Machine Learning

1 Disease Identification/Diagnosis – supervised learning - diagnosis, particularly in cancer research, depression treatment …

2 Personalized Treatment/Behavioral Modification – supervised learning – better disease assessment, allows physician select from limited sets of diagnosis – imagine a phone app that recognizes skin cancer

3 Drug Discovery/Manufacturing – unsupervised learning - identify patterns understand disease processes, design effective treatments

4 Clinical Trial Research – identifying candidates for clinical trials

5 Radiology and Radiotherapy – examine images to make medical determination

6 Smart Electronic records – advanced collection of records and patient data, OCR, making decisions on pulling information together

7 Epidemic Outbreak Prediction – collect data from various sources and predict epidemics in order to deliver vaccines and resource to prevent disasters, pandemics

Data governance is a big issue because medical data is private, so getting data for machine learning is hard

https://www.techemergence.com/machine-learning-in-pharma-medicine/

Machine learning has often been used in medical problems. It is tempting to think that if you have data on a number of patients, then you can use that to diagnose problems those patients are likely to have. We are going to start with the heart database and try to predict heart disease based on relevant patient data.

This is the Scenario

You work for the Acme Medical Analysis and Prediction Enterprises (AMAPE for short). As engineers for this company you are developing an app to be used by doctors to aid in their prediction of heart disease. Doctors who use this app will have given their patients a battery of tests and logged the results into the system. Each time they find that a patient they tested has or develops heart disease they log into the system and fill in that known.

You now have a database of patient data and outcomes from participating doctors. You have been tasked with developing AMAPE's official prediction model which will be used by subscribing doctors going forward.

As with any problem you first want to study the data.

Read in the heart database given in the problem. It is in the file heart1.csv provided to you. You may want to use panda read_csv which places the data in a dataframe, similar but not to be confused with a dictionary. This data set contains observations based on measurements made on a number of patients. These are medical measurements and personal information. The last column is whether the patient developed heart disease or not. Based on the measurements, you need to build a predictor to determine whether a patient will get heart disease and give the confidence of that prediction.

See reference on dataframes below.

These are the variables in the dataset:

Name	Num	Description
age	0	age
sex	1	sex
cpt	2	chest pain type (4 values)
rbp	3	resting blood pressure
sc	4	serum cholestoral in mg/dl
fbs	5	fasting blood sugar > 120 mg/dl
rer	6	resting electrocardiographic results (values 0,1,2)
mhr	7	maximum heart rate achieved
eia	8	exercise induced angina
opst	9	oldpeak = ST depression induced by exercise relative to rest
dests	10	the slope of the peak exercise ST segment
nmvcf	11	number of major vessels (0-3) colored by flourosopy
thal	12	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
a1p2	13	absence of heart disease = 1, presence = 2

NOTE: You must use, and keep, ALL of the features. Do not eliminate features. Do not use PCA. You will NOT be able to use plot_decision_regions because there are more than two features. That's ok.

Problem 1: Read the database in from this heart1.csv file and analyze the data.

Your analysis should include a statistical study of each variable: correlation of each variable, dependent or independent, with all the other variables. Determine which variables are most highly correlated with each other and also which are highly correlated with the variable you wish to predict.

Create a cross covariance matrix to show which variables are not independent of each other and which ones are best predictors of heart disease. Create a pair plot.

Based on this analysis you must determine what you think you will be able to do and which variables you think are most likely to play a significant roll in predicting the dependent variable, in this case occurrence of heart disease.

Your management at AMAPE want to be kept constantly updated on your progress. Write one paragraph based on these results indicating what you have learned from this analysis. We are looking for specific observations.

Problem 2: Split your heart1.cvs data into training and test datasets. Use every method specified below. Create a report containing a table where you compare prediction percentages and based on this data choose the best method of predicting heart disease on this database. Your written analysis should be one paragraph. Your management at AMAPE expect results and to have a succinct, compelling description of your work. We are looking for specific observations.

For this problem, you are looking for the best combination of parameters for each algorithm. For example, try many different values of K to find the best value to use for K-Nearest Neighbors. An approach would be to use loops for the various parameters. However, do NOT turn in the code you use to FIND the best parameters. Just turn in the code which uses the best parameters you found for each algorithm. And all of this code should be in a single file.

A note on grading: your grade will not be based on the accuracy you achieve unless that accuracy is significantly below what that algorithm is expected to achieve. The goal is to do well for this assignment, it is not to achieve perfection. And keep in mind that different algorithms can have wildly different accuracies.

All reports should be typed in a readable font, uploaded in pdf format, and well labeled. Anything the grader (a representative of AMAPE management) cannot find will be deemed due to omission, or poor organization and not included in the grade.

As in most cases AMAPE management is looking to you for answers, so you will not be able to review your report with them in advance. Note: this means we will not pre-grade this report or any other assignment, turn in what you think is right. (of course you can ask questions, just not: this what I plan to turn in, is it right?)

Machine learning methods to use:

Perceptron
Logistic Regression
Support Vector Machine (pick one version)
Decision Tree Learning
Random Forest
K-Nearest Neighbor (find the best value of K)

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html