2}$. Show that the problem reduces to a standard regression.

(a) QR decomposition

$X \in N \times (p+1)$

$\beta \in (p+1) \times 1$

$y \in N \times 1$ $\quad \left(\hat{\beta} = (X^T X)^{-1} y \right)$

$$L(\beta) = \| y - X\beta \|^2 = (y - X\beta)^T (y - X\beta) = y^T y - \underbrace{y^T X \beta}_{\in \mathbb{R}} - \underbrace{\beta^T X^T y}_{\in \mathbb{R}} + \overbrace{\beta^T X^T X \beta}^{\|X\beta\|^2}$$

$L(\beta)$: loss function

$\downarrow$ are the same

$$\frac{\partial L}{\partial \beta} = -2 X^T y + 2 X^T (X\beta) = 0 \Rightarrow \hat{\beta} = \underbrace{(X^T X)^{-1}}_{\text{problem if it is not isomorphism}} X^T y$$

So, QR decomposition:

$X = QR$ $\quad Q$ orthogonal $\quad R$ upper-triangular.

$$\min_{\beta} \| y - X\beta \|^2 = \min_{\beta} \overbrace{\| y - QR\beta \|^2}^{L(\beta)}$$

$$L(\beta) = (y - QR\beta)^T (y - QR\beta) = (y^T - \beta^T R^T Q^T)(y - QR\beta) =$$

$$= y^T y - \underbrace{y^T QR\beta}_{\in \mathbb{R}} - \underbrace{\beta^T R^T Q^T y}_{\in \mathbb{R}} + \underbrace{\beta^T R^T R \beta}_{\| R\beta \|^2}$$

$\downarrow$ the same

$$\frac{\partial L(\beta)}{\partial \beta} = -2 R^T Q^T y + 2 R^T R \beta = 0 \Rightarrow \beta = (R^T R)^{-1} R^T Q^T y = R^{-1} R^{-T} R^T Q^T y = R^{-1} Q^T y$$

(b) Bridge regression:

$$RSS(\beta, \lambda) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta = y^T y - \overbrace{\beta^T X^T y}^{=} - \overbrace{y^T X \beta}^{=} - \overbrace{\beta^T X^T X \beta}^{\|X\beta\|^2} + \lambda \overbrace{\beta^T \beta}^{\|\beta\|^2}$$

Bridge Regression:

$$\frac{\partial RSS}{\partial \beta} = -2X^T y + 2X^T X\beta + 2\lambda I\beta = 0$$

$$\Rightarrow \quad \text{BRIDGE} = \left(X^T X + \lambda I\right)^{-1} X^T y$$

Now if we consider the standard linear regression for $\tilde{X}$, $\tilde{y}$ we have:

$$\| \tilde{y} - \tilde{X}\beta \|^2 = \left\| \begin{pmatrix} y/\sigma \\ 0_{p\times 1} \end{pmatrix} - \begin{pmatrix} X/\sigma \\ \sqrt{\Lambda} \end{pmatrix} \beta \right\|^2 = \left\| \begin{pmatrix} y/\sigma - \beta X/\sigma \\ -\sqrt{\Lambda}\,\beta \end{pmatrix} \right\|^2 =$$

$$= \frac{1}{\sigma^2} \| y - \beta X \|^2 + \frac{1}{\lambda^2} \beta^T \beta \quad \text{that is a bridge regression.}$$

Question 3

SVD decomposition (***)

Consider a linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

(a) Apply SVD to $\boldsymbol{X}$ and calculate the least square estimate $\hat{\boldsymbol{\beta}}^{LS}$ using this decomposition. Show what this is an attractive way of solving the regression.

(b) Assume that the number of predictors is large compared to the number of observations. Penalize the problem using ridge regression and calculate the ridge estimate $\hat{\boldsymbol{\beta}}^{ridge}$ using SVD on $\boldsymbol{X}$.

(c) Show that the ridge regression computes the coordinates of y with respect to the orthonormal basis U and shrinks these coordinates more when the corresponding singular value is smaller.

(a) $\hat{\beta}^{LS} = (X^T X)^{-1} X^T y$, so if we write $X = UDV^T$ where U, V orthogonal and D diagonal, we have:

$$\hat{\beta}^{LS} = \left((UDV^T)^T UDV^T \right)^{-1} (UDV^T)^T y =$$

$$= \left(V \overset{D}{D^T} U^T UDV^T \right)^{-1} V \overset{D}{D^T} U^T y = \qquad V^{-T} = V^{T\,T} = V$$

$$= (V D^2 V^T)^{-1} V D^T U^T y = V^{-T} D^{-2} V^{-1} V D U^T y \overset{\uparrow}{=} V D^{-1} U^T y$$

I have only to compute $D^{-1} = \begin{pmatrix} \frac{1}{d_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{d_n} \end{pmatrix}$, easy.

(b) $\text{BRIDGE} = (X^TX + \lambda I)^{-1} X^T y = (V D \cancel{U^T} \cancel{U} D V^T + \lambda I)^{-1} V D U^T y =$

$= (V D^2 V^T + \lambda I)^{-1} V D U^T y = (V (D^2 + \lambda I) V^T)^{-1} V D U^T y =$

($I = V V^T$)

$= V^{-T} (D^2 + \lambda I)^{-1} \cancel{V^{-1}} \cancel{V} D U^T y = V (D^2 + \lambda I)^{-1} D U^T y$

($V^{-T} = V$)

but $D^2 + \lambda I = \begin{pmatrix} \lambda + d_1^2 & & 0 \\ & \ddots & \\ 0 & & \lambda + d_n^2 \end{pmatrix} \Rightarrow (D^2 + \lambda I)^{-1} = \begin{pmatrix} \frac{1}{\lambda + d_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda + d_n^2} \end{pmatrix}$

(c) $(\text{BRIDGE})_i = \Big(\underbrace{(V^1, \dots, V^n)}_{V} \underbrace{\begin{pmatrix} \frac{d_1}{\lambda + d_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{d_n}{\lambda + d_n^2} \end{pmatrix}}_{(D^2 + \lambda I)^{-1} D} \underbrace{\begin{pmatrix} U^1 \\ | \\ U^n \end{pmatrix}}_{U^T} y \Big)_i =$

(i-th coordinate; U^j: column of U seen as row-vector)

$= \Big(\Big(V^1 \frac{d_1}{\lambda + d_1^2} - \dots - V^n \frac{d_n}{\lambda + d_n^2} \Big) \begin{pmatrix} U^1 y \\ | \\ U^n y \end{pmatrix} \Big)_i =$

$= \Big(\sum_{j=1}^{n} V^j \underbrace{\frac{d_j}{\lambda + d_j^2}}_{\text{coefficient}} U^j y \Big)_i$

Question 4

Regression Tree versus Linear Regression

Figure 1 represents two two-dimensional problems $X_2 = f(X_1)$ (problem 1 is on row 1 and problem 2 on row 2).

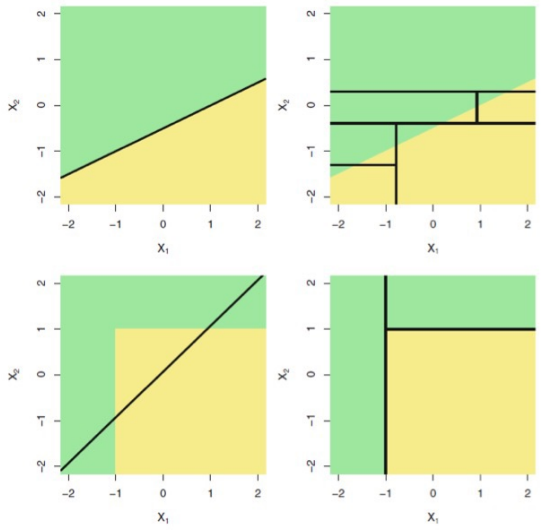


Figure 1: Left column: problem fit by linear regression. Right column: problem fit by regression tree.

(a) Comment on the performance of linear regression versus regression tree. When should you use one or the other?

(a) In the first problem the decision boundary is linear so linear regression is performing perfectly. On the other hand, regression tree is not performing so well since the output is a sum of indicator functions $\sum_{m=1}^{K} c_m \cdot \mathbb{1}_{\{X \in R_m\}}$.

In the second problem we have the opposite: the decision boundaries are parallel to the axis so a sum of indicators functions perform perfectly, instead linear regression doesn't work so well.

In general, when the decision boundary is close to linear is better linear regression, when it is not it is preferred regression trees since are more flexible (a well-known theorem of analysis says us that every positive function can be approximated pointwise by a sequence of indicators function).

Question 5

Boosting as gradient descent in prediction space

(a) The first step of a Gradient Boosting Machine (GBM) is to find a base prediction. Show that it makes sense, when the MSE is the loss function, to use the mean of the N observations as base prediction.

(b) Each step, a new tree is estimated using the residuals of the last step. Show that these residuals can be computed as follows:

$$r_{im} = -\left[\frac{\partial L(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)}\right]_{\hat{f}(x_i)=\hat{f}_{m-1}(x_i)}$$

where m denotes the index of the iteration and $\hat{f}_{m-1}(x_i)$ denotes the prediction made for observation i at the previous iteration $m-1$.

(c) If the MAE is used as loss function, how do the above results change?

possible threshold

(a) We know that the predictor $\hat{f}$ is built s.t. $\hat{f}(x) = \underset{\gamma}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, \gamma)$

where $L(y_i, \gamma) = (y_i - \gamma)^2$. So:

$$\frac{\partial}{\partial \gamma}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i-\gamma)^2\right) = \frac{1}{N}\left(\sum_{i=1}^{N} 2(y_i-\gamma)\cdot(-1)\right) = 0$$

$$\iff \sum_{i=1}^{N} y_i - N\gamma = 0 \Rightarrow \gamma = \sum_{i=1}^{N} y_i / N = \bar{y}.$$

(b) $$r_{im} = -\left[\frac{\partial L(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)}\right]_{\hat{f}(x_i)=\hat{f}_{m-1}(x_i)} =$$

$$= -\left[\frac{\partial\left((y_i - \hat{f}(x_i))^2\right)}{\partial \hat{f}(x_i)}\right]_{\hat{f}(x_i)=\hat{f}_{m-1}(x_i)} = -2\left(y_i - \hat{f}_{m-1}(x_i)\right)(-1) = 2\left(y_i - \hat{f}_{m-1}(x_i)\right)$$

but $y_i - \hat{f}(x_i)$ is the residuals itself so we can compute it doing $\frac{r_{im}}{2}$.

