Project 1: Case Studies
Iris Dataset and Bike Sharing Dataset

Professor Ahmad Namini
Analysis of Big Data 2

Exercise 1 Classification. The Iris Data set was created by R.A. Fisher and is perhaps the best known data set to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. Predicted attribute: class of iris plant.

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

Tasks
Using the prescriptive method of conducting classification problems, create one Python notebook that classifies this dataset utilizing the following models:

- K Nearest Neighbors (KNN)
- Naive Bayes
- Logistic Regression

Comment your notebook and in the end, rank the classification models from best to worst.

Exercise 2 Regression. Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Attribute Information Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

1. instant: record index
2. dteday: date
3. season: season (1:winter, 2:spring, 3:summer, 4:fall)
4. yr: year (0: 2011, 1: 2012)
5. mnth: month (1 to 12)
6. hr: hour (0 to 23)
7. holiday: weather day is holiday or not (extracted from week
8. day: day of the week
9. workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
10. weathersit: - 1: Clear, Few clouds, Partly cloudy, Partly cloudy - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
11. temp: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
12. atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
13. hum: Normalized humidity. The values are divided to 100 (max)
14. windspeed: Normalized wind speed. The values are divided to 67 (max)
15. casual: count of casual users
16. registered: count of registered users
17. cnt: count of total rental bikes including both casual and registered

Tasks

Using the prescriptive method of conducting regression problems, create one Python notebook that predicts the count of casual users (feature casual), count of registered users (feature registered), and the total count of both causal and registered users (feature cnt). Use the multi-regression model and comment your notebook.