Fall 2024

COSC522 – Homework #4 (Due 11/07)

Problem 1: (55 pts) This problem is 2-fold. First, it is designed to compare the decision boundaries made by the linear machines. How many linear machines we have covered in this semester? Hope your answer is 4 – minimum Euclidean distance classifier (MD), minimum Mahalanobis distance classifier, perceptron, and SVM. Use the 4 samples in an AND gate as the training set (assume the samples are located in the first quadrant and the labels are either 0 or 1).

(1) (20 pts) On the same figure (Fig. 1), plot the four samples of the AND gate and show the decision boundary from MD, Perceptron, and SVM on the training samples. Show details. That is, need detailed calculation or reasoning to support your plot. Note that the boundaries from MD and SVM are unique but that from Perceptron might vary (why?)

(2) (28 pts) On another figure (Fig. 2), show the projection direction derived from FLD and PCA. Note that FLD and PCA are dimensionality reduction methods that only output a projection direction. Additional classification methods need to be applied to find the decision boundary. On the same figure, show the decision boundaries from FLD+MD (10 pts) and PCA+MD (10 pts).

(3) (7 pts) Add the two decision boundaries in Fig. 2 back to Fig. 1 and comment on the result.

You can use whichever language that you feel comfortable (pencil & paper or Python).

Problem 2: (15 pts) An experienced machine learning researcher would be able to get a hint of potential issues with a classification model based on precision and recall. For a two-class classification problem, explain what might have caused the following scenarios:

a. High precision but low recall

b. Low precision but high recall

FYI, precision = TP/(TP+FP), recall = TP/(TP+FN)

Problem 3: (30 pts) The dataset is taken from Wikipedia's Logistic regression site, that was an interesting toy problem to solve.

A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?

Hrs (x_k)	1.0	2.0	3.0	4.0	5.0

Prob (p_k)	0.07	0.26	0.61	0.87	0.97
Pass (y_k)	0	0	0	1	1

a) (10 pts) Use simple linear regression to predict the probability of pass if a student studied for 3.25 hours. Use pencil and paper to write down the detailed steps. You can use calculator or numpy for the actual calculation though.

b) (20 pts) Use logistic regression to solve the above problem. Again, please provide details.