ELEG/FSAN 817
Large Scale Machine Learning
Homework 2

Due: 11:59PM, Monday, November 14, 2022

(You can work on the software/coding in pairs, but verbal answers should be your own.)

I. Would you eat that mushroom?

The point of this assignment is to investigate three types of classifiers (linear classifier with feature selection, linear classifierwithout feature selection, and a non-linear classifier) in a setting where precision and interpretability may matter.

The data set is adapted and modified but originally based on the UCI Machine Learning Repository[1]. In the modified version, the original categorical data has been converted to one-hot encoding for each categorical feature. The result is 112 binary features. There is a training set consisting of 8,124 instances and test set consisting of 200 novel instances. Each instance is labeled as ± 1, where negative instances are poisonous and positive instances are edible. The sparse feature matrix is stored as (row=instance, column=feature, value=1) tuples, one per line and comma separated in the files `train_X.csv` and `test_X.csv`. For example, if a line has $3, 4, 1$ then instance 3 has feature 4. Obviously, this format avoids the need for storing zeros. Labels are stored in `train_Y.csv` and `test_Y.csv` with one label per line. For interpretation you are provided the 112 feature names in the file `feature_names.csv`.

1) Eating a poisonous mushroom may cause more adverse effects than not eating an edible one. In this context, you are free to weight the classes differently, for example, negatives can be assigned higher weight. Note that you may select the performance metric for validation (and adjust the loss by weighting instances in the training set) by assigning different costs to false positives and false negatives or by assigning different weights to instances based on their classes. Even if you do not use this for the task, please describe why this could be important and what weighting you'd choose. [5 points]
2) The first classification model should be linear. Additionally, use Lasso (ℓ_1 regularization) to perform feature selection. (For a bonus you can implement BoLasso.) Theses leaves different choices of loss/optimization: State whether you will use a ℓ_1-regularized support vector machine (primal or dual), ℓ_1-regularized logistic regression, or if you will treat the labels as targets and use ℓ_1-regularized least squares. Justify your choice and explain the difference between these choices. [10 points + $\leq$ 10 point bonus for using BoLasso instead of just Lasso]
3) You may note that the data covariance is singular, which may induce problems depending on the choice of the the linear classifier. You have options based on regularization such as additional ℓ_2 regularization or elastic-net, adding small amount of noise, or principal component analysis. Which do you choose and why? [5 points]
4) After identifying the support using a shrinkage, you should optimize the linear classifier on the subset of features without regularization. Please report the size of the support for relevant features. List the selected feature (names or indices) along with their coefficients. [10 points]
5) The second classifier should be linear but does not need to use any feature selection, but you may still choose to use ℓ_2 regularization. Rather than report the support provide a scatter plot where the x-coordinate of each point is the coefficient for a feature using the first classifier (noting that some

[1] https://archive.ics.uci.edu/ml/datasets/Mushroom

of the coefficients for the first classifier will be exactly zero if there was feature selection) and the y-coordinate of the point is the coefficient value for the second classifier. [10 points]

6) The third classifier you are free to choose, but it should be non-linear (k-nearest neighbor, kernel support vector machine, random forest with or without gradient boosting, etc.). State the method (model and loss), any regularization, and any other design choices. Describe how you would explain the prediction mechanism of this non-linear model. [10 points]
7) For all three classifiers select the hyper-parameters using a valid experimental design—**You should not use any information from the test set**. For instance, you could perform 5-fold cross-validation on the training set to select the regularization parameter, or you could use a fixed validation set. In either case you can use a weighted validation metric of your choice based on your answer to (1) above [10 points].
8) To evalute your three final classifiers (linear with feature selection, linear without feature selection, and non-linear) report the correspondence of each predictions on the test set using a confusion matrix (results will be three 2-by-2 matrices). [10 points]
9) Write a short discussion of the results, your insights, and the limitations (5–10 sentences). [10 points]
10) Provide your code in a readable format. [20 points]