MESIIN481222 Advanced Machine Learning
Lab Assignment 2 – SVM or Regression in Healthcare
Graded assignment (Continuous assessment)

You are required to submit your solutions of this assignment on DeVinci Online, with respect to the deadline set by your instructor. Submit your own work. Cheating will not be tolerated and will be penalized.

Requirements: Lab activity 2 & 3

Before you solve this lab assignment, make sure you finished the lab activities 2 and 3.

SVM or Regression in Healthcare

Support vector machine

Support vector machine algorithms are critical algorithms promising to address the health problem with an accurate computational power. SVM works through regression, classification, and outlier detection of data.
Various researchers have reported SVM's ability to detect various health conditions such as cancer, blood pressure, and diabetes through medical data analytics.
SVM is likely to make a significant revolution with increased use in global health problems.

Noting that when you deal with big data, you are not recommended to use SVM because it is not efficient in handling large sets of datasets. In large datasets, the training time is very high, and the classification become difficult due to slower training. The target classes in a large dataset usually overlap, affecting the classification and predictability. As a result, the use of SVM in extensive medical data may be misleading and affect the results generated.

Regression

Regression analysis is an important statistical method that is commonly used to determine the relationship between several factors.

It can determine the relationship between several factors and disease outcomes or to identify relevant prognostic factors for diseases.

It can be used to investigate factors associated with, or treatments for disease and conditions to improve patient care and clinical practice.

Linear regression

Linear regression is used to quantify a linear relationship or association between a continuous response/outcome variable (that is also called dependent variable) with at least one independent or explanatory variable by fitting a linear equation to observed data.

When investigating the effect or association of a single independent variable on a continuous dependent variable, this type of analysis is called a simple linear regression. In many circumstances though, a single independent variable may not be enough to adequately explain the dependent variable. Often it is necessary to control for confounders and in these situations, one can perform a multivariable linear regression to study the effect or association with multiple independent variables on the dependent variable.

Logistic Regression

As with linear regression, logistic regression is used to estimate the association between one or more independent variables with a dependent variable.

Logistic regression is used to only solve classification problems.

Linear regression is used for predicting the continuous dependent variable using a given set of independent features whereas Logistic Regression is used to predict the categorical.

References:

Bzovsky, S., Phillips, M.R., Guymer, R.H. et al. The clinician's guide to interpreting a regression analysis, Eye 36, 1715–1717 (2022), Doi: 10.1038/s41433-022-01949-z; https://www.nature.com/articles/s41433-022-01949-z

Njoki, L., Effectiveness of Support Vector Machine in Analyzing Medical Data, Section.io, (2022); https://www.section.io/engineering-education/effectiveness-of-svm-on-health-asessment/

Assignment : Medical / Healthcare Case Study using SVM or Regression

You are asked to identify a case study of your choice related to medical or healthcare matter.

After your decide which case study is of interest to you, you select the dataset that is the most convenient for you to conduct your study. There are many clinical and healthcare datasets available online with historical data about patients, clinical centers, diseases, treatments, diagnosis, medical insurance, …

Your use case can be for example a study to detect specific disease from patient's symptoms, predict mental health issues from psychiatric and pathological symptoms and signs, or …

Here are two examples of interesting case studies done by data scientists and published online.

- A study that does insurance forecast by using Linear Regression can be found on https://www.kaggle.com/datasets/mirichoi0218/insurance/code. The medical costs dataset that is used contains personal information of insured persons and costs billed with medical health insurance, and can be found on:

https://www.kaggle.com/datasets/mirichoi0218/insurance/download?datasetVersionNumber=1

- Another case study that predict diabetes from medical records using multiple machine learning models including Support Vector Machine can be found on: https://www.kaggle.com/code/paultimothymooney/predict-diabetes-from-medical-records/notebook. The medical records dataset that is used contains measurements relating to pregnancies, glucose, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, age… and can be found on: https://www.kaggle.com/code/paultimothymooney/predict-diabetes-from-medical-records/data

Many medical and clinical datasets can be found online and can be downloaded for free. Here are some references:
https://www.kaggle.com/datasets?search=health+medical
https://data.world/datasets/health
https://paperswithcode.com/datasets?mod=medical

Other references for health datasets can be found in https://odsc.medium.com/15-open-datasets-for-healthcare-830b19980d9

Attention: You are not limited to the references that are given in this document. You may choose the dataset and case study that you want. The examples above are given to inspire you.

The deliverable should include : A Python Notebook with your source code and inline outputs.
NB: Do not forget to add internal comments for interpretation and explanation in each step. This is mandatory and will be graded.

Your work should follow the steps below:

1- Load your data (csv) in a dataframe
2- Explore your data
3- Print some statistical descriptive information about your data
4- Clean you data if needed
5- Transform your categorical variables if needed
6- You may include some visualization for further data exploration
7- Prepare your dataset for training and testing
8- Create your Regression or SVM model
9- Fit your model on the train dataset
10- Make predictions on the test dataset
11- Evaluate the performance of your model and give some insights

Deadline to share with your instructor the subject of the use case that is selected and the dataset that is chosen in on Tuesday 24th of January 2023 in class. It is recommended to start developing your solution so you can get early feedback from your instructor. Note that you will be given sometime during the sessions on Tuesday 24 January to work with your teammates on your assignment.

Deadline to submit your final work on Online Devinci is: Monday 30 January 2023 at 23:59
A quick demonstration of the code will be done in class on Tuesday 31st of January 2023.

Good luck!