CME 250 Homework 2

Instructor: Sherrie Wang (sherwang@stanford.edu)

Available: Wednesday, January 23
Due: Friday, February 1, 5:00pm

Please submit Part 1 of the homework via Google Forms, and submit Parts 2 and 3 below via Gradescope.

For each question, clearly identify your final answer and show any work or code used to arrive at the answer. Answers without the corresponding code or justification will not receive credit. You may use any programming language of your choice.

Part 2. Applied Exercise

There is no Part 2 this week. Yay! Let's focus on applying models to real data in Part 3.

Part 3. Classification on Real Data

Question 1. Young People Survey

Last week, I said we would try to predict the `Highest education achieved` variable using the other features. Well, I tried it and it doesn't work. Instead, we're going to try to predict something that the dataset has more signal for — respondent `Gender`.

To make sure we all start from the same encoding of categorical variables, download `hw2.csv` from the course website, and use this version of the Young People Survey for the remainder of this problem. The file `hw2_categorical_encodings.json` contains a dictionary that maps from numerical values to classes for the categorical variables, so you can see what the values correspond to.

(a) Data exploration [3 points]

Before applying any machine learning models to data, it's always a good idea to understand what is in the dataset. Often this is done through *data visualization.*

For your own understanding of this dataset, plot histograms (for numerical features) or bar graphs (for categorical features) for each column of the dataframe *by gender.* This will give you an idea of how feature distributions are similar and different for `female` respondents versus `male` respondents.

In your Homework 2 writeup, please show (1) a histogram for the `Reading` feature, (2) a bar graph for the `Internet usage` feature, and (3) a bar graph for the `Left - right handed` feature. In each, plot the `female` bars in one color and the `male` ones in another color. (You may have to make the bars transparent to allow you to see both.)

(b) Dataset split [3 points]

Using a random seed of `0` for reproducibility, split 20% of the data into the test set, and 16% (20% of the remaining 80% of non-test data) into the validation set. The remaining 64% will be the training set. This allows us to tune hyperparameters and estimate generalization error. Report the dimensions of the three splits.

(c) Logistic regression with ridge penalty [4 points]

Let's try some classification! Notice that we have 674 samples and 149 input features. This puts us in a high-dimensional setting, and warrants *regularization.*

Recall that least squares can be penalized with the coefficients' L_2-norm to become ridge regression:

$$\min_{\vec{\beta}} ||\mathbf{Y} - \mathbf{X}\vec{\beta}||_2^2 + \lambda||\vec{\beta}||_2^2$$

Logistic regression can also be penalized with an L_2-norm to become logistic ridge regression.[1] This creates a model with less variance and potentially better performance. The exact performance of the penalized model depends on the value of λ.

Using available packages/libraries, fit a logistic regression model with ridge penalty to predict `Gender` using the other variables, trying $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. Graph the training set accuracy and validation set accuracy against λ on one plot. Which value of λ results in the highest validation accuracy?

How many coefficients are exactly zero under this model?

(d) Logistic regression with lasso penalty [4 points]

Similarly, least squares can be penalized with the coefficients' L_1-norm to become the lasso:

$$\min_{\vec{\beta}} ||\mathbf{Y} - \mathbf{X}\vec{\beta}||_2^2 + \lambda||\vec{\beta}||_1$$

Logistic regression can also be penalized with an L_1-norm to become logistic lasso regression.[2] This creates a model with less variance *and sparse coefficients.* The exact performance of the penalized model depends again on the value of λ.

Using available packages/libraries, fit a logistic regression model with lasso penalty for each of $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$. Graph the training set accuracy and validation set accuracy against λ on one plot. Which value of λ results in the highest validation accuracy?

How many coefficients are exactly zero under this model?

(e) Generalization error [1 point]

Using the best model you found so far, train the model on your combined training and validation sets and apply it to your test set to obtain an estimate of performance on unseen data. What is the test set accuracy?

[1] If you're curious, this is done by adding $\lambda||\vec{\beta}||_2^2$ to the negative log-likelihood of observing the data. Logistic regression coefficients are found by minimizing this penalized negative log-likelihood.

[2] As you can guess, this is done by adding $\lambda||\vec{\beta}||_1$ to the negative log-likelihood of observing the data.