(c) With MAE as loss function we have:

$$L(y_i, \gamma) = |y_i - \gamma|$$

So:

$$r_{im} = -\left[\frac{\partial L(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)}\right]_{\hat{f}(x_i)=\hat{f}_{m-1}(x_i)} =$$

$$= -\left[\frac{\partial\left(|y_i - \hat{f}(x_i)|\right)}{\partial \hat{f}(x_i)}\right]_{\hat{f}(x_i)=\hat{f}_{m-1}(x_i)} = -\operatorname{sgn}\left(y_i - \hat{f}_{m-1}(x_i)\right)\cdot(-1) = \operatorname{sgn}\left(y_i - \hat{f}_{m-1}(x_i)\right)$$

FIN-407 Machine Learning in Finance

Exercise sheet 4 - Supervised learning (cont'd)

Assistant: Giuseppe Matera.

Faculty: Elise Gourier.

Question 1

Logistic model (**)

Let us assume that a response has value 0 or 1, and denote $p(X) = P(Y|X = 1)$. We could have a linear regression model:

$$p(X) = \beta_0 + \beta_1 X$$

The problem with this approach is that $p(X)$ can take negative values, or values that are larger than 1. This can happen any time a straight line is fit to a binary response. To avoid this problem, we choose a function that gives outputs between 0 and 1 for all values of X. In a logistic regression, we use the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

Maximum likelihood can be used to fit the model (1). The logistic function produces an S-shaped $P(X)$, which nevers goes beyond 0 or 1.

Show that

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

The left hand-side is called the *odds* and can take any value between 0 and ∞. In horse-racing, odds are used more often than probabilities because they relate more naturally to

the right betting strategy: the odds are close to 0 if $p(X)$ is close to 0, and close to ∞ if $p(X)$ is close to 1.

The logarithm of the odds is called the *log-odds* or *logit*. The logistic regression model has a logit that is linear in X.

Question 2

Classification with K-nearest neighbors

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X1 = X2 = X3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

(b) What is your prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

Question 3

Linear Discriminant Analysis for $p = 1$

Assume that $p = 1$, i.e. we only have 1 predictor. Let K denote the number of possible classes (=groups), and G be the random variable denoting the class of observations. Define the class-conditional density of the predictor $f_k(x)$ as

$$f_k(x) = P(X = x|G = k)$$

and let $p_k(x)$ be the probability of each observation belonging to a class (the probability that we want to predict):

$$p_k(x) = P(G = k|X = x).$$

According to Bayes theorem,

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)} \tag{2}$$

where $\pi_k = P(G = k)$.

We would like to obtain an estimate for $f_k(x)$ that we can plug into (2) in order to estimate $p_k(x)$. We will then classify an observation to the class for which $p_k(x)$ is greatest. That is, in the notation of the lecture, the class estimator $\hat{G}(x)$ is the value of k for which $p_k(x)$ is the largest. In order to estimate $f_k(x)$, we will first make some assumptions about its form.

(a) Assume that $f_k(x)$ is normal (Gaussian). Rewrite $p_k(x)$.

(b) Define the discriminant function:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k. \tag{4}$$

Show that classifying an observation to the class for which $p_k(x)$ is largest is equivalent to classifying an observation to the class for which (4) is largest.

This shows that under the assumption that the observations in the k^{th} class are drawn from a $\mathcal{N}(\mu_k, \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

Assume there are only two classes: $K = 2$, and $\pi_1 = \pi_2$.

(c) Derive the Bayes decision boundary.

(d) Figure 1 illustrates the example above. The two densities $f_1(x)$ and $f_2(x)$ overlap, and so given that $X = x$, there is some uncertainty about the class to which the observation belongs. How does a Bayes classifier assign an observation to a given class?

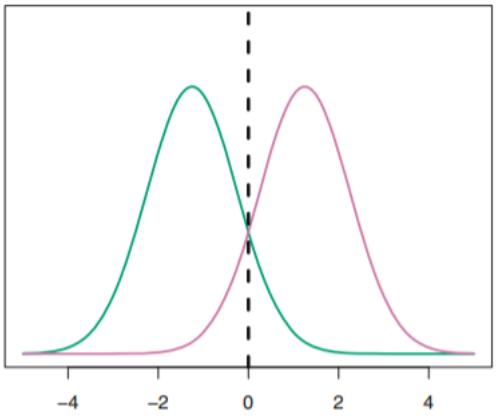


Figure 1: The two normal density functions that are displayed, $f_1(x)$ and $f_2(x)$, represent two distinct classes with same probabilities π_1 and π_2. The mean and variance parameters for the two density functions are μ_1 = -1.25, μ_2 = 1.25, and $\sigma_1 = \sigma_2 = 1$.

(e) In the above example, we are given all the parameters. In practice, we need to estimate them. Describe how you would estimate the parameters.

Question 4

LDA

Assume that the data is distributed in two classes and that they are normally distributed:

$$p(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1|\mu_1, \Sigma) \quad \text{and} \quad p(\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_2|\mu_2, \Sigma).$$

The vector μ_1 is the sample mean of class 1 data, and Σ is the sample covariance, similarly for class 2. We project the data in these two classes on two one dimension to obtain

$$y_1 = \omega^T \boldsymbol{x}_1 \quad \text{and} \quad y_2 = \omega^T \boldsymbol{x}_2$$

(a) What is the distribution of y_1 and y_2?

We search for a projection ω such that the projected distributions are as "separated" as possible. Fisher defines the separation between the two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{S_{between}}{S_{within}} = \frac{(\mathbb{E}(y_1) - \mathbb{E}(y_2))^2}{pVar(y_1) + (1-p)Var(y_2)}$$

where p represents the probability to be in class 1. This measure is, in some sense, a measure of the signal-to-noise ratio for the class labelling.

(b) Show that maximum separation is obtained for the $\omega \propto \Sigma^{-1}(\mu_1 - \mu_2)$.

Question 5

Logistic regression vs. LDA

Consider the two-class setting with $p = 1$ predictor, and let $p_1(x)$ and $p_2(x) = 1 - p_1(x)$ be the probabilities that the observation $X = x$ belongs to class 1 and class 2, respectively.

(a) Write the log odds in the LDA framework as a linear function of x.

(b) Compare with the log-odds in a logistic regression. Comment on the decision boundaries produced by LDA and the logistic regression. Is there a difference between the two approaches? How do you expect them to perform comparatively to each other?

(c) Compare the 2 above approaches with the k-nearest neighbor approach.

(d) How would you expect the QDA (quadratic discriminant analysis) to perform compared to the above approaches?

Question 6

Bias-variance trade-off

How you do expect KNN, LDA, QDA and logistic regression to perform in the below scenarios:

(a) Assume 2 classes, with 20 training observations in each of the classes. The observations within each class are uncorrelated random normal variables with a different mean in each class.

(b) Same scenario as in (a) except that within each class, the two predictors have a correlation of -0.5.

(c) Same scenario as in (a) except that X_1 and X_2 are generated from a t-distribution, with 50 observations in each class.

(d) The data are generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and a correlation of -0.5 between the predictors in the second class.

(e) Within each class, the observations are generated from a normal distribution with uncorrelated predictors. However, the responses are sampled from the logistic function using X_1^2, X_2^2 , and $X_1 \times X_2$ as predictors.

(f) Details are as in the previous scenario, but the responses are sampled from a more complicated non-linear function.

Question 1

Logistic model (**)

Let us assume that a response has value 0 or 1, and denote $p(X) = P(Y|X = 1)$. We could have a linear regression model:

$$p(X) = \beta_0 + \beta_1 X$$

The problem with this approach is that $p(X)$ can take negative values, or values that are larger than 1. This can happen any time a straight line is fit to a binary response. To avoid this problem, we choose a function that gives outputs between 0 and 1 for all values of X. In a logistic regression, we use the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \quad (1)$$

Maximum likelihood can be used to fit the model (1). The logistic function produces an S-shaped $P(X)$, which nevers goes beyond 0 or 1.

(1) Show that

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

The left hand-side is called the *odds* and can take any value between 0 and ∞. In horse-racing, odds are used more often than probabilities because they relate more naturally to the right betting strategy: the odds are close to 0 if $p(X)$ is close to 0, and close to ∞ if $p(X)$ is close to 1.

The logarithm of the odds is called the *log-odds* or *logit*. The logistic regression model has a logit that is linear in X.

(1) $\frac{p(X)}{1-p(X)} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1 - \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}} = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X} - e^{\beta_0+\beta_1 X}} = e^{\beta_0+\beta_1 X}$

Question 2

Classification with K-nearest neighbors

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

we have the label!

Obs	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X1 = X2 = X3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

(b) What is your prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

(a) $d_1 = \sqrt{(0-0)^2 + (0-3)^2 + (0-0)^2} = 3 \quad d_2 = 2 \quad d_3 = \sqrt{10} \quad d_4 = \sqrt{5} \quad d_5 = \sqrt{2}$

$d_6 = \sqrt{3}$

(b) $K = 1 \Rightarrow$ I have to choose the nearest that is Obs. 5 $\Rightarrow$ the prediction is "Green".

(c) $K = 3 \Rightarrow$ I have to choose the 3 nearest vectors: Obs. 2, 5, 6 and 2, 6 are Red instead 5 is green $\Rightarrow$ prediction is Red.

(d) KNN is very flexible but increasing K the complexity of this method reduces so, if the boundary is highly non linear, we want an high-complex method $\Rightarrow$ K small.

Question 3

Linear Discriminant Analysis for $p = 1$

Assume that $p = 1$, i.e. we only have 1 predictor. Let K denote the number of possible classes (=groups), and G be the random variable denoting the class of observations. Define the class-conditional density of the predictor $f_k(x)$ as

$$f_k(x) = P(X = x|G = k)$$

and let $p_k(x)$ be the probability of each observation belonging to a class (the probability that we want to predict):

$$p_k(x) = P(G = k|X = x).$$

According to Bayes theorem,

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)} \tag{2}$$

where $\pi_k = P(G = k)$.

We would like to obtain an estimate for $f_k(x)$ that we can plug into (2) in order to estimate $p_k(x)$. We will then classify an observation to the class for which $p_k(x)$ is greatest. That is, in the notation of the lecture, the class estimator $\hat{G}(x)$ is the value of k for which $p_k(x)$ is the largest. In order to estimate $f_k(x)$, we will first make some assumptions about its form.

(a) Assume that $f_k(x)$ is normal (Gaussian). Rewrite $p_k(x)$.

