CSE 6363 - *Machine Learning*

Project 1- Spring 2021

Due Date: Mar. 7 2021

Linear Regression

1. Consider the polynomial fit problem where you want to find the best polynomial approximation of order k to a set of data points.

   a) Implement the polynomial fit solver for 2-dimensional input data as a linear regression learner. Make sure your implementation can handle polynomial fits of different order (at least to 4th order).

   b) Apply your regression learner to the following data set and plot the resulting function for order 1, 2, 3, and 4. Plot the resulting polynomial surface together with the data points (using your favorite plotting program, e.g. Matlab, Octave, ...)

$$
\begin{array}{lllllll}
D = \{ & ((6.4432, 9.6309) & 50.9155), & ((3.7861, 5.4681) & 29.9852), & \\
 & ((8.1158, 5.2114) & 42.9626), & ((5.3283, 2.3159) & 24.7445), & \\
 & ((3.5073, 4.8890) & 27.3704), & ((9.3900, 6.2406) & 51.1350), & \\
 & ((8.7594, 6.7914) & 50.5774), & ((5.5016, 3.9552) & 30.5206), & \\
 & ((6.2248, 3.6744) & 31.7380), & ((5.8704, 9.8798) & 49.6374), & \\
 & ((2.0774, 0.3774) & 10.0634), & ((3.0125, 8.8517) & 38.0517), & \\
 & ((4.7092, 9.1329) & 43.5320), & ((2.3049, 7.9618) & 33.2198), & \\
 & ((8.4431, 0.9871) & 31.1220), & ((1.9476, 2.6187) & 16.2934), & \\
 & ((2.2592, 3.3536) & 19.3899), & ((1.7071, 6.7973) & 28.4807), & \\
 & ((2.2766, 1.3655) & 13.6945), & ((4.3570, 7.2123) & 36.9220), & \\
 & ((3.1110, 1.0676) & 14.9160), & ((9.2338, 6.5376) & 51.2371), & \\
 & ((4.3021, 4.9417) & 29.8112), & ((1.8482, 7.7905) & 32.0336), & \\
 & ((9.0488, 7.1504) & 52.5188), & ((9.7975, 9.0372) & 61.6658), & \\
 & ((4.3887, 8.9092) & 42.2733), & ((1.1112, 3.3416) & 16.5052), & \\
 & ((2.5806, 6.9875) & 31.3369), & ((4.0872, 1.9781) & 19.9475), & \\
 & ((5.9490, 0.3054) & 20.4239), & ((2.6221, 7.4407) & 32.6062), & \\
 & ((6.0284, 5.0002) & 35.1676), & ((7.1122, 4.7992) & 38.2211), & \\
 & ((2.2175, 9.0472) & 36.4109), & ((1.1742, 6.0987) & 25.0108), & \\
 & ((2.9668, 6.1767) & 29.8861), & ((3.1878, 8.5944) & 37.9213), & \\
 & ((4.2417, 8.0549) & 38.8327), & ((5.0786, 5.7672) & 34.4707) & \}
\end{array}
$$

   c) Evaluate your polynomial regression functions by computing the error on the following data points (generated from the original function. Compare the error results and try to determine for what polynomials overfitting might be a problem. Which order polynomial would you consider the best prediction function and why.

$$
\begin{array}{lllllll}
T = \{ & ((0.8552, 1.8292) & 11.5848), & ((2.6248, 2.3993) & 17.6138), & \\
 & ((8.0101, 8.8651) & 54.1331), & ((0.2922, 0.2867) & 5.7326), & \\
 & ((9.2885, 4.8990) & 46.3750), & ((7.3033, 1.6793) & 29.4356), & \\
 & ((4.8861, 9.7868) & 46.4227), & ((5.7853, 7.1269) & 40.7433), & \\
 & ((2.3728, 5.0047) & 24.6220), & ((4.5885, 4.7109) & 29.7602) & \}
\end{array}
$$

Logistic Regression

2. Consider again the problem from the first assignment where we want to predict the gender of a person from a set of input parameters, namely height, weight, and age. Assume the same training data:

$$
\begin{array}{lll}
D = \{ & ((170, 57, 32), & W), \\
 & ((192, 95, 28), & M), \\
 & ((150, 45, 30), & W), \\
 & ((170, 65, 29), & M), \\
 & ((175, 78, 35), & M), \\
 & ((185, 90, 32), & M), \\
 & ((170, 65, 28), & W), \\
 & ((155, 48, 31), & W), \\
 & ((160, 55, 30), & W), \\
 & ((182, 80, 30), & M), \\
 & ((175, 69, 28), & W), \\
 & ((180, 80, 27), & M), \\
 & ((160, 50, 31), & W), \\
 & ((175, 72, 30), & M), \quad \}
\end{array}
$$

a) Implement logistic regression to classify this data (use the individual data elements, i.e. height, weight, and age, as features). Your implementation should take different data sets as input for learning.

b) Use your logistic regression function to predict the classes for the same data items as in the first homework:

$$(155, 40, 35), (170, 70, 32), (175, 70, 35), (180, 90, 20)$$

Given that the correct class labels for these items are W, M, W, M, respectively, compare the results for your logistic regression server with the ones for KNN and Naïve Bayes (from your first assignment). Discuss what differences exist and why Logistic Regression can yield better results for this problem.

Linear Discriminant Analysis

3. Consider again the problem from Question 2. with the same training data and the test data in 2.b).

a) Implement Linear Discriminant Analysis and apply it to the data. Your implementation should take different data sets as input for learning.

b) Show the classification results obtained and compare them to the ones form Question 2. (logistic regression). Also compare the decision boundaries formed.

c) Use the learned model to generate datasets for the two classes (each one with approximately 50 data points) and plot these points as well as the original training points in terms of height and weight. Discuss similarities and differences of the training and generated distributions.