Introduction to Machine Learning Program assignment #3

Abstract

According to the class, we know what Decision Tree, K- nearest neighbor and naïve Bayes do. This time we use different classifiers/regressors to analyze two data sets.

Problem

- In this assignment you need to **use all the classifiers/regressors in the lecture (i.e. Decision Tree, K- nearest neighbor and naïve Bayes) to analyze two data sets**.
- In naïve Baye**s,** category features need to do **laplace smooth** and continuous features need to calculate the **PDF (probability density function)**.
- You need to submit your code and report. The report should include results, using library, language and explanation of your code. Also you need to tell us your idea about the results. (e.g. You can say why a classifier is better or worse than another)
- **Notice: You can call library this time**

Data set

Download the data sets from the following websites and split the data randomly to training data and test data (70% / 30%) then do your analysis

Iris data set
https://archive.ics.uci.edu/ml/datasets/Iris
Attribute Information:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Forest Fires Data Set

https://archive.ics.uci.edu/ml/datasets/Forest+Fires

Attribute Information:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi

If you want to know what features 5-8 are you can read this website

Grading

- Report(100%)

  For each data set Decision Tree(12.5%), K- nearest neighbor(12.5%) and naïve Bayes(25%)