SFU SCHOOL OF MECHATRONIC SYSTEMS ENGINEERING

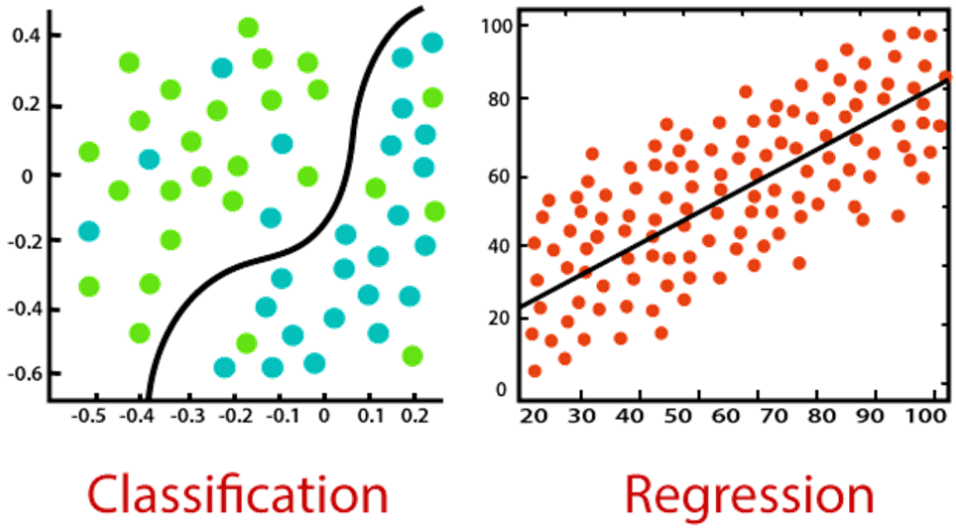
Simon Fraser University
School of Mechatronic Systems Engineering
MSE491 - Application of Machine Learning in Mechatronic Systems

Lab 2 – Classification
Due Date: Mar 07, 2021 23:59

Background

Classification is the process of recognizing, understanding, and grouping ideas and objects into preset categories or sub-populations. Classification algorithms use pre-categorized training datasets to classify future datasets into categories. These techniques use input training data to predict the likelihood that subsequent data will fall into one of the predetermined categories.

Regression and classification are categorized under the same umbrella of supervised machine learning. The main difference between these techniques is that the regression task's output variable is numerical (or continuous) while it is categorical (or discrete) for classification



© https://www.javatpoint.com/regression-vs-classification-in-machine-learning

There are two major classes of classification problems: Binary-class and Multi-class. In Binary-class classifications, the given dataset is categorized into two categories, whereas in Multi-class classification, the given dataset is categorized into several classes based on the classification rules.

MSE 491- Assignment 2 (Spring 2021)

Binary Classification

Heart Failure Clinical Records Data Set

This dataset contains the medical records of 299 heart failure patients collected at an institute of Cardiology in Faisalabad-Pakistan, during April–December 2015. The dataset contains 11 features, which report clinical, body, and lifestyle information (Table1).

Table 1 Meanings, measurement units, and intervals of each feature of the dataset

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40, ..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood	mcg/L	[23, ..., 7861]
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14, ..., 80]
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01, ..., 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50, ..., 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114, ..., 148]
Smoking	If the patient smokes	Boolean	0, 1
(target) death event	If the patient died during the follow-up period	Boolean	0, 1

mcg/L: micrograms per liter, mL: microliter, mEq/L: milliequivalents per litre

1. Getting Started

Load the Heart Failure Clinical Records Data Set (heart_failure_dataset.csv) into Python. According to the target variable (death event), answer the following questions:

1. What percentage of patients who ended up passing away had anemia and were smokers? **(5 points)**
2. The t-test tests whether samples from two independent populations provide evidence that the populations have different means. It produces a "p-value", which can be used to decide whether there is evidence of a difference between the two population means. The p-value is the probability that the difference between the sample means is at least as large as what has been observed, under the assumption that the population means are equal. The smaller the p-value, the more surprised we would be by the observed difference in sample means if there really was no difference between the population means. Therefore, the smaller the p-value, the stronger the evidence is that the two populations have different means. Typically a threshold (known as the significance level) is chosen, and a p-value less than the threshold is interpreted as indicating evidence of a difference between the population means. The most common choice of the significance level is 0.05.
   a. It is suggested that the level of Creatinine Phosphokinase (CPK) in the blood is associated with heart attack risk. Calculate the mean of CPK for both classes (dead vs. survived) in the death event. Generate a box plot (you can use the boxplot option provided in pandas) to show the CPK difference between these groups visually. **(5 points)**
   b. Using the "***scipy.stats.ttest_ind***" function, test the difference in CPK between dead and survived patients. Is the difference significant (p-value less than 0.05)? **(5 points)**

2. Logistic Regression

Split the dataset into training (80%) and test set (20%). Consider random effects in splitting the dataset.

1. Train a logistic regression classifier (with default parameters) on the train set. (the goal is to predict the patients survival) **(5 points)**
2. Evaluate the model performance on both the train set and test set by providing the following metrics. **(10 points)**
   a. Four classification metrics: accuracy, sensitivity, specificity, and f1 score.
   b. A visual normalized confusion matrix.
3. According to the model performance and confusion matrix, do you think it is a reliable model to be used in a real-world application? Provide your reasons. **(5 points)**

3. K-Nearest Neighbors

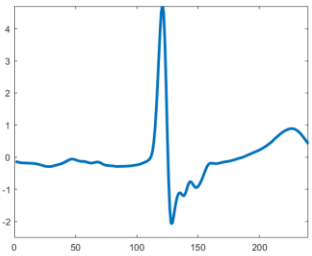
Using the same train set and test set from the previous part, answer the following questions.