(b) Define the discriminant function:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k. \tag{4}$$

Show that classifying an observation to the class for which $p_k(x)$ is largest is equivalent to classifying an observation to the class for which (4) is largest.

This shows that under the assumption that the observations in the k^{th} class are drawn from a $\mathcal{N}(\mu_k, \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

Assume there are only two classes: $K = 2$, and $\pi_1 = \pi_2$. $=: \pi$

(c) Derive the Bayes decision boundary.

(d) Figure 1 illustrates the example above. The two densities $f_1(x)$ and $f_2(x)$ overlap, and so given that $X = x$, there is some uncertainty about the class to which the observation belongs. How does a Bayes classifier assign an observation to a given class?

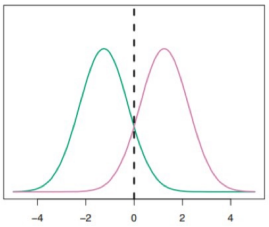


Figure 1: The two normal density functions that are displayed, $f_1(x)$ and $f_2(x)$, represent two distinct classes with same probabilities π_1 and π_2. The mean and variance parameters for the two density functions are μ_1 = -1.25, μ_2 = 1.25, and $\sigma_1 = \sigma_2 = 1$.

(e) In the above example, we are given all the parameters. In practice, we need to estimate them. Describe how you would estimate the parameters.

(a) $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma_k^2}\right)$, in LDA $\sigma_1 = \dots = \sigma_K =: \sigma$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma^2}\right)}{\sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu_j)^2}{\sigma^2}\right)} = \frac{\pi_k \exp\left(-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma^2}\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\frac{1}{2}\frac{(x-\mu_j)^2}{\sigma^2}\right)}$$

(b) Suppose for index s :

$p_s(x) \geq p_k \quad \forall k \in \{1, \dots, K\}$. This is true iff:

$$\frac{\pi_s \exp\left(-\frac{1}{2}\frac{(x-\mu_s)^2}{\sigma^2}\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\frac{1}{2}\frac{(x-\mu_j)^2}{\sigma^2}\right)} \geq \frac{\pi_k \exp\left(-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma^2}\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\frac{1}{2}\frac{(x-\mu_j)^2}{\sigma^2}\right)} \quad \forall k \iff$$

> 0 $\qquad$ > 0

$$\iff \pi_s \exp\left(-\frac{1}{2}\frac{(x-\mu_s)^2}{\sigma^2}\right) \geq \pi_k \exp\left(-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma^2}\right) \quad \forall k$$

$$\iff \log\pi_s - \frac{1}{2}\frac{(x-\mu_s)^2}{\sigma^2} \geq \log\pi_k - \frac{1}{2}\frac{(x-\mu_k)^2}{\sigma^2} \quad \forall k$$

$$\iff \log\pi_s - \cancel{\frac{1}{2}\frac{x^2}{\sigma^2}} - \frac{1}{2}\frac{\mu_s^2}{\sigma^2} + \frac{x\mu_s}{\sigma^2} \geq \log\pi_k - \cancel{\frac{1}{2}\frac{x^2}{\sigma^2}} - \frac{1}{2}\frac{\mu_k^2}{\sigma^2} + \frac{x\mu_k}{\sigma^2} \quad \forall k$$

$$\iff \delta_s(x) \geq \delta_k(x) \quad \forall k$$

(c) We want to maximize δ_k with $k = 1, 2$, so the decision boundary is when $\delta_1 = \delta_2$

$$\delta_1(x) = \delta_2(x) \iff x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \cancel{\log\pi} = x\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \cancel{\log\pi}$$

$$\iff x(\mu_1 - \mu_2) = \frac{\mu_1^2 - \mu_2^2}{2} \iff x = \frac{\mu_1 + \mu_2}{2}$$

(d) We note from the graph that $\mu_1 = -\mu_2 \Rightarrow$ the boundary is $x = 0 \Rightarrow$

if $x > 0 \Rightarrow$ class 2, if $x < 0 \Rightarrow$ class 1.

(e) We have one predictor (vector of dimension 1) so:

$$\hat{\mu}_k := \frac{1}{N_k}\sum_{x_i \in G_k} x_i \quad (N_k = \#G_k,\ x_i \text{ elements in } G_k)$$

$$\hat{\sigma}^2 := \frac{1}{N-K}\sum_{k=1}^{K}\sum_{x_i \in G_k}(x_i - \hat{\mu}_k)^2 \qquad \hat{\pi}_k = \frac{N_k}{N}$$

√ **Question 4**

LDA

Assume that the data is distributed in two classes and that they are normally distributed:

$$p(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1|\mu_1, \Sigma) \quad \text{and} \quad p(\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_2|\mu_2, \Sigma).$$

The vector μ_1 is the sample mean of class 1 data, and Σ is the sample covariance, similarly for class 2. We project the data in these two classes on two one dimension to obtain

$$y_1 = \omega^T \boldsymbol{x}_1 \quad \text{and} \quad y_2 = \omega^T \boldsymbol{x}_2$$

(a) What is the distribution of y_1 and y_2?

We search for a projection ω such that the projected distributions are as "separated" as possible. Fisher defines the separation between the two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{S_{between}}{S_{within}} = \frac{(\mathbb{E}(y_1) - \mathbb{E}(y_2))^2}{pVar(y_1) + (1-p)Var(y_2)}$$

where p represents the probability to be in class 1. This measure is, in some sense, a measure of the signal-to-noise ratio for the class labelling.

(b) Show that maximum separation is obtained for the $\omega \propto \Sigma^{-1}(\mu_1 - \mu_2)$.

(a) $p(y_1) = \omega^T N(\mu_1, \Sigma) = N(\omega^T \mu_1, \omega^T \Sigma \omega)$

$p(y_2) = \omega^T N(\mu_2, \Sigma) = N(\omega^T \mu_2, \omega^T \Sigma \omega)$

(b) $S = \dfrac{(\omega^T \mu_1 - \omega^T \mu_2)^2}{p\,\omega^T \Sigma \omega + (1-p)(\omega^T \Sigma \omega)} = \dfrac{(\omega^T \mu_1 - \omega^T \mu_2)^2}{\omega^T \Sigma \omega}$

$$\nabla_\omega S = \frac{2(\omega^T\mu_1 - \omega^T\mu_2)(\mu_1 - \mu_2)\,\omega^T\Sigma\omega - (\omega^T\mu_1 - \omega^T\mu_2)^2 \cancel{2}\Sigma\omega}{(\omega^T\Sigma\omega)^2} = 0$$

(numbers: $(\omega^T\mu_1 - \omega^T\mu_2)$, $\omega^T\Sigma\omega$, $(\omega^T\mu_1 - \omega^T\mu_2)^2$)

$$\underbrace{\omega^T\Sigma\omega}_{\text{number}}(\mu_1 - \mu_2) - \underbrace{\omega^T(\mu_1 - \mu_2)}_{\text{number}}\Sigma\omega = 0$$

$$\omega = \underbrace{\frac{\omega^T\Sigma\omega}{\omega^T(\mu_1 - \mu_2)}}_{\text{number}} \Sigma^{-1}(\mu_1 - \mu_2) \quad \text{(it's the same result of the solution!)}$$

√ **Question 5**

Logistic regression vs. LDA

Consider the two-class setting with $p = 1$ predictor, and let $p_1(x)$ and $p_2(x) = 1 - p_1(x)$ be the probabilities that the observation $X = x$ belongs to class 1 and class 2, respectively.

(a) Write the log odds in the LDA framework as a linear function of x.

(b) Compare with the log-odds in a logistic regression. Comment on the decision boundaries produced by LDA and the logistic regression. Is there a difference between the two approaches? How do you expect them to perform comparatively to each other?

(c) Compare the 2 above approaches with the k-nearest neighbor approach.

(d) How would you expect the QDA (quadratic discriminant analysis) to perform compared to the above approaches?

(a) $p_1(x) = P(G=1 \mid X=x) \qquad p_2(x) = P(G=2 \mid X=x)$

$$\log\left(\frac{P(G=1\mid X=x)}{P(G=2\mid X=x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = \log\left(\frac{P(G=1)\,\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}\right)}{P(G=2)\,\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2}\right)}\right) =$$

$$= \log\left(P(G=1)/P(G=2)\right) - \frac{1}{2\sigma^2}\left((x-\mu_1)^2 - (x-\mu_2)^2\right) =$$

$$= \log\left(P(G=1)/P(G=2)\right) - \frac{1}{2\sigma^2}\left(\cancel{x^2} + \mu_1^2 - 2x\mu_1 - \cancel{x^2} - \mu_2^2 + 2x\mu_2\right) =$$

$$= \underbrace{\log\left(P(G=1)/P(G=2)\right) - \frac{\mu_1^2 - \mu_2^2}{2\sigma^2}}_{c_0} + x\underbrace{\frac{(\mu_1 - \mu_2)}{\sigma^2}}_{c_1} = c_0 + c_1 x \Rightarrow \text{linear}$$

(b) $\log\left(\frac{p_1(x)}{p_2(x)}\right) = \log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{\exp(w^T x)}{1+\exp(w^T x)} \cdot \frac{1+\exp(w^T x)}{1}\right)$

$w^T = (1, w_1)$

$= w^T x = (1, w_1)\begin{pmatrix}\beta \\ x_1\end{pmatrix} = \beta + w_1 x_1 \Rightarrow$ linear.

The difference is that β, w_1 are estimated with maximum likelyhood, instead μ, C_1 are used estimating mean and variance from a gaussian distribution.

In practice, they're very similar but we know that in LDA we assume feature are gaussian r.v. and Σ is the same so when the hyp. of the problem are these LDA performs better, the opposite when hyp don't fit.

(c) Given $X = x$, KNN find the K training observation that are nearest to x.

Then x is assigned to the class to which the plurality of these observations belong.

So KNN is not parametric and no assumptions are made on boundary (small K works for no linear boundary) On the other hand, KNN does not tell us which predictors are important

(d) QDA compromises between KNN and the linear LDA and logistic regression.

The boundary of QDA are quadratic so more flexible than LDA and logistic but less than KNN.

Question 6

Bias-variance trade-off

How you do expect KNN, LDA, QDA and logistic regression to perform in the below scenarios:

(a) Assume 2 classes, with 20 training observations in each of the classes. The observations within each class are uncorrelated random normal variables with a different mean in each class.

(b) Same scenario as in (a) except that within each class, the two predictors have a correlation of -0.5.

(c) Same scenario as in (a) except that X_1 and X_2 are generated from a t-distribution, with 50 observations in each class.

