15.076 Homework 4

Due: April 4, 11:59 pm

Note: we strongly recommend using Python and the corresponding libraries (like scikit-learn, pandas etc.) for both problems. If you want to use a different language, you need to obtain approval from the TAs.

Problem 1: Understanding Heart Failure

Data from many sources have suggested that heart failure is an increasingly challenging task affecting public health in the United States. It is estimated that nearly 6.5 million Americans over the age of 20 have heart failure, with more than 960,000 new heart failure cases annually. It is also a leading cause of subsequent conditions, contributing to about 36% of cardiovascular deaths. In this problem, we will take a preliminary attempt at understanding why a patient develops heart failure.

The data set we use comes from Kaggle, one of the largest data-science and machine learning community that hosts thousands of freely available data to users. It is a great place for you to explore problems that interest and challenge you. Download the heart.csv data set from Canvas, which comes from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data/. This file contains demographic as well as clinical information of 303 anonymous patients. For each patient, we have the following 12 variables:

- Age
- Anaemia: Decrease of red blood cells or hemoglobin
- Creatinine Phosphokinase: Level of the CPK enzyme in the blood (mcg/L)
- Diabetes: If the patient has diabetes
- Ejection Fraction: Percentage of blood leaving the heart at each contraction
- High Blood Pressure: If the patient has hypertension
- Platelets: Platelets in the blood (kiloplatelets/mL)
- Serum Creatinine: Level of serum creatinine in the blood (mg/dL)
- Serum Sodium: Level of serum sodium in the blood (mEq/L)
- Sex: Woman or man

- Smoking: Smoking or not

a

Developing meaningful models relies on an in-depth understanding of the data. In most data sets, we have two types of variables: quantitative and categorical. Quantitative variables refers to continuous numerical values that represent some measurements (such as age), and categorical variables represent discrete groups (such as sex). Explore the variables of patients, output the min, mean, median and max of quantitative values, and output the counts of each group of the categorical variables (i.e., how many people have sex = 0, and how many have sex = 1). Ignore the “times” variable.

b

Make a histogram of age, and make a bar plot of sex. Do you think the data is representative of each group? Explain your reasoning.

c

Randomly split your data into 70% in train and 30% in test using random seed 1234. Build a logistic regression model to predict whether or not a patient died from heart failure. Use all potential variables except “times”. What is the coefficient for ”smoking” in your logistic regression model, and how would you interpret the relationship between it and whether a patient dies because of heart failure?

d

Plot the ROC curve, and report the AUC of your model in your test set. Explore some current literature on what is the state-of-the-art methods, how do we perform in comparison, and how do you think we can improve the model?

Problem 2: Climate Change

Everyone loves adorable videos of polar bear cub cuddling with their mother or colorful underwater footage of coral reefs. However, due to the constant detrimental effects of climate change, this world of life and beauty is dying at a rapid rate. In this problem, we will look at the factors that may have contributed to the rise of climate change.

Download the climatechange.csv from Canvas, which contains the following features:

- Year: the observation year.
- Month: the observation month.
- Temp: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.

- CO2, N2O, CH4, CFC.11, CFC.12: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division.
- Aerosols: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.
- TSI: the total solar irradiance (TSI) in W/m2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.
- MEI: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

a

Split the data into a training set, consisting of all the observations up to and including 2006, and a testing set consisting of the remaining years. Next, build a linear regression model to predict the dependent variable Temp, using MEI, CO2, CH4, N2O, CFC.11, CFC.12, TSI, and Aerosols as independent variables (Year and Month should NOT be used in the model). What is the R^2 value of this model?

b

Consider a variable significant only if the p-value is below 0.05, which variables are significant in the model? Have you noticed anything counter-intuitive about the coefficients of nitrous oxide and CFC-11? If so, what do you think caused this behavior?

c

Compute the correlations between all the variables in the training set. Which of the independent variables is N2O highly correlated with (absolute correlation greater than 0.7)?

d

What can you do to avoid the problem above and improve the results? Implement a new model with your approach and report the new R^2 value.