IE 7275 Data Mining in Engineering
Homework 3
Deadline: 11/05/2021

Note:

- **Submit all necessary files along with your solution sheets on Canvas.**
- **You use Python or equivalent program languages such as R only.**

Problem 1 (Multiple Linear Regression) [15 points]

See Problem 6.1 in the textbook: Predicting Boston Housing Prices. The dataset was provided in Homework 02.

Problem 2 (Logistic Regression) [15 points]

See Problem 10.1 in the textbook: Financial Condition of Banks

Problem 3 (Linear Discriminant Analysis) [15 points]

See Problem 12.1 in the textbook: Personal Loan Acceptance

Problem 4 (Linear Discriminant Analysis) [20 points]

Consider the data in the following table. Answer the following questions:

(a) Compute $\boldsymbol{\mu}_{+1}$ and $\boldsymbol{\mu}_{-1}$, and $\mathbf{S}_B$, the between-class scatter matrix.

(b) Compute $\mathbf{S}_{+1}$ and $\mathbf{S}_{-1}$, and $\mathbf{S}_W$, the within-class scatter matrix.

(c) Find the best direction $\mathbf{w}$ that discriminates between the classes. Use the fact that the inverse of the matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given as $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

(d) Having found the direction $\mathbf{w}$, find the point on $\mathbf{w}$ that best separates the two classes.

Note: you need to elaborate all details although you are allowed to verify your solution with Python or R.

	x1	x2	Class
1	0.54	0.78	1
2	0.78	0.71	1
3	0.59	0.92	1
4	0.56	0.88	1
5	0.79	1.00	1
6	1.00	0.81	1
7	0.98	0.75	1
8	0.75	0.95	1
9	0.63	0.85	1
10	0.84	0.88	1
11	0.46	0.93	1
12	0.74	0.85	1
13	0.68	0.83	0
14	0.48	0.88	0
15	0.59	0.73	0
16	0.39	0.86	0
17	0.76	0.75	0
18	0.45	0.75	0
19	0.54	0.68	0
20	0.60	0.78	0
21	0.43	0.69	0
22	0.30	0.80	0
23	0.46	0.59	0
24	0.57	0.63	0

Problem 5 (SVM) [15 points]

Consider the dataset in the following table.

(a) Write down a soft-margin SVM model with the training data.

(b) Solve the SVM model using Python package or equivalent programming language and provide the optimal hyperplane and support vectors.

(c) Check the optimality using KKT conditions.

(d) Draw all data points and the hyperplane on a 2-dimensional space, and highlight the support vectors.

A	B	Class
3.5	4.0	1
2.0	4.0	1
2.0	6.0	1
1.5	7.0	1
7.0	6.5	1
2.1	2.5	0
8.0	4.0	0
9.1	4.5	0

Problem 6 (SVM) [20 points]

Consider a dataset of Personal Loan Acceptance (1500 records)

Apply a cross validation to build a SVM model and to make a prediction

(a) Show the function for SVM hyperplane from 1000 training records. Specify your setting in building a SVM model.

(b) What are the support vectors?

(c) What are training and testing accuracy, respectively?