

**Instructions**

- This homework assignment is worth 90 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: October 28, 2022 by 11:59 pm.**

**Exercise 1**

(5 points) What is a perceptron? Be specific.

**Exercise 2**

(5 points) What are the different types of perceptrons? Briefly describe each of them.

**Exercise 3**

(5 points) What is a hard margin in a support vector machine model? Be specific.

**Exercise 4**

(4 points) The effectiveness of a support vector machine model depends on:

- (a) kernel
- (b) kernel parameters
- (c) penalty cost parameter
- (d) All of the above
- (e) None of the above

**Exercise 5**

(4 points) What is/are **true** about kernel in SVM?

- (a) Kernel function map low dimensional data into high dimensional space.
- (b) Kernel function map high dimensional data into low dimensional space.
- (c) It is a similarity function.
- (d) (a) and (c)

- (e) (b) and (c)
- (f) (a), (b) and (c)
- (g) None of the above

## Exercise 6

Consider the `framingham.csv` data file. The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

- Demographic:
  - Sex: male or female (Nominal)
  - Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- Behavioral
  - Current Smoker: whether or not the patient is a current smoker (Nominal)
  - Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- Medical (history)
  - BP Meds: whether or not the patient was on blood pressure medication (Nominal)
  - Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
  - Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
  - Diabetes: whether or not the patient had diabetes (Nominal)
- Medical (current)
  - Tot Chol: total cholesterol level (Continuous)
  - Sys BP: systolic blood pressure (Continuous)
  - Dia BP: diastolic blood pressure (Continuous)
  - BMI: Body Mass Index (Continuous)
  - Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
  - Glucose: glucose level (Continuous)

- Predict variable (desired target)
  - 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

**In Python**, answer the following:

- (4 points) Load the data file to you S3 bucket. Using the pandas library, read the csv data file and create a data-frame called **heart**.
- (3 points) Remove observations with missing values.
- (40 points) Using **age**, **currentSmoker**, **totChol**, **BMI**, and **heartRate** as the predictor variables, and **TenYearCHD** as the target variable, do the following:
  - Split the data into train (80%) and test (20%).
  - Using **MinMaxScaler**, transform all the input variables in the train and test datasets to 0-1 scale.
    - Build a multi-layer perceptron model with one single hidden layer with 4 neurons (hyperbolic tangent as the activation function) and softmax as the activation function for the output. Use the stochastic descent gradient as the method to estimate the weights (**optimizer = 'sgd'**) and **metrics = ['accuracy']**. Use **epochs = 100** and **batch\_size = 500** to build the model. After that, use the model to predict on the test dataset. Using 15% as the cut-off value, report the recall of this model.
    - Build a multi-layer perceptron model with one single hidden layer with 4 neurons (ReLU as the activation function) and softmax as the activation function for the output. Use the stochastic descent gradient as the method to estimate the weights (**optimizer = 'sgd'**) and **metrics = ['accuracy']**. Use **epochs = 100** and **batch\_size = 500** to build the model. After that, use the model to predict on the test dataset. Using 15% as the cut-off value, report the recall of this model.
    - Build a support vector machine model using **rbf** as the kernel. After that, use the model to predict on the test dataset. Using 15% as the cut-off value, report the recall of this model.
    - Build a support vector machine model using **poly** as the kernel. After that, use the model to predict on the test dataset. Using 15% as the cut-off value, report the recall of this model.
- (20 points) Repeat part (c) 100 times. Create a visualization that shows the recall value for each of the models at each iteration. Also, report the average recall of each of the model for the 100 repetitions. What model would use to predict **TenYearCHD**?