(d) The data are generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and a correlation of -0.5 between the predictors in the second class.

(e) Within each class, the observations are generated from a normal distribution with uncorrelated predictors. However, the responses are sampled from the logistic function using X_1^2, X_2^2, and $X_1 \times X_2$ as predictors.

(f) Details are as in the previous scenario, but the responses are sampled from a more complicated non-linear function.

(a) Small dataset $\Rightarrow$ complex models overfit $\Rightarrow$ KNN and QDA not so performing

(KNN less than QDA since is more complex). LDA the best, also logistic regression good but worse than LDA since random variables are normal distributed.

(b) Little change. no significant

(c) t-distribution similar to gaussian but more extreme points ⇒ logistic regression better than LDA. The boundary are always linear (not complex dataset) ⇔ they're better than KNN and QDA (overfitting problems)

(d) Different variance ⇒ QDA is the best

(e) Quadratic decision boundary ⇒ QDA and KNN the best. Linear models bad performance

(f) KNN is the most flexible ⇒ the best model

FIN-407 Machine Learning in Finance

Exercise sheet 5 - Support Vector Machines

Assistant: Giuseppe Matera.

Faculty: Elise Gourier.

Question 1

Decision boundary (*)

Consider the decision boundary corresponding to $(\boldsymbol{w}, b)$.

(a) Show that $\boldsymbol{w}$ is orthogonal to the decision boundary.

(b) Consider point A $(\boldsymbol{x}_i, y_i)$ above the decision boundary $(y_i = 1)$. Calculate the distance γ_i of A to the decision boundary.

(c) Calculate, more generally, the distance γ_i of any point in the training set to the decision boundary.

Question 1

Decision boundary (*)

Consider the decision boundary corresponding to $(\boldsymbol{w}, b)$.

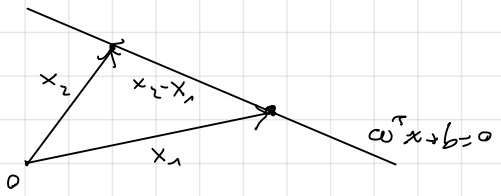
(a) Show that $\boldsymbol{w}$ is orthogonal to the decision boundary.

(b) Consider point A $(\boldsymbol{x}_i, y_i)$ above the decision boundary $(y_i = 1)$. Calculate the distance γ_i of A to the decision boundary.

(c) Calculate, more generally, the distance γ_i of any point in the training set to the decision boundary.

output is binary: +1 or -1 , +1 is above, -1 is below

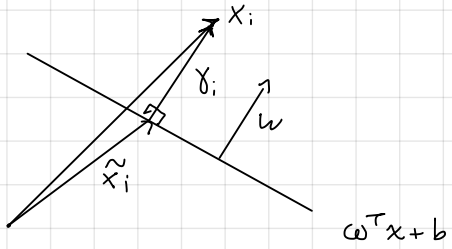
(a) The decision boundary is $\omega^T x + b = 0$



I take ω_1, ω_2 points of the decision boundary:

$$\begin{cases} \omega^T x_1 + b = 0 \\ \omega^T x_2 + b = 0 \end{cases} \Rightarrow \omega^T (x_1 - x_2) = 0 \Rightarrow \omega \perp x_1 - x_2$$ but $x_1 - x_2$ lies on the decision boundary $\Rightarrow \omega \perp$ decision boundary

(b)



$\tilde{x}_i \in$ decision boundary $\Rightarrow \omega^T \tilde{x}_i + b = 0$. At the same time $\tilde{x}_i - x_i \parallel w$ and the same versus (is above the boundary) so:

$$x_i - \tilde{x}_i = \gamma_i \frac{w}{\|w\|}$$

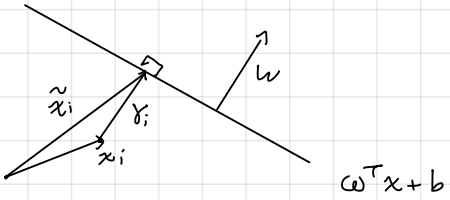
So:

$$\omega^T \left(x_i - \gamma_i \frac{w}{\|w\|} \right) + b = 0$$

$$\omega^T x_i + b = \frac{\gamma_i}{\|w\|} w w^T = \frac{\gamma_i}{\|w\|} \|w\|^2$$

$$\gamma_i = \frac{\omega^T x_i}{\|w\|} + \frac{b}{\|w\|}$$

(c) We have to compute the distance for a point below the decision boundary:



Now $\tilde{x}_i = x_i + \gamma_i \frac{w}{\|w\|} \Rightarrow w^T \left(x_i + \gamma_i \frac{w}{\|w\|} \right) + b = 0$

$\Rightarrow w^T x_i + \gamma_i \|w\| + b = 0 \Rightarrow \gamma_i = \frac{-w^T x_i}{\|w\|} - \frac{b}{\|w\|}$

So, in general, $\gamma_i = y_i \left(\frac{w^T x_i}{\|w\|} + \frac{b}{\|w\|} \right)$

FIN-407 Machine Learning in Finance

Exercise sheet 6 - Unsupervised learning

Assistant: Giuseppe Matera.

Faculty: Elise Gourier.

Question 1

K-means clustering

(a) Remember that for each cluster $\mathcal{C}_k$, the center $\boldsymbol{m}_k$ is given by:

$$\boldsymbol{m}_k = \arg\min_{\boldsymbol{m}} \sum_{i \in \mathcal{C}_k} ||\boldsymbol{x}_i - \boldsymbol{m}||_2^2.$$

Show that $\boldsymbol{m}_k = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} \boldsymbol{x}_i$.

(b) Check that the within-cluster dissimilarity is:

$$W(\mathcal{C}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j \in \mathcal{C}_k} ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = \sum_{k=1}^{K} N_k \sum_{i \in \mathcal{C}_k} ||\boldsymbol{x}_i - \boldsymbol{m}_k||_2^2$$

Let us use the K-means algorithm and Euclidian distance to cluster the 8 data points given in Figure 1 into $K = 3$ clusters. The distance matrix based on the Euclidian distance is given in Table 1. The coordinates of the data points are:

$$x^{(1)} = (2, 8) \quad x^{(2)} = (2, 5) \quad x^{(3)} = (1, 2) \quad x^{(4)} = (5, 8)$$
$$x^{(5)} = (7, 3) \quad x^{(6)} = (6, 4) \quad x^{(7)} = (8, 4) \quad x^{(8)} = (4, 7)$$

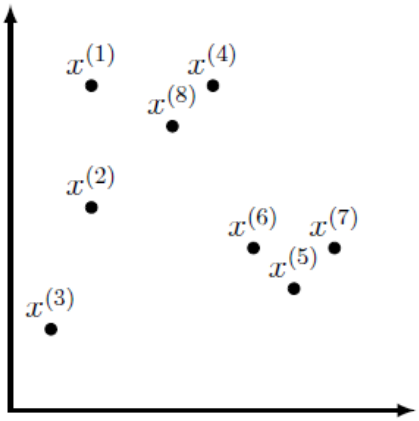


Figure 1: Training data set for K-means clustering.

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$
$x^{(1)}$	0	3.0000	6.0828	3.0000	7.0711	5.6569	7.2111	2.2361
$x^{(2)}$	3.0000	0	3.1623	4.2426	5.3852	4.1231	6.0828	2.8284
$x^{(3)}$	6.0828	3.1623	0	7.2111	6.0828	5.3852	7.2801	5.8310
$x^{(4)}$	3.0000	4.2426	7.2111	0	5.3852	4.1231	5.0000	1.4142
$x^{(5)}$	7.0711	5.3852	6.0828	5.3852	0	1.4142	1.4142	5.0000
$x^{(6)}$	5.6569	4.1231	5.3852	4.1231	1.4142	0	2.0000	3.6056
$x^{(7)}$	7.2111	6.0828	7.2801	5.0000	1.4142	2.0000	0	5.0000
$x^{(8)}$	2.2361	2.8284	5.8310	1.4142	5.0000	3.6056	5.0000	0

Table 1: Distance matrix for training data from Table 1.

(c) Suppose you initialize the cluster centers instead of initializing the clusters, as in the slides. The steps will be the same (just starting with step 2 in the slides, then iterating through 1 and 2). Assume that points $x^{(3)}$, $x^{(4)}$ and $x^{(6)}$ are chosen as cluster centers. Perform one iteration of the K-means and report the coordinates of the resulting centroids.

(d) Calculate the within-cluster dissimilarity at initialization and after the first iteration.

Comment on your results.

Question 2

K-means clustering 2

Consider the following points:

Pt	X	Y
Pt1	2	3
Pt2	3	1
Pt3	4	2
Pt4	11	5
Pt5	12	4
Pt6	12	6
Pt7	7	5
Pt8	8	4
Pt9	8	6

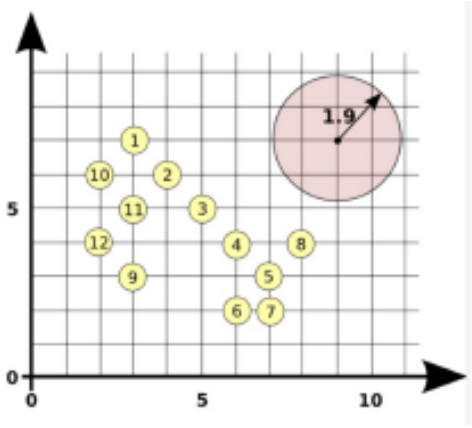
(a) Apply K-means starting with the centroids Pt1 and Pt8

(b) Select a pair of initial centroids such that we get two different clusters.

Question 3

Hierarchical clustering

(a) Execute single-linkage and complete-linkage hierarchical clustering on the following similarity matrix. The more similar two points are, the smaller the distance between them.



	P1	P2	P3	P4	P5
P1	1	0.10	0.41	0.55	0.35
P2	0.10	1	0.64	0.47	0.98
P3	0.41	0.64	1	0.44	0.85
P4	0.55	0.47	0.44	1	0.76
P5	0.35	0.98	0.85	0.76	1

Question 4

DBSCAN

Apply the DBSCAN algorithm with radius 1.9 and MinPts=4 (3 neighbors + the point we are considering as center for computing the density). Indicate what is the nature of each point. Draw the clusters.

