Mario Moreno

Machine Learning and Cancer

PA 4

1. How did you prepare the data for modelling?

In order to prepare the data for modelling, I read in the county income and cancer incidence data into a pandas dataframe, merged them together on FIPS, dropped all the columns except our label column (cancer incidence), the county column, and the income column. I then turned all variables such as ‘\*’ and ‘data not available’ into NaN. I created histograms of the income feature and noticed it was left-skewed when I was expecting a normal distribution, so I took the logarithm and returned a normally distributed histogram which I then used to create features.

Dollars histogram prior to log Dollars histogram after log

 

1. What features did you use and how did you compute them?

Using the counties column and the logged dollars column, I then proceeded to build a set of features. I discretized the counties into states, so that each county belonging to Arizona for instance, would return a value of 1 under a column for the state of Arizona. This allowed me to measure the impact that location, which is somewhat loosely correlated with income, would have on cancer incidence. Next, I binned the income column into bins equivalent to the number of standard deviations from the mean. I then discretized the binned values which allowed me to turn a continuous variable into a categorical variable for easier computing.

1. The model you built to model the relationship, if any.

I then ran a train-test split and passed the data through a linear regression model in scikit-learn. I chose a linear regression model after drawing a scatterplot of income to incidence, which hinted at the possibility of a linear relationship between the two. Given that I kept my label variable as a continuous variable, I had to choose an algorithm that works on continuous outcomes such as linear regression.



1. How did you evaluate the model?

Given that this was a model designed to predict a continuous variable, methods such as accuracy are not adequate. Instead, I used mean squared error and r2 as the two metrics of choice. The MSE is big: 152.17, and the r2 is small: 0.47. Both of these suggest that the model we’ve built is a poor predictor of cancer incidence given the feature set.