1. Train ten different KNN classifiers with K=[1, 2, 3, …, 9, 10]. **(10 points)**
2. Evaluate the performance of each model on both training set and test set by providing the confusion matrix and metrics of classification (accuracy, sensitivity, specificity, and f1 score). **(10 points)**
3. Compare the accuracy of your trained models by plotting accuracy versus K (horizontal axis: K, vertical axis: accuracy). Which K resulted in the highest accuracy? Discuss the results. **(5 points)**

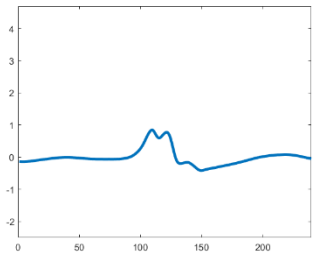
Multi-Class Classification

MIT-BIH Arrhythmia Database

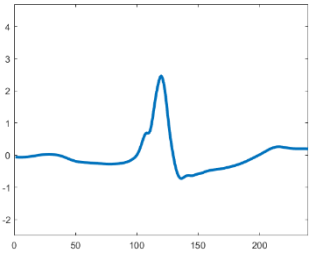
The MIT-BIH Arrhythmia database contains Electrocardiography (ECG) recordings obtained from 47 subjects who were either healthy or suffered from heart arrhythmia (heart rhythm problem). It was studied and annotated by cardiologists by the BIH Arrhythmia Laboratory between 1975 and 1979. Original data is continuous and can be explored using a web application (https://www.physionet.org/lightwave/?db=mitdb/1.0.0). However, five different beat types are extracted for this assignment and they are provided separately in the mat format. There are 2000 beats of each kind (NORMAL, APC, PVC, LBBB, and RBBB). In addistion, a ground truth is provided for each beat. An example of each rhythm is depicted below.



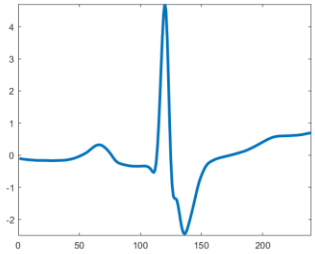
Atrial Premature Beat (APC)



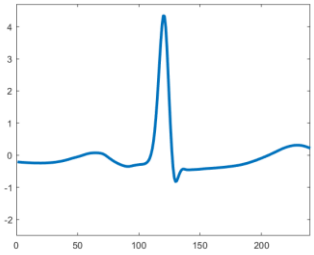
Premature Ventricular Contraction Beat (PVC)



Left Bundle Branch Block Beat (LBBB)



Right Bundle Branch Block Beat (RBBB)



Normal ECG Beat

4. Feature Extraction

1. Load ECG beats (e.g., NORMAL_BEATS.mat) and their ground truths (e.g., NORMAL.mat) into Python.
2. It is reported that each ECG heartbeat can be observed as a series of deflections away from the baseline. Therefore, the time-domain graph of an ECG beat associated with a specific condition (e.g., an arrhythmia) usually generates similar morphological features (appearance). Dynamic time warping (DTW) is a robust algorithm for measuring the similarity between two temporal sequences. Using a DTW implementation, ***"tslearn.metrics.dtw"*** compute each beat's similarity with the provided ground truths. Save the similarity measures for each beat type as the classification features. **(10 points)**

Hints:

1. Install ***tslearn*** package using: ***conda install -c conda-forge tslearn***
2. Load mat files using ***scipy.io.loadmat***
   example: ***normal_beats = loadmat('NORMAL_BEATS.mat')['NORMAL_BEATS']***
3. Compute the DTW similarity measure for each beat and all ground truths. (you should expect to have five features, each containing 10,000 values)
4. Create a label array for your features.
5. The final dataset should have 6 columns (F1, F2, F3, F4, F5, Class) and 10000 rows.

5. Multi-Class Evaluation

Split the dataset into training (80%) and test set (20%). (add some random effects in splitting the dataset)

1. Using the extracted features (F1 to F5), train four classifiers to predict the beat type using the following methods (use default settings for all models): **(10 points)**
    a. K-Nearest Neighbor
    b. Decision Tree
    c. Gaussian Naive Bayes
    d. Support Vector Machine
2. Provide the classification report (***sklearn.metrics.classification_report***) and normalized confusion matrix for each case in part a. **(5 points)**
3. Suppose you are going to select one of these models in a real-world application, which one you choose. Provide a detailed explanation for your choice. **(5 points)**

6. Raspberry Pi

In question 3, find the KNN model with the highest accuracy and implement your trained model on the Raspberry Pi (RPi). You should record a short video showing you can successfully establish a connection between RPi and your PC. The trained model should be sent to RPi, and the classification task has to be done on the RPi. The prediction results should be sent back to your PC to check the performance. All these steps should be included and explained in the video. **(10 points)**

Submitting your Assignment

The assignment must be submitted online at Canvas with the following structure:

For each group, you must submit **one zipped folder** that contains:

1. **A report**- The assignment report must be in **PDF format**, called **report.pdf**. This report must contain all the figures and discussions/explanations for the questions.
2. **A video**- The recorded video from Raspberry Pi function in the **MPEG4 format**, called **video.mp4**. This video should not exceed 3 minutes. Make sure to introduce all the members of your group at the beginning. This video must clearly indicate you could successfully connect the Raspberry Pi kits to your system, and you are able to run the code and produce results. You can also discuss the performance of the model in the video.
3. **A zipped file that contains all your codes**- This zipped folder must be called **code.zip** and must have a single directory called code. No sub-directories and leading path names are accepted. For each question, a separate script must be included in this folder. You must name each script based on the question number in the assignment.