Data Science with Kaggle Decal

Project 1: Wine Classification

Due at 11:59 PM on March 17th, 2017

Daniel's Big Party

Daniel is planning the frat party of the century. He has tons of wine, but labels for 497 of the bottles fell off. He needs to quickly figure out what the correct labels (1 for Red, 0 for White) are so that people who "only drink red" don't cry out in misery. Luckily, he still has 6,000 labeled bottles as well as the wines' nutrition facts that detail the chemical compositions. Your job is to help Daniel determine what the labels should be for the 497 unknown bottles using their nutritional information.

Basic Logistic Regression

A binary logistic regression model is of the form:

$$P(y=1|x) = h_w(x) = \frac{1}{1+e^{-w^T x}} = \sigma(w^T x)$$
$$P(y=0|x) = 1 - P(y=1|x) = 1 - h_w(x)$$

Question 1

Read in the wine dataset into Python. (Hint: use Pandas)

Question 2

Examine the first few rows of the data. Do a little bit of exploratory data analysis (open ended). You can find and compare summary statistics for different classes, plot scatter plots and histograms, and/or anything else you'd like

Question 3

Separate your labels from your features. That is, extract the last column of the wine dataset and call it "wine.labels"

Question 4

Split the wine data into a training and validation set. Common practice is to select 20% for your validation set, and leave 80% to train on. You should have 4 components here: "wine.training.feats", "wine.training.labs", "wine.valid.feats", and "wine.valid.labs" (feel free to choose your own variable names).

Question 5

It is often very interesting to see how the number of training samples used to train affects the performance of the model. For this reason, please prepare to train on the following number of sample points: [100, 200, 500, 1000, 2000, 4800] (remember, 4800 should be the size of your training set after Question 4).

For each of the number of training samples above, please fit a logistic regression model that classifies wine as red or white based on its nutritional facts. You should use Sklearn's LogisticRegression. At this point, please do not change any default parameters.

Question 6

For each of the number of training samples, please validate your model on the full validation set. Please report a plot that shows how the validation accuracy changes as a function of the number of training samples.

Regularization

Regularization is a commonly used technique to mitigate the problem of overfitting. There are many different types of regularization, but two common types which we will use here are L_1 and L_2 penalization. Essentially, by doing regularization, we are adding a penalty term to the cost function which we are trying to minimize. To adjust your logistic regression models to include regularization, simply alter the "penalty" parameter. See `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

Question 7

Please redo Questions 5 and 6 twice - once with L_1 regularization and once with L_2. Report the validation error vs. number of training samples plots again, and comment on the results. Which regularization method performed better? (For experts: why?)

(Optional) Question 8

When doing regularization, there is a hyperparameter "C" called the inverse of regularization strength. If you have time, please fiddle around with it to determine what values of C give us the smallest error. Sklearn's "KFold" and "CVGridSearch" would be very useful here.

Kaggle

Question 9

Choose your best model and predict the labels of the 497 mystery wines. Output your predictions in a .csv file with 2 columns. The first column should be "IDs" numbered 1-497, and the second column should be labels 0 or 1. Submit your predictions to Kaggle. Good luck!

Deliverables

Please submit:
(1) A Jupyter notebook with your answers to Question 1-9
(2) Your predictions to Kaggle