AU 332 Homework 5 Due DEC 15 11:59pm

> *Adhere to the Code of Academic Integrity.* You may discuss background issues and general strategies with others and seek help from course staff, but the implementations that you submit must be your own. In particular, you may discuss general ideas with others but you may not work out the detailed solutions with others. It is never OK for you to see or hear another student's code and it is never OK to copy code from published/Internet sources. If you feel that you cannot complete the assignment on you own, seek help from the course staff.

When submitting your assignment, follow the instructions summarized in Section 2 of this document.

This homework can be done in a group of two or one.

1 Forest Categories Classification

This exercise will familiarize you with scikit-learn and its use to classify forest categories. Please use scikit-learn(python) to finish the homework. Scikit-learn is a simple and efficient tools for data mining and data analysis. You will use it to classify forest categories.

1.1 Description

In this homework you are asked to predict the forest cover type (the predominant kind of tree cover) from strictly cartographic variables (as opposed to remotely sensed data). The actual forest cover type for a given 30 × 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. This 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 a result of ecological processes rather than forest management practices.

1.2 Dataset

The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m × 30m patch. You are asked to predict an integer classification for the forest cover 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 training set (15120 observations) contains both features and the *Cover_Type*. The test set contains only the features. You must predict the *Cover_Type* for every row in the test set (565892 observations).
The Data Fields are as follows:

- 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.

1.3 Tasks

Select three or four different algorithms in Figure 1 to solve the above problems and compare their strengths and weaknesses.

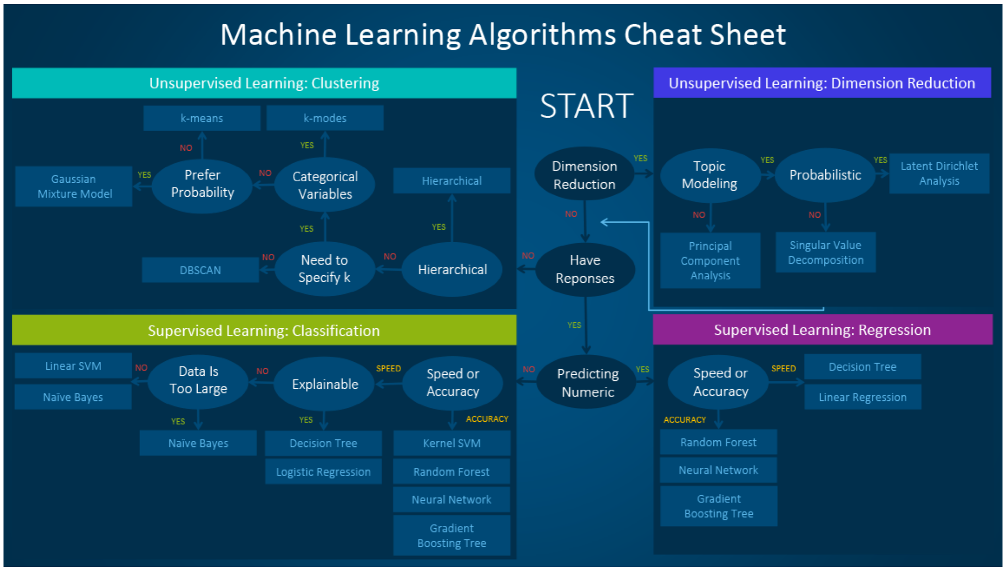
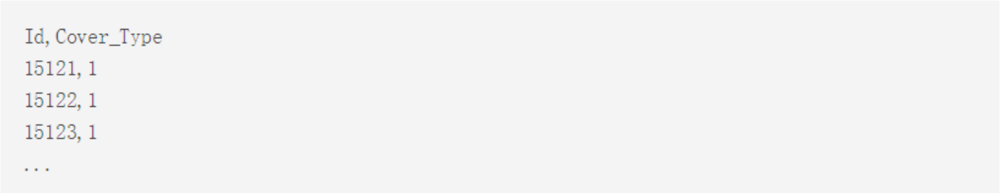


Figure 1: Machine learning cheat sheet

Noticed:

1. Please read the demo file **LogisticRegression.py** and scikit-learn help files before your start coding.

2. Your submission file should have the observation Id and a predicted cover type (an integer between 1 and 7, inclusive). Please check our provided sample submission files called **sampleSubmission.csv**. The file should contain a header and have the following format:



```
Id, Cover_Type
15121, 1
15122, 1
15123, 1
...
```

Figure 2: predict.csv

3. In your implementation, we highly recommend using k-fold Cross-Validation. Calculate the mean squared error with different choices of k.

4. You are required to submit a rar file including all of your source code and your best prediction result for submission in csv format.

5. The evaluation of your homework is based on your report including comparing different algorithms and tuning hyper-parameters.

2 Submission instructions

1. Zip all your python files, `predict.csv`, and `HW5.pdf` to a folder called *homework*5_*name.zip*

2. Send the zip file to TA 121103451@*qq.com*