The University of Texas at Austin

Business Data Science - Fall 2022

HOMEWORK 4

Dimakis Due: Monday, October 3rd, Midnight.

Problem 0. As we have covered in class, we are training a logistic regression model to predict if someone will click on an advertisement. Consider the logistic regression model with 3 features and weights $w = [1, -30, 3]$. For the dataset with features x_1=[20,0,0], y_1=1 and x_2=[23,1,1], y_2=0

- Compute the probabilities that the logistic regression assigns to these two customers clicking on the advertisement (i.e. $P(y = 1)$)
- Compute the cross entropy loss of this logistic regression.
- Design a decision stump (a decision tree of depth 1) that splits on the first feature. What is the Gini impurity of the root? What is the Gini impurity after the best split that you find?

Problem 1: Logistic Regression and CIFAR-10. In this problem you will explore the dataset CIFAR-10, and you will use multinomial (multi-label) Logistic Regression to try to classify it. You will also explore visualizing the solution.

- (Optional) You can read about the CIFAR-10 and CIFAR-100 datasets here: `https://www.cs.toronto.edu/~kriz/cifar.html`.
- (Optional) OpenML curates a number of data sets. You will use a subset of CIFAR-10 provided by them. Read here for a description: `https://www.openml.org/d/40926`.
- Use the `fetch_openml` command from `sklearn.datasets` to import the CIFAR-10-Small data set.
- Figure out how to display some of the images in this data set, and display a couple. While not high resolution, these should be recognizable if you are doing it correctly.
- There are 20,000 data points. Do a train-test split on 3/4 - 1/4.
- You will run multi-class logistic regression on these using the cross entropy loss. You have to specify this specifically (`multi_class='multinomial'`). Use cross validation to see how good your accuracy can be. In this case, cross validate to find as good regularization coefficients as you can, for ℓ_1 and ℓ_2 regularization (called penalties), which are naturally supported in `sklearn.linear_model.LogisticRegression`. I recommend you use the solver `saga`.
- Report your training and test loss from above,
- How sparse can you make your solutions without deteriorating your testing error too much? Here, we ask for a sparse solution that has test accuracy that is close to the best solution you found.

Problem 2: Multi-class Logistic Regression – Visualizing the Solution. You will repeat the previous problem but for the MNIST dataset which you will find here: `https://www.openml.org/d/554`. MNIST is a dataset of handwritten digits, and is considered one of the easiest image recognition problems in computer vision. We will see here how well logistic regression does, as you did above on the CIFAR-10 subset. In addition, we will see that we can visualize the solution, and that in connection to this, sparsity can be useful.

- Use the `fetch_openml` command from `sklearn.datasets` to import the MNIST data set,
- Choose a reasonable train-test split, and again run multi-class logistic regression on these using the cross entropy loss, as you did above. Try to optimize the hyperparameters.
- Report your training and test loss from above,
- Choose an ℓ_1 regularizer (penalty), and see if you can get a sparse solution with almost as good accuracy.
- Note that in Logistic Regression, the coefficients returned (i.e., the $\boldsymbol{\beta}$'s) are the same dimension as the data. Therefore we can pretend that the coefficients of the solution are an image of the same dimension, and plot it. Do this for the 10 sets of coefficients that correspond to the 10 classes. You should observe that, at least for the sparse solutions, these "kind of" look like the digits they are classifying.

Problem 3: Revisiting Logistic Regression and MNIST.
Here we throw the kitchen sink of classical ML (i.e. pre-deep learning) on MNIST.

- Use Random Forests to try to get the best possible test accuracy on MNIST. Use Cross Validation to find the best settings. How well can you do? You should use the accuracy metric to compare to logistic regression. What are the hyperparameters of your best model?
- Use Gradient Boosting to do the same. Try your best to tune your hyper parameters. What are the hyperparameters of your best model?

Problem 4: Revisiting Logistic Regression and CIFAR-10.
As before, we'll throw the kitchen sink of classical ML (i.e. pre-deep learning) on CIFAR-10. Keep in mind that CIFAR-10 is a few times larger.

- What is the best accuracy you can get on the test data, by tuning Random Forests? What are the hyperparameters of your best model?
- What is the best accuracy you can get on the test data, by tuning any model including Gradient boosting? What are the hyperparameters of your best model?