MAE551_451 Fall 2022 – Problem Set 1

Due 9/26/2022 at 11:59 PM

1. **(Sigmoid function)** Compute the derivative of the Sigmoid function; write a Python/MATLAB function to make two plots: one for the Sigmoid function and the other for its derivative.
2. **(Probability theory)** The Dercum disease is an extremely rare disorder of multiple painful tissue growths. In a population in which the ratio of females to males is equal, 5% of females and 0.25% of males have the Dercum disease. A person is chosen at random and that person has the Dercum disease. Calculate the probability that the person is male.
3. **(Gradient descent)** Write a Python/MATLAB program to implement the gradient descent method to minimize the function $f(x) = 2x^2$
4. **(Polynomial curve fitting)** Use Python/MATLAB and follow the steps below to perform polynomial curve fitting:
   a. Generate 100 equal-spaced numbers (denoted as x) between 0 and 1
   b. Plug in the above 100 numbers to the function $\cos(2\pi x)$ to obtain 100 new numbers denoted as y
   c. For each y value add a random noise; the added random noises follow a Gaussian distribution; This step gives another 100 new numbers denoted as z
   d. Plot the above data of x and z in the xz coordinate system
   e. Randomly choose 80 (x,z) data as the training data and the rest as testing data. Fit the training data to a polynomial function: $z(x,w) = w_0+w_1x+w_2x^2+\ldots w_Mx^M$ using M = 1, 2, 3, ….. Plot the fitted data. Applied the trained model to the testing data. Plot the root mean square error for the training and testing data as a function of M. Record the critical M value where overfitting occurs.
   f. Apply different regularization coefficients λ and determine λ values that can be applied to remedy the overfitting problem. Try both the Lasso and Ridge regularization methods.
5. (**Linear regression**) The attached **HW1_P5_weatherHistory.csv** file records an hourly/daily summary of weather for Szeged, Hungary area, between 2006 and 2016. Train a **linear regression** model to uncover the relationship between humidity and temperature. Available machine learning packages such as Scikit-learn is encouraged to be used.
   Consider the following factors in your solution:
   a. Visualization of the data
   b. Effect of learning rate
   c. Different optimization algorithms
   d. 3-fold cross validation
   e. Possible overfitting
   Discuss your findings on the relationship between humidity and temperature.
6. (**Logistic regression**) The attached **HW1_P6_candy-data.csv** file records 85 brands of candy and their two attributes (sugarpercent and pricepercent) and rankings (winpercent). The meanings of the attributes and ranking are:
   sugarpercent: The percentile of sugar it falls under within the data set; pricepercent: The unit price percentile compared to the rest of the set; winpercent: The overall win percentage.
   Train a **logistic regression** model to determine whether a brand of candy is popular depending on the sugarpercent and pricepercent values. Available machine learning packages such as Scikit-learn is encouraged to be used. **Hint:** label the data according to the winpercent values. For example, if a candy has winpercent greater than 0.5, the candy is labeled as popular. Otherwise, not popular. Consider similar factors as in Problem 5 and discuss your findings.