Submission Date: 05th November 2021

1 LDA general

1. What are the assumptions made by the LDA algorithm regarding the data distribution of the different classes?

2. Could you mention a condition/scenario that would facilitate the application of LDA? How about a condition that makes it more difficult?

3. What would you say are some of the main differences between the LDA and the KNN algorithm?

2 LDA calculation

We want to use Linear Discriminant Analysis for a real world problem. Suppose the RELEA lab wants to open a new Doner place where you can pick your own ingredients for your Doner. The lab members notice that most customers are comming in wearing either a NASA or ALDI t-shirt. They record the selections made by 12 random customers using ounces (because they love UK).

Customer	Meat	Vegetatrian	Logo
1	4.9	10	NASA
2	4.7	15	NASA
3	4.5	12	NASA
4	5.8	25	NASA
5	4.9	10	NASA
6	4	15	NASA
7	5	43	ALDI
8	8.2	45	ALDI
9	8.7	50	ALDI
10	6.9	55	ALDI
11	7.2	52	ALDI
12	9	51	ALDI

1. Calculate means $\boldsymbol{m}_k$ and covariance matrices $\boldsymbol{S}_k$ and $\boldsymbol{S}_W$.

2. Calculate $\boldsymbol{w}$ and b using the analytical form.

3. Use your estimations to guess the logo for customer 13: Meat 6, Veg. 25.

4. Sketch the data and the query point.

3 Ordinary Least Squares Regression

The formulation of Ordinary Least Squares Regression (OLSR) algorithm in augmented vector notation:

$$y = Xw + \epsilon \tag{1}$$

1. In its simplest form, what kind of interactions can the OLSR algorithm model? Can you mention an interaction that the algorithm cannot model well? How can the algorithm perform better in the latter situation?
2. What are the assumptions that are made to generate the analytical solution? Can you derive the formula for the analytical solution?
3. Can you mention a situation where the squares error loss function would not perform well? What other loss functions can be used in that scenario?

4 Logistic Regression

1. What is the input and the output space of logistic regression? Hint: we are looking for something like $X \in \mathbb{N}$
2. Recap gradient descent: What is it and why do we use it?
3. Starting from negative loglikelihood (binary cross entropy loss), derive the update rule for w for gradient descent. Ignore the bias term for the moment. Why is there no closed form solution for logistic regression as there is for linear regression? Notes:

   Negative loglikelihood:

   $$J = -\sum_{n=1}^{N} y_n \log p_n + (1 - y_n) \log(1 - p_n) \tag{2}$$

   with $p_n = h_{\boldsymbol{w}}(\boldsymbol{x}_n) = P(y = 1 | X = \boldsymbol{x}_n; \boldsymbol{w}) = \sigma(\boldsymbol{x}_n)$
   For sigmoid use

   $$\sigma(\boldsymbol{x}_n) = \frac{e^{\boldsymbol{x}_n \boldsymbol{w}}}{1 + e^{\boldsymbol{x}_n \boldsymbol{w}}} \tag{3}$$

   which is equivalent to $\frac{1}{1+e^{-\boldsymbol{x}_n \boldsymbol{w}}}$ (try it)
4. Perform one gradient descent update step for using the initial parameters as $\boldsymbol{w}^{(0)} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$ and learning rate $\alpha = 0.5$ using the following data:

x_1	x_2	y
2	4	1
3	3	1
-4	-2	0
-2	-6	0

Note: Think about a good way to include the bias in the data instead of deriving its update rule on its own.

Predict $P(y = 1|X = \begin{bmatrix}-1 & 1\end{bmatrix})$

5 Linear Separability

Given the following data, where x_1, x_2 are the inputs and y is the output:

x_1	x_2	y
1	0	1
0	0	0
0	1	1
1	1	0

We want to learn a linear classification model such that w_1, w_2, b are the parameters:

$$\hat{y} = \begin{cases} 1 & x_1 \cdot w_1 + x_2 \cdot w_2 + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(4)

1. Why is this problem not linearly separable using the proposed model?
2. How would you modify the problem to do it linearly separable?
3. Specify one set of weights (w_1, w_2, b) that classify the input correctly