CS-233 Theoretical Exercise 4

1 When will CFF trains break down?

You are given the task to predict whether a CFF train will break down or not under certain weather conditions. The dataset, represented by $\{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$, has $N = 1000$ data entries, and is explained as follows: For each data entry $(\boldsymbol{x}^{(i)}, y^{(i)})$,

- $\boldsymbol{x}^{(i)}$ has 5 features: Train line, time of the day, temperature of the day, precipitation of the day, and maximum wind speed of the day.
- $y^{(i)}$ is either 0, which means that the train reaches its final destination with a delay of less than 3 minutes (negative - not broken down), or 1 if the train is delayed by more than 3 minutes or is cancelled (positive - broken down).

The dataset has 900 cases where $y^{(i)} = 0$ and 100 cases where $y^{(i)} = 1$. On this dataset, your model has the following performance:

- Among the 900 cases where the train doesn't break down, i.e., $y^{(i)} = 0$, you successfully predicted 800 cases;
- Among the 100 cases where the train breaks down, i.e., $y^{(i)} = 1$, you successfully predicted 10 cases.

Now let us evaluate your prediction model!

Question 1: Draw the confusion matrix.

Question 2: What is the accuracy of the model?

Question 3: What is the precision of the model?

Question 4: What is the recall of the model?

Question 5: What is the F1 score of the model?

Question 6: Do you think this model is a good model or a bad model ? Explain your reasoning.

2 Multiclass classification

Your boss wants to differentiate a canceled train from a delayed train. Therefore, the dataset is now labeled with three classes:

- 0 - a train on time;
- 1 - a train delayed;
- 2 - a train canceled.

FOr this task, you decide to use a least-square classifier with a one-hot encoding of the label. The prediction for the entire dataset (same dataset as in the previous question) can be formalized in matrix form as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X} \cdot \boldsymbol{W} .$$

Question 1. Write down the one-hot encoding of each class.

Question 2. What are the shapes of $\hat{\boldsymbol{Y}}$, $\boldsymbol{X}$ and $\boldsymbol{W}$? Note that we append a 1 to each input when performing classification to account for the bias.

Question 3. Write down the loss function for the MSE loss. Given a learning rate η, what is the update of a single gradient descent step? When the gradient descent algorithm converges (i.e., the gradient goes to 0), what is the final $\boldsymbol{W}$? (Hint: Solve for each column vector of $\boldsymbol{W}$.)

3 Connection between binary and multi-class logistic regression

Consider a binary logistic regression task. Your model (model 1) using a sigmoid function has a weight vector $\boldsymbol{w}$. If you now use another model (model 2) with a one-hot representation and a soft-max function, is there a weight matrix $\boldsymbol{W}$ such that model 2 has the same decision boundary as model 1? If not, explain your reasoning; if yes, compute $\boldsymbol{W}$.