Assignment 2

***Note 1**: for answers with Python, display both codes and results clearly. You may explain answers with either comments, markdowns or print() function.*

***Note 2**: for answers with manual calculation, please display all calculation steps clearly.*

***Note 3**: round all numerical answers to 2 decimal places.*

Question 1. Logistic Regression [30 points]

Suppose we collect data for a group of students in a course with variables $X_1 = hours\ studied$, $X_2 = Cumulative\ GPA$, and $y = 1$ if the student receives an A in this course and $y = 0$ otherwise. We trained a logistic regression model on this dataset to predict whether a student can receive A with features X_1 and X_2, the estimated parameters are $w_0 = -7, w_1 = 0.05, w_2 = 1$. Answer the following questions with either Python or manual calculation.

(a) What is the probability that a student who studies for 40 hours and has a cumulative GPA of 3.5 gets an A in the class?

(b) How many hours would the student in step (a) need to study to have a 50% chance of getting an A in the class?

(c) What is the odds ratio and log-odds for the student in (a)? And what does the odds ratio mean?

(d) Visualize the model (the linear hyperplane) with a simple line plot where X_1 is on the x-axis and X_2 on the y-axis. Also indicate the region for A grade (positive, $y = 1$) and region for non-A grade (negative) in the figure. (***Hint**: you can draw the plot manually or with the help of any software.*)

Question 2. Odds-ratio and Probability [20 points]

Please answer the following questions either Python or manual calculation.

(a) Suppose that an individual has a 20% chance of defaulting on her credit card (positive, $y = 1$) payment. What is the odds-ratio of defaulting?

(b) Suppose the odds-ratio of defaulting on credit card payment for a man is 0.4, what is the probability this person will default on his credit card payment?

Question 3. Support Vector Machine [20 points]

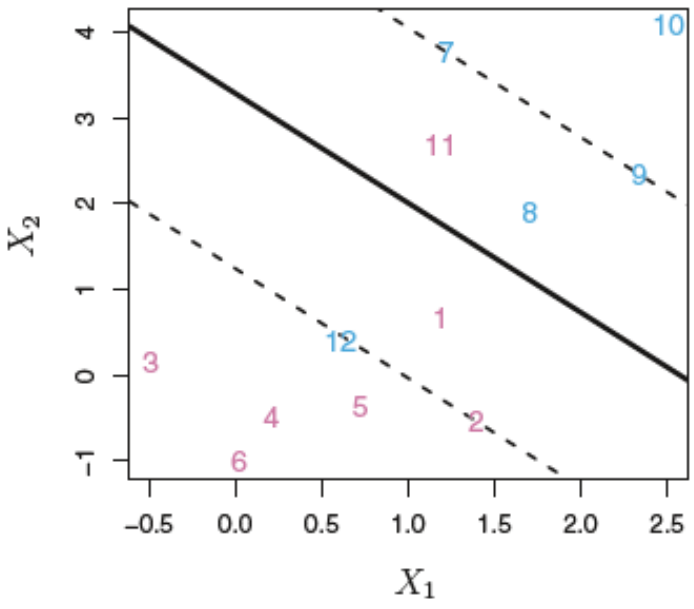
A support vector classifier was fitted to a small data set with 12 instances, each represented by a colored number in the following figure. Their colors indicate their classes (blue represents positive, red represents negative). The hyperplane (solid line) and the two margins (dashed lines) are plotted as well. Please answer the following questions with your own words.

(a) List the number of instances that are support vectors in this model.

(b) Suppose instance 4 (i.e., the red number 4 in the figure) moves closer to its margin. Will that affect the position of the hyperplane?

(c) List the number of instances that will bring hinge loss to but NOT be misclassified.

(d) List the number of instances that would be misclassified.



Assignment 2

Question 4. Logistic Regression [30 points]

We'd like to use the dataset *smoking.csv* to build a logistic regression model to predict whether a person smoke or not (***smoker***) with three features: ***smkban*** (i.e., whether smoking was banned in this person's country), ***female*** (i.e., whether the person is female or not) and ***age*** (i.e., age of this person)**.** Please answer the following questions with Python.

(a) Load the data into Python. How many instances are there? How many smokers (i.e., $smoker = 1$) and non-smokers (i.e., $smoker = 0$) respectively?

(b) Select relevant features and target as listed above, split the data into train (80%) and test (20%) set, and set random_state = 2025. Scale the features with *MinMaxScaler*.

(c) Train a logistic regression model and display all parameter values. How would you interpret three coefficients?

(d) Check model accuracy on both the train and test set.

(e) A male is aged 48 and living in a country where smoking was not banned, will he smoke? Please also estimate how likely this person will smoke as well. *(**Hint**: you may need to arrange his feature values in a 2D array and transform with the scaler before making predictions.)*