Question 5

Principal component analysis

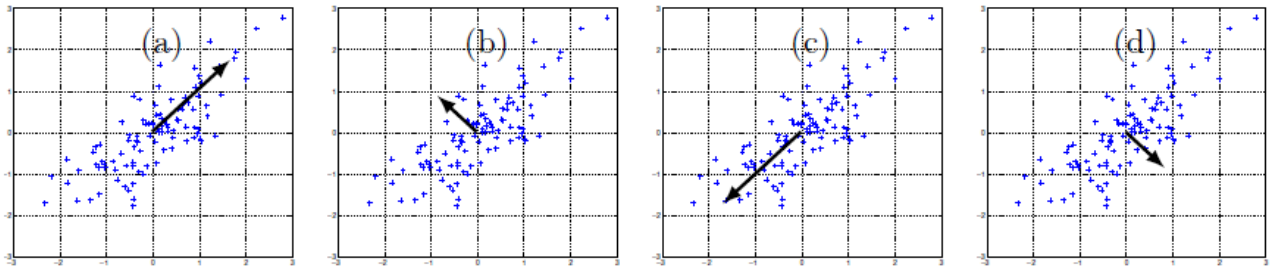
(a) Let Σ be a positive semi-definite d-dimensional matrix, with eigenvalues λ_i, $i = 1, ..., d$. Show that

$$\text{trace}(\Sigma) = \sum_{i=1}^{d} \lambda_i.$$

(b) Let $\boldsymbol{x} = (x_1, ..., x_d)^T$ denote a d-dimensional random vector with variance-covariance matrix Σ. Let $p_1 = \gamma_1^T \boldsymbol{x}$ be the first principal component of $\boldsymbol{x}$. Show that γ_1 solves

$$\max_{\boldsymbol{a}} Var(\boldsymbol{a}^T \boldsymbol{x})$$
$$\text{subject to } \boldsymbol{a}^T \boldsymbol{a} = 1.$$

(c) Which of the following figures correspond to possible values that PCA may return for the first principal component?



Question 1

K-means clustering

(a) Remember that for each cluster $\mathcal{C}_k$, the center $\boldsymbol{m}_k$ is given by:

$$\boldsymbol{m}_k = \arg\min_{\boldsymbol{m}} \sum_{i \in \mathcal{C}_k} ||\boldsymbol{x}_i - \boldsymbol{m}||_2^2.$$

Show that $\boldsymbol{m}_k = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} \boldsymbol{x}_i$.

(b) Check that the within-cluster dissimilarity is:

$$W(\mathcal{C}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j \in \mathcal{C}_k} ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = \sum_{k=1}^{K} N_k \sum_{i \in \mathcal{C}_k} ||\boldsymbol{x}_i - \boldsymbol{m}_k||_2^2$$

Let us use the K-means algorithm and Euclidian distance to cluster the 8 data points given in Figure 1 into $K = 3$ clusters. The distance matrix based on the Euclidian distance is given in Table 1. The coordinates of the data points are:

$$x^{(1)} = (2, 8) \quad x^{(2)} = (2, 5) \quad x^{(3)} = (1, 2) \quad x^{(4)} = (5, 8)$$
$$x^{(5)} = (7, 3) \quad x^{(6)} = (6, 4) \quad x^{(7)} = (8, 4) \quad x^{(8)} = (4, 7)$$

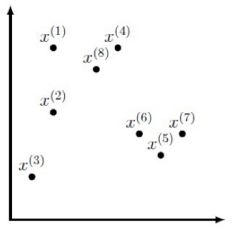


Figure 1: Training data set for K-means clustering.

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$
$x^{(1)}$	0	3.0000	6.0828	3.0000	7.0711	5.6569	7.2111	2.2361
$x^{(2)}$	3.0000	0	3.1623	4.2426	5.3852	4.1231	6.0828	2.8284
$x^{(3)}$	6.0828	3.1623	0	7.2111	6.0828	5.3852	7.2801	5.8310
$x^{(4)}$	3.0000	4.2426	7.2111	0	5.3852	4.1231	5.0000	1.4142
$x^{(5)}$	7.0711	5.3852	6.0828	5.3852	0	1.4142	1.4142	5.0000
$x^{(6)}$	5.6569	4.1231	5.3852	4.1231	1.4142	0	2.0000	3.6056
$x^{(7)}$	7.2111	6.0828	7.2801	5.0000	1.4142	2.0000	0	5.0000
$x^{(8)}$	2.2361	2.8284	5.8310	1.4142	5.0000	3.6056	5.0000	0

Table 1: Distance matrix for training data from Table 1.

(c) Suppose you initialize the cluster centers instead of initializing the clusters, as in the slides. The steps will be the same (just starting with step 2 in the slides, then iterating through 1 and 2). Assume that points $x^{(3)}$, $x^{(4)}$ and $x^{(6)}$ are chosen as cluster centers. Perform one iteration of the K-means and report the coordinates of the resulting centroids.

(d) Calculate the within-cluster dissimilarity at initialization and after the first iteration.

(a)

$$m_K = \underset{m}{\operatorname{argmin}} \sum_{i \in C_K} \|x_i - m\|_2^2$$

$$\nabla_m \sum_{i \in C_K} \|x_i - m\|_2^2 = \sum_{i \in C_K} 2(x_i - m)(-1) = 0$$

$$\Rightarrow \sum_{i \in C_K} x_i = n_K m \Rightarrow m_K = \frac{\sum_{i \in C_K} x_i}{n_K}$$

(b)

$$\frac{1}{2} \sum_{K=1}^{K} \sum_{i,j \in C_K} \|x_i - x_j\|_2^2 = \frac{1}{2} \sum_{K=1}^{K} \sum_{i,j \in C_K} \|(x_i - m_K) - (x_j - m_K)\|_2^2 =$$

$$= \frac{1}{2} \sum_{K=1}^{K} \sum_{i,j C_K} \left(\|x_i - m_K\|^2 + \|x_j - m_K\|_2^2 + 2\langle x_i - m_K, x_j - m_K \rangle \right) =$$

$$= \frac{1}{2} \sum_{K=1}^{K} \sum_{i,j \in C_K} \|x_i - m_K\|^2 + \frac{1}{2} \sum_{K=1}^{K} \sum_{i,j \in C_K} \|x_j - m_K\|_2^2 + \sum_{K=1}^{K} \sum_{i,j \in C_K} \langle x_i - m_K, x_j - m_K \rangle =$$

$$= \frac{1}{2} N_K \sum_{K=1}^{K} \sum_{i \in C_K} \|x_i - m_K\|^2 + \frac{1}{2} N_K \sum_{K=1}^{K} \sum_{j \in C_K} \|x_j - m_K\|_2^2 + \sum_{K=1}^{K} \left\langle \sum_{i \in C_K} (x_i - m_K), \sum_{j \in C_K} (x_j - m_K) \right\rangle =$$

$$= N_K \sum_{K=1}^{K} \sum_{i \in C_K} \|x_i - m_K\|^2 + \sum_{K=1}^{K} \left\langle \sum_{i \in C_K} x_i - N_K m_K, \sum_{j \in C_K} x_j - N_K m_K \right\rangle =$$

$$= 0$$

$$= N_K \sum_{K=1}^{K} \sum_{i \in C_K} \|x_i - m_K\|^2$$

(c) · We associate point x to the cluster that has the nearest center to x. (Notation C_i := cluster of center x_i):

$x_1 \rightsquigarrow C_4$ $x_2 \rightsquigarrow C_3$ $x_3 \rightsquigarrow C_3$ $x_4 \rightsquigarrow C_4$ $x_5 \rightsquigarrow C_6$ $x_6 \rightsquigarrow C_6$ $x_7 \rightsquigarrow C_6$ $x_8 \rightsquigarrow C_4$

· Now we re-compute the centers:

$$m_3 = \frac{x_2 + x_3}{2} = \frac{(2,5) + (1,2)}{2} = \frac{(3,7)}{2} = (1.5, 3.5)$$

$$m_4 = \frac{x_1 + x_4 + x_8}{3} = \frac{(2,8) + (5,8) + (4,7)}{3} = \frac{(11, 23)}{3} = (3.67, 7.67)$$

$$m_6 = \frac{x_5 + x_6 + x_7}{3} = \frac{(7,3) + (6,4) + (8,4)}{3} = \frac{(21, 11)}{3} = (7, 3.67)$$

(d) · At initialization:

$$W(C)_{BEG.} = 2\left(\|x_2 - x_3\|_2^2 + \|x_3 - x_3\|_2^2 \right) + 3\left(\|x_1 - x_4\|_2^2 + \|x_4 - x_4\|_2^2 + \|x_4 - x_8\|_2^2 \right) + 3\left(\|x_5 - x_6\|_2^2 + \|x_6 - x_6\|_2^2 + \|x_7 - x_6\|_2^2 \right) =$$

$$= 2\left((3.1623)^2 + (0^2) \right) + \dots$$

· After the iteration we compute the new distances from the new centers:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
m_3	4.53	1.58	1.58	5.70	5.52	4.53	6.52	4.3
m_4	1.70	3.15	6.26	1.37	5.74	4.35	5.68	0.75
m_6	6.61	5.17	6.22	4.77	0.67	1.05	1.05	4.48

So:

$x_1 \rightsquigarrow C_4$ $x_2 \rightsquigarrow C_3$ $x_3 \rightsquigarrow C_3$ $x_4 \rightsquigarrow C_4$ $x_5 \rightsquigarrow C_6$ $x_6 \rightsquigarrow C_6$ $x_7 \rightsquigarrow C_6$ $x_8 \rightsquigarrow C_4$

The clusters remain the same. The new dissimilarity is:

$$W(C)_{BEG.} = 2\left(\|x_2 - m_3\|_2^2 + \|x_3 - m_3\|_2^2 \right) + 3\left(\|x_1 - m_4\|_2^2 + \|x_4 - m_4\|_2^2 + \|x_8 - m_4\|_2^2 \right) + 3\left(\|x_5 - m_6\|_2^2 + \|x_6 - m_6\|_2^2 + \|x_7 - m_6\|_2^2 \right) =$$

The within cluster dissimilarity decreases as we expected.

Question 2

K-means clustering 2

Consider the following points:

Pt	X	Y
Pt1	2	3
Pt2	3	1
Pt3	4	2
Pt4	11	5
Pt5	12	4
Pt6	12	6
Pt7	7	5
Pt8	8	4
Pt9	8	6

(a) Apply K-means starting with the centroids Pt1 and Pt8

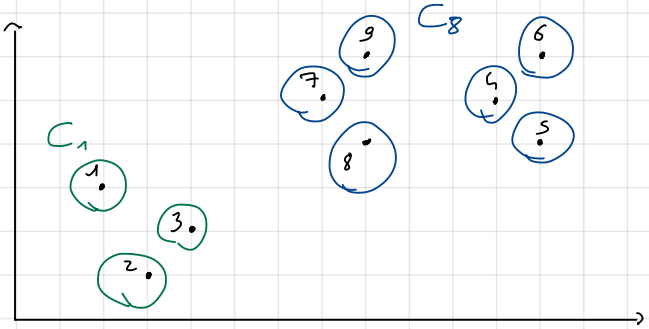
(b) Select a pair of initial centroids such that we get two different clusters.

