COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

Machine Learning, I

CS 5805

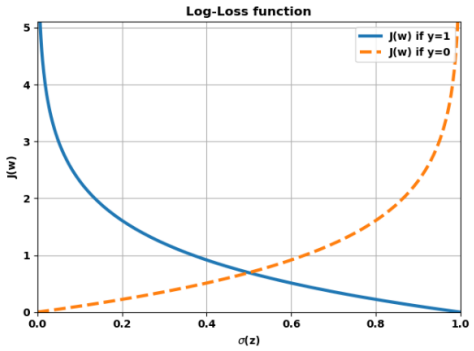
Homework # 6

1. Consider the Riding Mower example with the following information as described in the table below:

Customer#	Income	Lost size	Ownership
1	60	18.4	Owner
2	64.8	21.6	Owner
3	84	17.6	Nonowner
4	59	16	Nonowner
5	108	17.6	Owner
6	75	19.6	Nonowner

   I. Construct a logistic regression model with two predictors for riding mower example with $\beta_0 = -25.9382, \beta_1 = 0.1109, \beta_2 = 0.9638$ where β_1 and β_2 are for the "Income" and "Lot_size" variables, respectively. Write down the model.
   II. Using the logistic regression model with probability cutoff = 0.75, classify the customers in the above table as "owner" or "non-owner."
   III. Develop the confusion matrix for this problem.
   IV. Find the accuracy, precision and recall for this classification problem.

2. Plot the cross-entropy function (logistic regression) versus $\sigma(z)$ where $\sigma(z) = \frac{1}{1+e^{-z}}$. The final figure should be like below figure: (linewidth =3). The cross-entropy function equation can be found on the lecture notes. The input is x that ranges form -inf to +inf. [5pts]



3. Using python and sklearn package generate a synthetic data (make_classifcation) with the following information:
   a. n_samples : 1000
   b. n_features : 2

c. n_cluster_per_class : 2
d. n_informative : 2
e. n_repeated : 0
f. n_redundant : 0
g. random_state:5805

Split the dataset into train-test with 80-20, respectively. Develop a regression model that classifies the target.

I. Plot the confusion matrix.
II. Plot the ROC curve with the AUC as the legend.
III. Display the accuracy, recall and precision for this classification problem.

4. Using python load the 'Smarket.csv' from the course GitHub. The objective is to find a machine learning model [logistic regression] that predicts 'Direction' using 'Lag1' through 'Lag5' and 'Volume'.

   a. Is this dataset balanced or imbalanced? If the dataset is imbalanced, then make it balance using the SMOTE method. Plot the imbalanced and balanced dataset.
   b. Split the dataset into train-test 80-20 then standardize the dataset. Turn the shuffle ON and use the random_state = 5805. Display the first 5 rows of the train and test set.
   c. Find the logistic regression model that predicts the 'Direction' using the default parameters of logistic regression. [only set the random state to 5805].
      i. Print the best score for the train and test.
      ii. Plot the confusion matrix.
      iii. Plot the ROC curve with the AUC as the legend.
      iv. Display the accuracy, recall, precision and f1 score for this classification problem.
   d. Repeat part c. using the grid search with cross validation = 5 to find the best parameter :
      i. penalty{'L1', 'L2', 'elastic net'}
      ii. L1_ratio : search for [0,1] with 30 numbers in between
      iii. C : search between [0.001, 10] with 10 numbers in between
      iv. Display the result of the grid search with CV with the best parameters listed above on the console.
      v. Fit the train dataset to a Logistic Regression model using the fined tuned parameter from the grid search and display part i, ii, iii, iv. listed in part c. Did the grid search improve the performance of this classification?