HW3 Report

Weigeng Li G44406413

Problem 1

a) I would like to choose a y value with minimum MSE. Solving $y = b$ with Least Squares.

$\hat{b} = \frac{\sum y}{N} = \frac{14}{4} = 3.5$. Therefore, I would choose the average of $y = 3.5$ as prediction. The

$MSE = \frac{(1-3.5)^2+(3-3.5)^2+(4-3.5)^2+(6-3.5)^2}{4} = 3.25$

b) By Least Squares, $\hat{\alpha} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{22}{17}$, $MSE = \frac{43}{34}$

c) By Least Squares,

$$\hat{a} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - x)^2} = 1\,, \qquad \hat{b} = \bar{y} - \hat{a}\bar{x} = 1, MSE = 1$$

Note: MSE in problems 2 and 3 are the mean MSE of 10-fold cross-validation

Problem 2

variable	MSE	Intercept	coefficients	Model	Error=sqrt(MSE)
RBI	857556.4	14.18	28.04	28.04*RBI14.18	926.0434099
Runs	906051.5	-34.29	27.47	27.47*Runs-34.29	951.867378
Hits	948824.5	-129.49	14.84	14.84*Hits-129.49	974.0762343
HomeRuns	1009302	531.49	78.81	78.81*HomeRuns+531.49	1004.64009
Doubles	1038189	106.31	68.50	68.5*Doubles+106.31	1018.91576
Walks	1052320	257.31	28.31	28.31*Walks+257.31	1025.826418
Strikeouts	1291287	405.67	14.86	14.86*Strikeouts+405.67	1136.348138
OnBasePct	1382653	-1535.90	8594.62	8594.62*OnBasePct-1535.9	1175.862835
BattingAvg	1436805	-985.71	8665.70	8665.7*BattingAvg-985.71	1198.667954
StolenBases	1456266	1026.69	26.90	26.9*StolenBases+1026.69	1206.758675
Triples	1464413	980.34	114.69	114.69*Triples+980.34	1210.129339

The model with the minimum MSE is 28.04*RBI14.18. Therefore, I would like RBI as my single indicator.

Problem 3

These are the top 9 models with the highest MSE

Variables	MSE	Error=sqrt(MSE)
['BattingAvg', 'OnBasePct', 'Hits', 'HomeRuns', 'Strikeouts', 'StolenBases']	778625.8	882.3977
['BattingAvg', 'Runs', 'HomeRuns', 'RBI', 'Strikeouts']	779771.6	883.0468
['BattingAvg', 'Runs', 'HomeRuns', 'RBI', 'Strikeouts', 'StolenBases']	780850.7	883.6576
['BattingAvg', 'Hits', 'HomeRuns', 'Walks', 'Strikeouts', 'StolenBases']	780869.3	883.6681
['OnBasePct', 'Runs', 'HomeRuns', 'RBI', 'Walks', 'Strikeouts', 'StolenBases']	781602.7	884.0829
['OnBasePct', 'Hits', 'HomeRuns', 'Walks', 'Strikeouts', 'StolenBases']	781808.5	884.1994
['BattingAvg', 'OnBasePct', 'Runs', 'HomeRuns', 'RBI', 'Walks', 'Strikeouts', 'StolenBases']	782043	884.3319
['BattingAvg', 'Hits', 'HomeRuns', 'Strikeouts', 'StolenBases']	782396.8	884.532
['Runs', 'HomeRuns', 'RBI', 'Strikeouts']	782816.7	884.7693

he Best Model has variables 'BattingAvg', 'OnBasePct', 'Hits', 'HomeRuns', 'Strikeouts' and 'StolenBases'. The coefficients of this model are:

variable	coefficients
BattingAvg	-287.779
OnBasePct	189.9063
Hits	682.629
HomeRuns	676.9383
Strikeouts	-497.125
StolenBases	120.3222
Intercept	1248.528

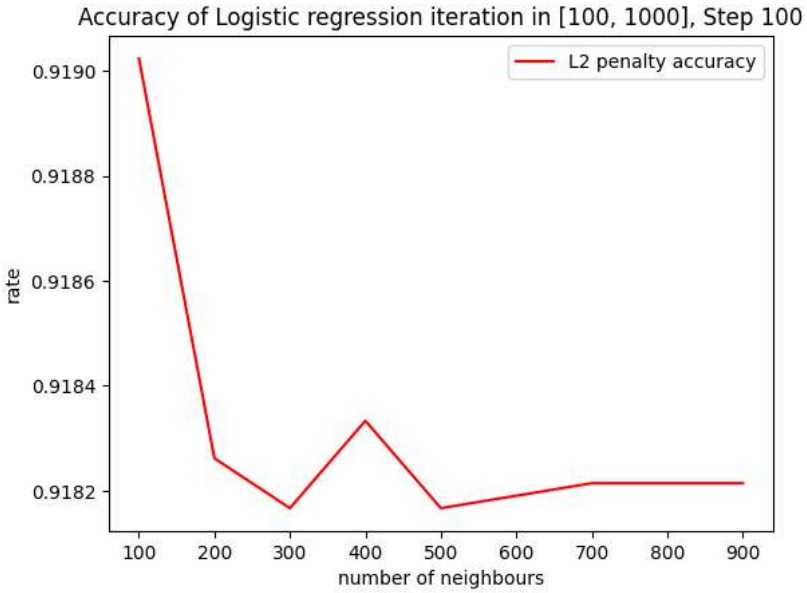
The model is:

$$Salary = -287.779\,\text{BattingAvg} + 189.9063\text{OnBasePct} + 682.629\text{Hits} + 676.9383\text{HomeRuns} - 497.125\text{Strikeouts} + 120.3222\text{StolenBases} + 1248.528$$

The Mean Error of this model is 882.3977.

Problem 4

I evaluate the model with mean MSE of 5-fold cross-validation. The 5-fold cross-validation takes the whole training set. Randomly split into 5 same-size sets, fit model on 4 sets, and evaluate the model in the rest of one set. I take the average of the 5 accuracies as the cross-validation score. The cross-validation score is the green line in the graph.



The Logistic Regression model closure at the 706 iterations. The overall accuracy of the model is among 0.91 to 0.92, which is smaller than the KNN model.

The accuracy of model is dropping as iterations increase. I think it will be better to exam the model with R-square score. However running these bathes takes a lot of time so I can't run it again.

The decision boundary of logistic regression is linear. This means the model can't fit more complicate boundary. In contract the decision boundary of knn is more complicated. That lead to KNN model fit the MNIST dataset better.