DSC 190: Homework 1

January 6, 2021

Instructions

Please submit your solution **by the beginning of the week 2 lecture (Jan 12 9:30 AM PT)**. Submissions should be made through **Gradescope**. Please complete homework **individually**. Please include the code of your solutions in the submission along a write-up explaining the logic of your code. Feel free to use any third party libraries.

Nutrition Facts for McDonald's Menu data: Please download menu.csv from the following link https://www.kaggle.com/mcdonalds/nutrition-facts

1. **Data Exploration (20 points):**

   (a) Plot the histogram of the Calories. Comment on the datatype of the features.

   (b) Plot the correlation heatmap between features and Calories (Hint: It will be a 21 $\times$ 21 matrix). You may notice that the diagonal elements are always 1. Explain the reason.

   (c) List the features which have the second and third largest positive correlation with Calories. Note: if you encounter multiple features actually mean the same thing, only list the feature with the largest correlation. For instance, if you see Sodium and Sodium (%Daily Value) and the former has larger correlation with Calories, list Sodium only.

   (d) Report all features which have negative correlation with Calories. Does your result meet your expectation?

2. **Plotting (20 points):**

   (a) Plot the scatter plot for 'features vs. Calories' for all features found in 1(c) and 1(d).

   (b) Plot the box plot for all features found in 1(c) and 1(d) correspondingly.

3. **Data Pre-processing: missing values (20 points):**

   (a) Report the median and standard deviation for all numerical features.

   (b) Write the code to replace outliers of all numerical features in (a) with NaN. (Hint: Use the 3 sigma deviation to find outliers). Report the total number of NaNs corresponding to each feature.

   (c) Write the code to replace the missing values (NaN) with mean values. Report the median and standard deviation. Compare your result with (a) and write one sentence to explain your discovery.

4. **Linear Regression (20 points):**

Note: Please use the original dataset for this question.

   (a) Train a predictor to predict the Calories as follows:

$$\text{Calories} = \theta_0 + \theta_1 \times [\text{Carbohydrates}] + \theta_2 \times [\text{Protein}] + \theta_3 \times [\text{Total Fat}]$$

   Report the values of θ_0, θ_1, θ_2 and θ_3. Briefly describe your interpretation of these values, i.e., what do θ_0, θ_1, θ_2 and θ_3 represent? Explain these in terms of the features and labels.

(b) Train another predictor to predict the Calories as follows:

$$\text{Calories} = \theta_0 + \theta_1 \times [\text{Total Fat}]$$

Report the values of θ_0 and θ_1. Note that the coefficient here might be different than the one from (a) though they refer to the same feature. Provide an explanation as to why these coefficients might vary significantly.

(c) Split the data into two fractions – the first 90% for training, and the remaining 10% testing (based on the order they appear in the file). Train the model using all the features available in the training set only. What is the model's MSE on the training and on the test set? Did it perform too well on the training set than the test set? If yes, what could be the reason?

5. **Logistic Regression (20 points):**

In this question, let's decipher the interpretation of weights in Logistic Regression. Consider a binary classification problem where possible labels are 0, 1. If we use logistic regression to perform this classification task where θ_1, θ_2, ... θ_n are the weights and x_1, x_2, ... x_n are the features:

(a) Express odds of predicting label 1 in terms of weights and features.
Hint: odds of an event E = $\frac{\text{probability of the event E}}{1-\text{probability of the event E}}$. In our case, E = predicting label 1.

(b) Increase one feature value(x_i) by 1 and let all other features remain the same. What is the ratio of new odds of predicting label 1 to the old odds of predicting label 1? What can we infer from this expression? Can we deduce anything about the weights?