CME 250 Homework 3

Instructor: Sherrie Wang (sherwang@stanford.edu)

Available: Wednesday, February 6
Due: Friday, February 15, 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 again, so we can focus on analyzing the dataset in Part 3.

Part 3. Classification on Real Data

Question 1. Young People Survey

Recall that we are interested in predicting respondent `Gender` from the other features.

To make sure we all start from the same encoding of categorical variables, download `hw3.csv` from the course website, and use this version of the Young People Survey for the remainder of this problem. The corresponding file `hw3_categorical_encodings.json` contains a dictionary that maps from numerical values to classes for the categorical variables.

The difference between `hw2.csv` and `hw3.csv` is that many of the survey samples with missing values have been retained in `hw3.csv` but discarded in `hw2.csv`. The dataset in `hw3.csv` contains 976 samples, compared to 674 in `hw2.csv`.

(a) Mean Imputation [3 points]

Before we impute missing data, we should test for whether data is missing completely at random (MCAR). However, statistical tests are outside the scope of this course, so we will just proceed with imputation to get some practice.

Perform mean imputation on the missing values in the dataset. That is, for every missing value, fill in the mean value across all samples for that corresponding feature.

Hint: As always, instead of implementing this yourself, search for existing implementations in your language of choice.

(b) Effect of Mean Imputation on Variance [2 points]

What is the sample standard deviation of the `Height` feature before mean imputation? What about after imputation? Why did this happen?

(c) Non-Nested Cross-validation, SVM [8 points]

Using a random seed of `0` for reproducibility, split 20% of the data into a test set, leaving the rest for training and validation.

We are going to fit support vector machines (SVMs) with radial basis function (RBF) kernels to predict `Gender`. As we saw in lecture, SVMs can have a number of hyperparameters in need of tuning. To keep things simple, we are going to keep C, the cumulative budget for points violating the margin, at the default value in whichever SVM function you are using, and focus on tuning γ, a hyperparameter in the RBF.

Recall that the RBF kernel, also called a Gaussian kernel, takes the form

$$K(x, x') = \exp\left(-\gamma||x - x'||^2\right)$$

Using `GridSearchCV` in Python or analogous functions in R and Matlab, perform 5-fold cross-validation to find the best value of γ in the set $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. Here "best" is defined as highest accuracy on validation set `Gender` prediction.

Graph validation set accuracy (mean and standard errors across the 5 folds) against γ and report the best value of γ found.

(d) Generalization Error [2 points]

Using an SVM with RBF kernel and optimal value of γ found in part (c), report the prediction accuracy on the test set.