(a) We compute distances from Pt 1 and Pt 8

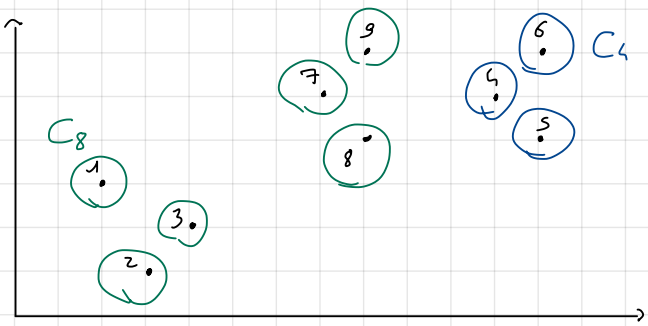
	Pt 1	Pt 2	Pt 3	Pt 4	Pt 5	Pt 6	Pt 7	Pt 8	Pt 9
Pt 1	0	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{85}$	$\sqrt{101}$	$\sqrt{109}$	$\sqrt{29}$	$\sqrt{37}$	$\sqrt{45}$
Pt 8	$\sqrt{37}$	$\sqrt{34}$	$\sqrt{20}$	$\sqrt{10}$	4	$\sqrt{20}$	$\sqrt{2}$	0	2

So Pt 1 $\rightsquigarrow C_1$ Pt 2 $\rightsquigarrow C_1$ Pt 3 $\rightsquigarrow C_1$ Pt 4 $\rightsquigarrow C_8$ Pt 5 $\rightsquigarrow C_8$ Pt 6 $\rightsquigarrow C_8$

Pt 7 $\rightsquigarrow C_8$ Pt 8 $\rightsquigarrow C_8$ Pt 9 $\rightsquigarrow C_8$



(b) For example choose as centers Pt 8 and Pt 4:



Question 3

Hierarchical clustering

(a) Execute single-linkage and complete-linkage hierarchical clustering on the following similarity matrix. The more similar two points are, the smaller the distance between them.

	P1	P2	P3	P4	P5
P1	1	0.10	0.41	0.55	0.35
P2	0.10	1	0.64	0.47	0.98
P3	0.41	0.64	1	0.44	0.85
P4	0.55	0.47	0.44	1	0.76
P5	0.35	0.98	0.85	0.76	1

↑ similarity matrix

- Single linkage: if G and H are two clusters $d_{SL}(G,H) = \min_{x_i \in G,\ x_j \in H} d(x_i, x_j)$, in terms of similarity:

$$s_{SL}(G,H) = \max_{x_i \in G,\ x_j \in H} s(x_i, x_j)$$

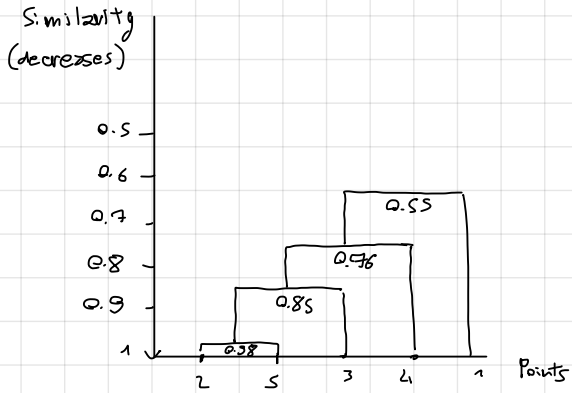
So the first cluster is (P_2, P_5) (similarity 0.98)

After we have $s_{SL}(P_3, (P_2, P_5)) = 0.85$ so we merge them together.

After we have $s_{SL}(P_4, (P_3, P_2, P_5)) = 0.76$ so we merge them together.

After we have $s_{SL}(P_1, (P_2, P_3, P_4, P_5)) = 0.55$

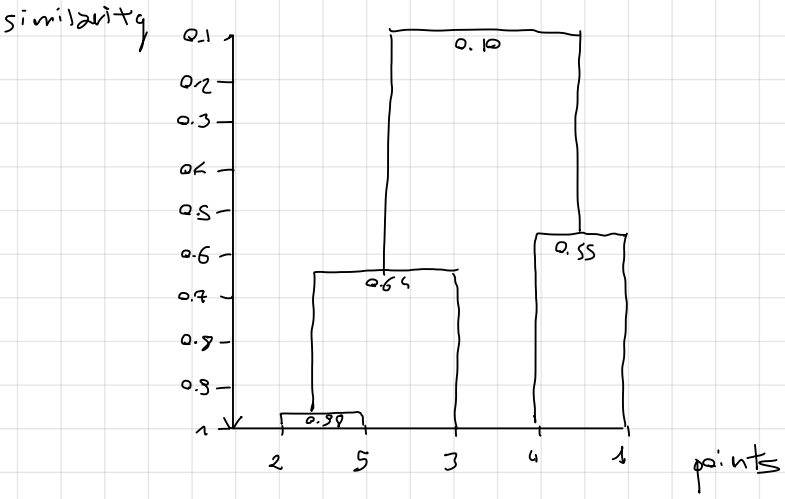
So the dendogram is:



- Complete linkage: if G and H are two clusters $d_{CL}(G,H) = \max_{x_i \in G,\ x_j \in H} d(x_i, x_j)$, in terms of similarity: $s_{CL}(G,H) = \min_{x_i \in G,\ x_j \in H} s(x_i, x_j)$

So we start always with the two most similar points (in the first step clusters are points so max or min is the same). After we want the highest similarity between two clusters using s_{CL}: for example $d_{CL}(P_3, (P_2, P_5)) = 0.64$ and is the highest because $s_{CL}(P_4, (P_2, P_5)) = 0.47$ (it could be tricky because $s(x_5, x_5) = 0.76$ but in complete linkage we take the min.).

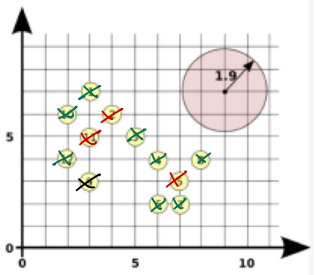
So we merge P_3 with P_2, P_5. For the same reason after we merge P_1, P_4 (similarity 0.55). And at the end we merge all together (similarity 0.10):



Question 4

DBSCAN

Apply the DBSCAN algorithm with radius 1.9 and MinPts=4 (3 neighbors + the point we are considering as center for computing the density). Indicate what is the nature of each point. Draw the clusters.



Core points: 2, 5, 11

Border point: 1, 3, 4, 6, 7, 8, 10

Outliers: 9

Question 5

Principal component analysis

is it symmetric?

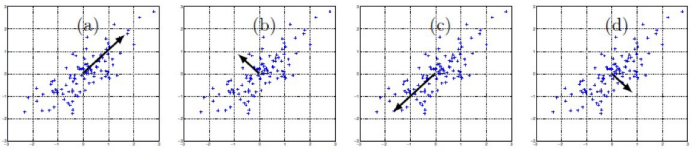
(a) Let Σ be a positive semi-definite d-dimensional matrix, with eigenvalues λ_i, $i = 1, ..., d$. Show that

$$\text{trace}(\Sigma) = \sum_{i=1}^{d} \lambda_i.$$

(b) Let $\boldsymbol{x} = (x_1, ..., x_d)^T$ denote a d-dimensional random vector with variance-covariance matrix Σ. Let $p_1 = \gamma_1^T \boldsymbol{x}$ be the first principal component of $\boldsymbol{x}$. Show that γ_1 solves

$$\max_{\boldsymbol{a}} Var(\boldsymbol{a}^T \boldsymbol{x})$$
$$\text{subject to } \boldsymbol{a}^T \boldsymbol{a} = 1.$$

(c) Which of the following figures correspond to possible values that PCA may return for the first principal component?



(a) Since Σ is symmetric admits : $\Sigma = Q^T D Q$ where $QQ^T = id$ and $D = diag(\lambda_1, \dots, \lambda_n)$

$\{\lambda_i\}_{i=1}^n$ eigenvalues. So:

$$tr(\Sigma) \underset{\text{def.}}{=} \sum_{i=1}^n e_i^T \Sigma e_i = \sum_{i=1}^n \underbrace{e_i^T Q^T}_{(Qe_i)^T} D \underbrace{Q e_i}_{Q^i} =$$

$$= \sum_{i=1}^n Q^{i\,T} D Q^i = \sum_{i=1}^n \lambda_i \underbrace{Q^{i\,T} Q^i}_{1} = \sum_{i=1}^n \lambda_i$$

(b) $\max_a Var(a^T x)$ subject to $a^T a = 1$ ← normalization :

$$Var(a^T x) = Cov\left(\sum_{i=1}^d a_i x_i, \sum_{j=1}^d a_j x_j\right) = \sum_{i,j=1}^d a_i a_j Cov(x_i, x_j) = a^T \Sigma a$$

So we can write the lagrangian

$$L(a, \lambda) = a^T \Sigma a - \lambda(a^T a - 1)$$

$$\nabla_a L = 2\Sigma a - 2\lambda a = 0 \Rightarrow \Sigma a = \lambda a \quad (a \text{ has to be an eigenvector of } \Sigma)$$

So $Var(a^T x) = a^T \Sigma a = \lambda a^T a = \lambda$, so to maximize it we have to choose $\lambda = \lambda_1$ (the greatest eigen value) $\Rightarrow a = \gamma_1$ (the corresponding eigen vector).

(c) The maximum variance is along $y = x$ so (a) and (c) are correct, (b) and (d) incorrect.

FIN-407 Machine Learning in Finance

Exercise sheet 7 - Neural networks

Assistant: Giuseppe Matera.

Faculty: Elise Gourier.

Question 1

Representation of functions

(a) The Fourier series of a function F has the form

$$f(x) = \sum_{i=0}^{\infty} (a_i \cos(ix) + b_i \sin(ix))$$

Show that an artificial neural network with the sine as primitive function can implement a finite number of terms in the above expression.

(b) What is the main difference between Taylor or Fourier series and artificial neural networks?

