Week 6 Exercises: Statistics for Data Science

Part 1: Linear Regression

The file bikes.csv contains a cleaned version of the dataset available at https://www.kaggle.com/marklvl/bike-sharing-dataset. It contains a two-year historical log of bike rentals from the Capital Bikeshare system in Washington, D.C.

Your task in this exercise is to learn how the number of rentals on a given day (rentals) is related to environmental and seasonal factors.

The columns for this dataset are as follows:

- dteday: record's date
- season: the season
- mnth: the month
- holiday: whether the day was a holiday or not
- weekday: day of the week
- temp: temperature in Fahrenheit
- atemp: "Feels like" temperature in Fahrenheit
- hum: humidity
- windspeed: wind speed
- rentals: count of total rental bikes (our target variable)

Read in this data as a dataframe named *bikes*.

1. Create boxplots of rentals vs. season, rentals vs. mnth, rentals vs. holiday, and rentals vs. weekday. Do you notice any trends?

2. Create scatterplots of rentals vs. temp, rentals vs. atemp, rentals vs. hum, and rentals vs. windspeed. Do you notice any trends.

3. Create a linear regression model to predict rentals based on all other features (besides dteday). Note that you do you need to dummyize the categorical predictors (season, mnth, holiday, and weekday) prior to modeling. Also, make sure to do a train/test split prior to fitting your model so that we have a test set on which to evaluate.

4. What coefficient do you get for temp? Explain, in English, the meaning of this coefficient. Does this make sense considering the scatterplot you created earlier?

5. What coefficients do you get for the season variables? Explain, in English, the meaning of these coefficients. Do they make sense, give the boxplots you created earlier?

6. Calculate both the R^2 score and the mean absolute error on the test set. Explain, in English, what these metrics are telling you about your model.

7. What potential problems do you see in regard to interpreting the coefficients you got?

Part 2: Logistic Regression

The dataset breast_cancer.csv comes from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. n the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

The columns are as follows:

- id: ID number of the patient
- diagnosis: Diagnosis, where M = malignant and B = benign
- radius_mean: mean of distances from center to points on the perimeter
- texture_mean: standard deviation of gray-scale values
- perimeter_mean: perimeter
- area_mean: area
- smoothness_mean: local variation in radius lengths
- compactness_mean: perimeter^2 / area - 1.0
- concavity_mean: severity of concave portions of the contour
- concave points_mean: number of concave portions of the contour
- symmetry_mean: symmetry
- fractal_dimension_mean: "coastline approximation" - 1

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

We will consider the following set of features: 'radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean', 'concave points_mean', 'symmetry_mean','fractal_dimension_mean'

Read in this data as a dataframe named *cancer*.

1. Look at box plots for each variable against diagnosis. For each one, describe the relationship that you see (eg. malignant tend to have higher average).

2. Look at the scatterplots of each variable against each other variable. What do you see?

3. Run a logistic regression model with response variable diagnosis and using all of the mean variables as predictors.

4. Describe the performance of your model in terms of calibration and discrimination.

5. Look at the coefficients. What is the sign of the coefficients for radius_mean? This seems unexpected. Can you explain why this happened?