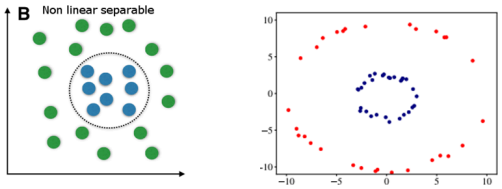
Machine learning
Fall 2023
HW2

PartA - theory questions

1. How can you ensure that your linear regression model is well-generalized and performs well on new data?

2. How can binary classification models be used for multi-class classification tasks? Explain two common methods for using binary classifiers for multi-class classification. (Hint: one-vs-one, one-vs-rest)

3. Why can't linear classification algorithms be used for the data depicted in the image below?



4. What strategies do you suggest for building an effective model when dealing with imbalanced data? How can you ensure the best performance from your model with this type of data and be confident in its performance?

PartB - practical questions

Note: The majority of the points for this exercise are allocated to the data engineering domain; therefore, it is recommended to proceed with appropriate methods such as one-hot encoding, binning, data normalization, handling imbalanced data techniques, and selecting appropriate evaluation metrics for your models and other relevant approaches in performing this task.

1. Stroke, also known as a cerebrovascular accident or CVA, occurs when a part of the brain is deprived of its blood supply, and the part of the body controlled by that area of the brain stops functioning. This loss of blood supply can be due to a lack of blood flow or due to bleeding into brain tissue. Stroke is a medical emergency as it can lead to death or permanent disability. Treatments are available for this type of stroke, but they must be initiated within the first few hours after the onset of stroke symptoms.

The strokes.csv dataset contains information about individuals and their history of stroke. In this problem, perform the following steps:

a) Perform preprocessing operations considering the problem's objective.
b) Split the data into appropriate training and testing datasets.
c) Train a suitable model using the pre-implemented SVM algorithm in the sklearn library.

2. In the health insurance industry, insurance companies often face challenges in determining accurate insurance premiums for each insured individual. Errors in assessing health risks of patients can lead to significant financial losses. Therefore, precise determination of health insurance premiums is crucial for maintaining financial stability for insurance companies and providing fair services to policyholders.

In this project, we will utilize a dataset containing information about health insurance policyholders, including age, gender, Body Mass Index (BMI), number of children, smoking habits, residential area, and individual medical expenses covered by insurance. This dataset serves as a valuable data source for developing predictive models that can assist health insurance companies in assessing risks and determining accurate insurance premiums. This project directly impacts the operational and business strategies of health insurance companies and ultimately provides benefits for both companies and their policyholders.

The objectives of this project include developing machine learning models that can help health insurance companies in the following areas:

- Accurate premium determination: Use policyholder data to calculate insurance premiums based on health risks faced by each insured individual. As a result, insurance companies can minimize financial losses resulting from incorrect insurance premiums.

- Health risk assessment: Identify risk factors affecting individual medical expenses, such as age, BMI, number of children, and smoking habits. This can help insurance companies assess and manage risks more effectively.

The insurance.csv dataset contains information about insured individuals collected by a health insurance company. In this problem, perform the following steps:

a) Perform preprocessing operations considering the problem's objective.

b) Split the data into appropriate training and testing datasets. Report the reason for choosing the percentage split of training and testing data.

c) Implement a linear regression model from scratch based on the concepts taught in class, and report the accuracy percentage of the model on the test data.

d) Extend your previous model to a polynomial regression model and report the accuracy percentage.