Question 2

Terminology of neural networks

Assume you have a dataset with 200 samples and you choose a batch size of 5 and 1,000 epochs.

(a) How many batches are there? How many samples does each batch contain? Based on how many batches will the model weights be updated?

(b) How many batches does one epoch contain? How many updates of the model are done within one epoch?

(c) How many times will the model pass through the whole dataset? How many batches will the model pass through during the training phase?

Question 3

Feedforward networks

Suppose you have a feedforward network composed of one input layer with 10 neurons, followed by one hidden layer with 15 neurons, and finally one output layer with 3 neurons. All neurons in the hidden layer and output layer use the ReLU activation function.

(a) What is the shape of the input matrix $\boldsymbol{X}$?

(b) What are the shapes of the hidden layer's weight vector $\boldsymbol{W}_h$ and its bias vector $\boldsymbol{b}_h$?

(c) What are the shapes of the output layer's weight vector $\boldsymbol{W}_o$ and its bias vector $\boldsymbol{b}_o$?

(d) What is the shape of the network's output matrix $\boldsymbol{Y}$?

(e) Write the equation that computes the network's output matrix $\boldsymbol{Y}$ as a function of $\boldsymbol{X}$, $\boldsymbol{W}_h$, $\boldsymbol{b}_h$, $\boldsymbol{W}_o$ and $\boldsymbol{b}_o$.

Question 4

Feedforward networks and autoregressive models

(a) Design a neural network that reproduces an AR(3) model.

(b) How would you change this network to account for seasonality?

(c) What is the main limitation of this model and how can a neural network overcome this limitation?

(d) Consider a neural network with activation function

$$g(h) = \begin{cases} 1 & h > 0 \\ 0 & h \leq 0 \end{cases}$$

Why is this activation function not used with backpropagation?

(e) Consider the sigmoid activation function. Show that it satisfies

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Justify why it is important to standardize variables when using the sigmoid function

Question 5

Logistic regression

Remember that logistic regression is a classification model that learns to predict from a set of features x to which class a data point belongs. It is a linear and discriminative model.

(a) Draw a neural network which takes as inputs the component of x and produces as output the output of a logistic regression, $P(y = 1|x)$.

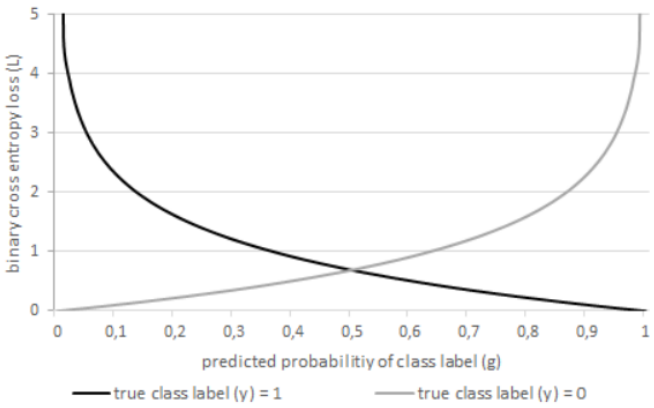
Consider the binary cross entropy as loss function:

$$L(y, g) = -y \ln(g) - (1 - y) \ln(1 - g)$$

where g is the predicted probability that the class label is 1 and y is the true class label. The total loss over the training sample is $\sum_i L(y_i, g_i)$.

The binary cross entropy loss function decreases if the predicted probability for the class labels get closer to the true class labels and is therefore an appropriate measure to monitor the learning progress and convergence of the model.

(b) You are given a set of training data and want to estimate the weights and biases of the network from these data. Describe the first forward pass through the network.



(c) Describe the backward pass through the network, used to adjust weights and biases in the direction of a local minimum.

(d) Describe how the forward pass and the backward pass are applied in practice. Does the algorithm always converge to a global minimum?

We create a random data set with 100,000 samples and two different classes generated by two bivariate normal distributions with $\mu = (0, 0)$ and $\mu = (2, 8)$, with no correlation between x_1 and x_2. Two independent models are trained for 100 epochs. In the first one, the data is centered around 0. In the second one it is not.

(e) What do you think is preferable? Would you apply the first or the second model? Why?

(f) Which plots can you use to study the convergence of the training method?

Question 6

Recurrent networks

(a) Consider an LSTM architecture. Suppose you want the memory cell to sum its inputs over time in the long-term state. What values should the input gate and forget gate take?

Question 1

Representation of functions

(a) The Fourier series of a function F has the form

$$f(x) = \sum_{i=0}^{\infty} (a_i \cos(ix) + b_i \sin(ix))$$

Show that an artificial neural network with the sine as primitive function can implement a finite number of terms in the above expression.

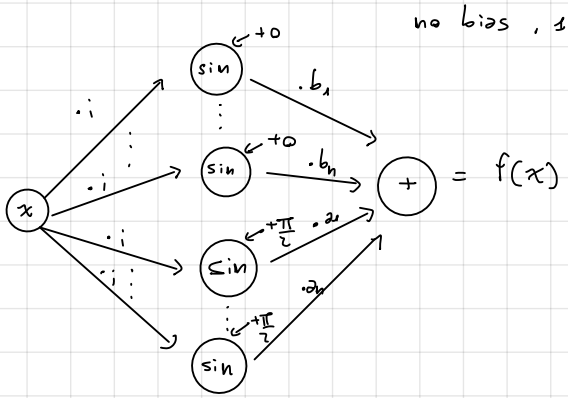
(b) What is the main difference between Taylor or Fourier series and artificial neural networks?

(a) Suppose we truncate the Fourier Series at n:

$$f(x) = \sum_{i=0}^{n} \left(a_i \cos(ix) + b_i \sin(ix) \right)$$

So, we note that $\cos(ix) = \sin\left(ix + \frac{\pi}{2}\right)$, then:

no bias, 1 hidden layer



(b) The neural networks do not know the function f a priori, we just know some points and we want to find the "right" parameters that describe well the data we have and at the same time can do good prediction.

Question 2

Terminology of neural networks

Assume you have a dataset with 200 samples and you choose a batch size of 5 and 1,000 epochs.

(a) How many batches are there? How many samples does each batch contain? Based on how many batches will the model weights be updated?

(b) How many batches does one epoch contain? How many updates of the model are done within one epoch?

(c) How many times will the model pass through the whole dataset? How many batches will the model pass through during the training phase?

(a) • # batches = $\frac{200}{5} = 40$

- size of batch = 5
- Update every batch : 1

(b) Epoch : n° of batches to do all the dataset ⇒ 40 . Same for update: 40

(c) Every epoch the model pass through the whole dataset ⇒ 1000 times.

batches for 1000 epochs = 40 · 1000 = 40.000

Question 3

Feedforward networks

Suppose you have a feedforward network composed of one input layer with 10 neurons, followed by one hidden layer with 15 neurons, and finally one output layer with 3 neurons. All neurons in the hidden layer and output layer use the ReLU activation function.

(a) What is the shape of the input matrix $\boldsymbol{X}$?

(b) What are the shapes of the hidden layer's weight vector $\boldsymbol{W}_h$ and its bias vector $\boldsymbol{b}_h$?

(c) What are the shapes of the output layer's weight vector $\boldsymbol{W}_o$ and its bias vector $\boldsymbol{b}_o$?

(d) What is the shape of the network's output matrix $\boldsymbol{Y}$?

(e) Write the equation that computes the network's output matrix $\boldsymbol{Y}$ as a function of $\boldsymbol{X}$, $\boldsymbol{W}_h$, $\boldsymbol{b}_h$, $\boldsymbol{W}_o$ and $\boldsymbol{b}_o$.

(a) Input matrix $X \in \mathbb{R}^{m \times 10}$ where in each column I have a point from the batch (m = # batch)

(b) $W_h \in \mathbb{R}^{10 \times 15}$ (each column is the vector weight of a node)

$b_h \in \mathbb{R}^{15}$

(c) $W_o \in \mathbb{R}^{15 \times 3}$ and $b_o \in \mathbb{R}^3$

(d) $Y \in \mathbb{R}^3$

(e) $\hat{Y} = \text{ReLu}(F_{W,b}(X)) = \text{Relu}\left(f^{(o)}_{W_o, b_o} \circ f^{(h)}_{W_h, b_h}(X)\right) = f^{(o)}_{W_o, b_o}\left(W_h\left(\underbrace{\text{Id}(X)}_{\text{activation function of input layer} = \text{Id}}\right) + b_h\right) =$

$\hat{Y}$ without error; $f^{(o)}_{W_o, b_o}$ output; $f^{(h)}_{W_h, b_h}$ hidden

$= \text{Relu}\left(f^{(o)}_{W_o, b_o}(XW_h + b_h)\right) = \text{Relu}\left(W_o\left(\underbrace{\text{ReLu}(XW_h + b_h)}_{\in \mathbb{R}^{m \times 15}\ (\text{Relu set to 0 negative entrances of the matrix})}\right) + b_o\right) =$

$W_h \in \mathbb{R}^{10 \times 15}$, $X \in \mathbb{R}^{m \times 10}$

$\Rightarrow W_h(X) = XW_h$

we add $b_h \in \mathbb{R}^3$ to every single row of $XW_h \in \mathbb{R}^{m \times 15}$

$$= \text{Relu}\left(\underbrace{\left(\text{Relu}\left(X W_h + b_h\right)\right) W_o}_{\in \mathbb{R}^{m\times 3}} + b_o \right) \quad \left(Y^* \in \mathbb{R}^{m\times 3}\right)$$

Question 4

Feedforward networks and autoregressive models

(a) Design a neural network that reproduces an AR(3) model.

(b) How would you change this network to account for seasonality?

(c) What is the main limitation of this model and how can a neural network overcome this limitation?

(d) Consider a neural network with activation function

$$g(h) = \begin{cases} 1 & h > 0 \\ 0 & h \leq 0 \end{cases}$$

Why is this activation function not used with backpropagation?

