Programming Assignment *5*:

Predict the predominant kind of tree coverage in a forest

Here, you are asked to predict the forest cover type (the predominant kind of tree cover) from strictly cartographic variables (i.e., geographic mapping data etc.). The actual forest cover type for a given 30 x 30 meter cell was determined from US Forest Service (USFS) Region 2 Resource Information System data. Independent variables were then derived from data obtained from the US Geological Survey and USFS. The data is in raw form (not scaled) and contains binary columns of data for qualitative independent variables such as wilderness areas and soil type.

Data Description

The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more of a result of ecological processes rather than forest management practices. Each observation is a 30m x 30m patch. You are asked to predict an integer classification for the forest coverage type. The seven types are:

1 - Spruce/Fir

2 - Lodgepole Pine

3 - Ponderosa Pine

4 - Cottonwood/Willow

5 - Aspen

6 - Douglas-fir

7 - Krummholz

The dataset, “dataset.csv” with 15120 observations contains both features and the Cover_Type.

Description of the Data Fields

Elevation - Elevation in meters

Aspect - Aspect in degrees azimuth

Slope - Slope in degrees

Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features

Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features

Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway

Hillshade_9am (0 to 255 index) - Hillshade index at 9am, summer solstice

Hillshade_Noon (0 to 255 index) - Hillshade index at noon, summer solstice

Hillshade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice

Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition points

Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation

Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation

Cover_Type (7 types, integers 1 to 7) - Forest Cover Type designation

The wilderness areas are:

1 - Rawah Wilderness Area

2 - Neota Wilderness Area

3 - Comanche Peak Wilderness Area

4 - Cache la Poudre Wilderness Area

The soil types are:

1 Cathedral family - Rock outcrop complex, extremely stony.

2 Vanet - Ratake families complex, very stony.

3 Haploborolis - Rock outcrop complex, rubbly.

4 Ratake family - Rock outcrop complex, rubbly.

5 Vanet family - Rock outcrop complex complex, rubbly.

6 Vanet - Wetmore families - Rock outcrop complex, stony.

7 Gothic family.

8 Supervisor - Limber families complex.

9 Troutville family, very stony.

10 Bullwark - Catamount families - Rock outcrop complex, rubbly.

11 Bullwark - Catamount families - Rock land complex, rubbly.

12 Legault family - Rock land complex, stony.

13 Catamount family - Rock land - Bullwark family complex, rubbly.

14 Pachic Argiborolis - Aquolis complex.

15 unspecified in the USFS Soil and ELU Survey.

16 Cryaquolis - Cryoborolis complex.

17 Gateview family - Cryaquolis complex.

18 Rogert family, very stony.

19 Typic Cryaquolis - Borohemists complex.

20 Typic Cryaquepts - Typic Cryaquolls complex.

21 Typic Cryaquolls - Leighcan family, till substratum complex.

22 Leighcan family, till substratum, extremely bouldery.

23 Leighcan family, till substratum - Typic Cryaquolls complex.

24 Leighcan family, extremely stony.

25 Leighcan family, warm, extremely stony.

26 Granile - Catamount families complex, very stony.

27 Leighcan family, warm - Rock outcrop complex, extremely stony.

28 Leighcan family - Rock outcrop complex, extremely stony.

29 Como - Legault families complex, extremely stony.

30 Como family - Rock land - Legault family complex, extremely stony.

31 Leighcan - Catamount families complex, extremely stony.

32 Catamount family - Rock outcrop - Leighcan family complex, extremely stony.

33 Leighcan - Catamount families - Rock outcrop complex, extremely stony.

34 Cryorthents - Rock land complex, extremely stony.

35 Cryumbrepts - Rock outcrop - Cryaquepts complex.

36 Bross family - Rock land - Cryumbrepts complex, extremely stony.

37 Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.

38 Leighcan - Moran families - Cryaquolls complex, extremely stony.

39 Moran family - Cryorthents - Leighcan family complex, extremely stony.

40 Moran family - Cryorthents - Rock land complex, extremely stony.

Methods

You need to work on the given dataset.csv for the training purpose. You can either decide on 80%-20% splitting strategy, or k-fold cross validation to build your model.

In this assignment, you will experiment with AdaBoost using any classification algorithm (of your choice) as the base learner (e.g., logistic regression, etc.), except an ensemble learner. You can leverage library/packages for the base learners. But, you need to write from scratch the AdaBoost function:

- **`def training_adaboost(XTrain, yTrain, num_rounds):`** the function runs num_rounds of AdaBoost on the training set "XTrain" and "yTrain" using the base learner you picked above. It returns a data structure to represent the ensemble of base learners (and their weights) computed by AdaBoost.

Write another function which computes predictions for a given test/judge dataset, and parameters computed by the training_adaboost() function.

- **`def testing_adaboost(adaboost_params, XTest, yTest=None):`** Using the adaboost_params learned from the training_adaboost() function, and test data (Xtest feature data, and optional yTest target true class labels), and returns the accuracy.

Tasks

1. Run AdaBoost you developed above with the given dataset, and num_rounds=1, 2, 3, 4, …, 100. For every value of num_rounds, compute the training accuracy, and test accuracy.
2. Run the base learner alone on the given datasets, and record the training and test accuracy.
3. Plot curves (on a single graph preferably) of the training and test accuracy (of AdaBoosting, as well as single base learning classifer) on the y-axis with the number of rounds on the x-axis, and include the graph in the jupyter notebook submission.
4. In your experiments, you may have noticed something interesting about the accuracy (either training or test) when you run AdaBoost for 1 round or 2 rounds or more. Please provide a note why this happens.
5. **(MS Students Only)** The judge set, "judge-no-labels.csv" contains features without the true labels. Leveraging the entire dataset given, build an AdaBoost model to predict the Cover_Type for every row in the judge set (565892 observations), and record your predictions in a file "judge-predictions.csv", and make a Kaggle entry at the following URL and try(!) to dominate the leaderboard by improving the performance of the classifier through boosting. There is a limit of 10 submissions per day until the deadline. Wishes!

   https://www.kaggle.com/c/f21pa5/overview