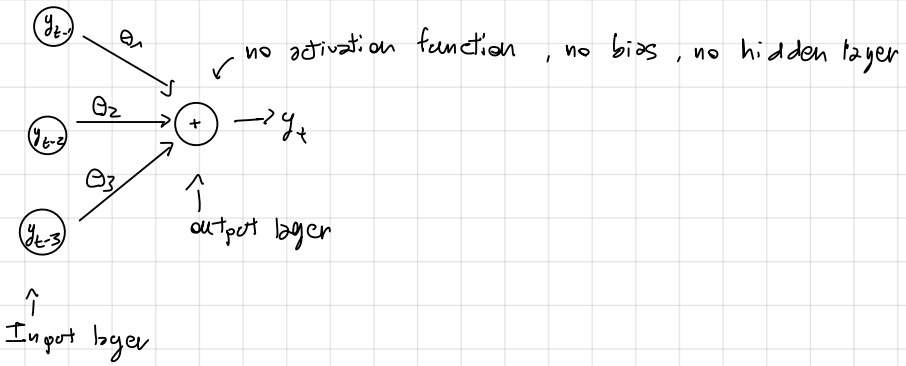
(e) Consider the sigmoid activation function. Show that it satisfies

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Justify why it is important to standardize variables when using the sigmoid function

(a) The AR(3) is a model s.t. $y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \theta_3 y_{t-3}$.

So:

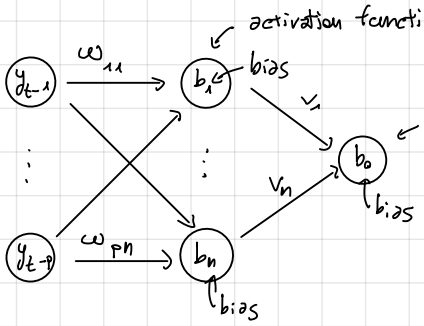


(b) If you want to capture seasonality (if t are the months and you suppose seasonality of 1 year) you can add to the input nodes y_{t-12}, y_{t-24} ...

(c) The main limitation of AR(3) model is the linearity. A neural network can improve this method: suppose we have an AR(p) model:

$$y_t = \theta_1 y_{t-1} + \dots + \theta_p t_{t-p}$$

Now we can choose a neural network of this form:



$\{\omega_{ij}\}_{\substack{i=1,n \\ j=1,p}}$, $v_1, \dots, v_n$ weights.

Now, the output in node b_i is:

$$z_i = \sigma\left(\sum_{j=1}^{p} \omega_{ji}\, y_{t-j} + b_i\right)$$

So the final output is:

$$\hat{y}_t = \beta_0 + \sum_{i=1}^{n} v_i z_i = \beta_0 + \sum_{i=1}^{n} v_i\, \sigma\left(b_i + \sum_{j=1}^{p} \omega_{ij}\, y_{t-j}\right)$$

(d) We want an activation function that is differentiable and such that his derivative does not vanish.

(e)

$$\sigma(x) = \frac{e^x}{e^x + 1}$$

$$\sigma'(x) = \frac{e^x(e^x+1) - e^x(e^x)}{(e^x+1)^2} = \frac{e^x\left(\cancel{e^x}+1-\cancel{e^x}\right)}{(e^x+1)^2} = \sigma(x)\,\frac{1}{e^x+1}$$

but $1-\sigma(x) = 1 - \frac{e^x}{e^x+1} = \frac{\cancel{e^x}+1-\cancel{e^x}}{e^x+1}$ $\Rightarrow \sigma'(x) = \sigma(x)\cdot\left(1-\sigma(x)\right)$

So, when $x \to +\infty \Rightarrow \sigma'(x) \longrightarrow 1\left(1-1\right) = 0$ and when $x \longrightarrow -\infty$ $\sigma'(x) \longrightarrow 0\left(1-0\right) = 0$, so there is the vanishing gradient problem $\Rightarrow$ we have to standardize x.

Question 5

Logistic regression

Remember that logistic regression is a classification model that learns to predict from a set of features x to which class a data point belongs. It is a linear and discriminative model.

(a) Draw a neural network which takes as inputs the component of x and produces as output the output of a logistic regression, $P(y = 1|x)$.

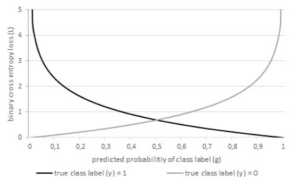
Consider the binary cross entropy as loss function:

$$L(y, g) = -y \ln(g) - (1 - y) \ln(1 - g)$$

where g is the predicted probability that the class label is 1 and y is the true class label. The total loss over the training sample is $\sum_i L(y_i, g_i)$.

The binary cross entropy loss function decreases if the predicted probability for the class labels get closer to the true class labels and is therefore an appropriate measure to monitor the learning progress and convergence of the model.

(b) You are given a set of training data and want to estimate the weights and biases of the network from these data. Describe the first forward pass through the network.



(c) Describe the backward pass through the network, used to adjust weights and biases in the direction of a local minimum.

(d) Describe how the forward pass and the backward pass are applied in practice. Does the algorithm always converge to a global minimum?

We create a random data set with 100,000 samples and two different classes generated by two bivariate normal distributions with $\mu = (0, 0)$ and $\mu = (2, 8)$, with no correlation between x_1 and x_2. Two independent models are trained for 100 epochs. In the first one, the data is centered around 0. In the second one it is not.

(e) What do you think is preferable? Would you apply the first or the second model? Why?

(f) Which plots can you use to study the convergence of the training method?

(a) Remember that in logistic regression $P(y=1|x) = \frac{e^{w^T x}}{1+e^{w^T x}}$

Input — Output layer

$x \xrightarrow{w} \boxed{\begin{matrix} b=0 \\ \sigma \end{matrix}} \longrightarrow$ Output: $P(y=1|x)$

bias $b=0$

activation function $\sigma(z) = \frac{e^z}{1+e^z}$

(b) Suppose we train also with a bias:

- we initialize w, b randomly
- forward: use mini-batch, so take an input x: we multiply it by the weight w obtaining $w^T x$ and adding the bias term $w^T x + b$.

After we apply the sigmoid function that in this case is $\sigma(z) = \frac{e^z}{1+e^z}$

$\Rightarrow$ the output is $\frac{e^{w^T x + b}}{1 + e^{w^T x + b}}$

- Compute the error $L(y, g) = -y \ln\left(\frac{e^{w^T x + b}}{1 + e^{w^T x + b}}\right) - (1-y) \ln\left(1 - \frac{e^{w^T x + b}}{1 + e^{w^T x + b}}\right) =$

$= -y \ln\left(\frac{e^{w^T x + b}}{1 + e^{w^T x + b}}\right) - (1-y) \ln\left(\frac{1}{1 + e^{w^T x + b}}\right)$

- Also, in this step we save the output (a^k) and $w^T x + b$ (z^k).

(c) Now we have to compute:

$\frac{\partial L}{\partial w} = \delta \cdot x$ ← is a_k^{l-1} where $\delta = \frac{\partial L}{\partial z}$ that in our case

$z = w^T x + b$ so we have to compute:

$$\frac{\partial L(y,g)}{\partial z} = \frac{\partial}{\partial z}\left(-y \ln\left(\frac{e^z}{1+e^z}\right) - (1-y)\ln\left(\frac{1}{1+e^z}\right)\right) =$$

$$= -y \frac{1+e^z}{e^z} \cdot \frac{e^z(1+e^z) - e^z e^z}{(1+e^z)^2} - (1-y)(1+e^z)\left(-\frac{e^z}{(1+e^z)^2}\right) =$$

$$= -y \cdot \frac{1}{1+e^z} + (1-y)\frac{e^z}{1+e^z}$$

So $\frac{\partial L}{\partial w} = \left((1-y)\frac{e^z}{1+e^z} - \frac{y}{1+e^z}\right) w$

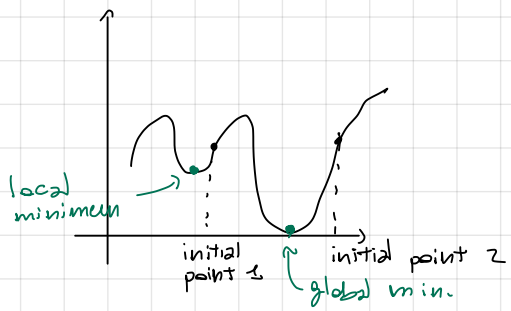
So $w_{new} = w_{old} - lr \cdot \frac{\partial L}{\partial w}$ with lr the learning rate.

For the bias: $\frac{\partial L}{\partial b} = \delta = \left((1-y)\frac{e^z}{1+e^z} - \frac{y}{1+e^z}\right)$ so

$$b_{new} = b_{old} - lr \cdot \frac{\partial L}{\partial b}$$

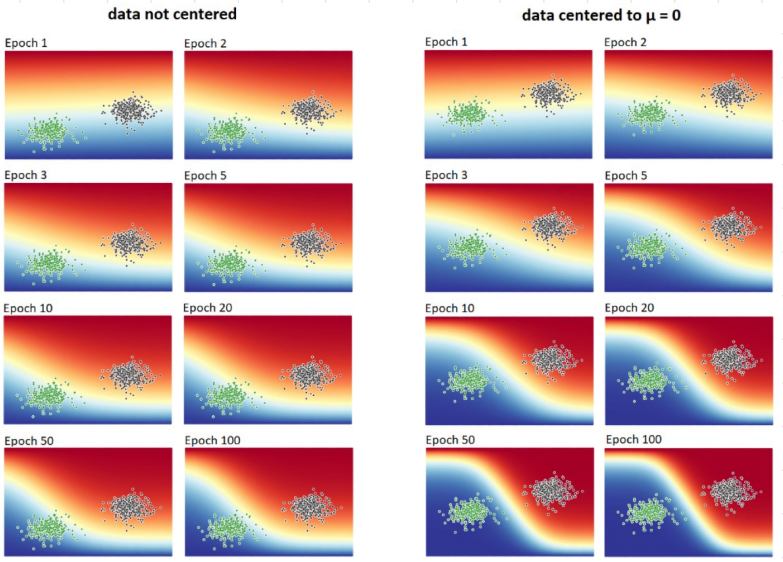
(d) We apply these steps many times to reach a minimum that is not always global.

It could depend on the initial point:



If the function is strictly convex it is a global minimum.

(e) When we apply a neural network is always better to standardize the data $\Rightarrow$ the data is centered, to improve convergence:



(f) We can plot the loss function with respect to the epochs.

Question 6

Recurrent networks

(a) Consider an LSTM architecture. Suppose you want the memory cell to sum its inputs over time in the long-term state. What values should the input gate and forget gate take?

(a) Start at $t=0 \Rightarrow S_0 = x_0$, then $t=1$ $S_1 = x_1 + S_0$ ...

time n $S_n = x_n + S_{n-1}$ so I want to remember all the output S_{n-1}

$\Rightarrow f_n = 1$ (function of forget gate, when $=1$ it does not eliminate anything)

and $i_n = 1$ (function of input gate, we want to update all the vector $\Rightarrow = 1$)