Alex Wolf
5/8/17

ML Regression Model Notes

Regression Modeling Steps/Notes

1. Things to check in Data
   a. if high dimensional
      -may want to use Lasso
   b. Check Initial feature correlation
      -use scatterplot
      -Data should be correlated with Regression Label
      -Ridge Regression works well if highly correlated data
   c. Get rid of any features with very low correlation with Label mainly, and other feats

2. Regression Model Testing
   a. Try RidgeReg as works well with good data
      c. Lasso first if very high dimensional Data or poor correlaton between all features
   b. Try Lasso Next
   c. Remove features
      - Plot Lasso Parameter coefficents and get rid of low features
   d. Try Ridge Regression again
   e. Try Elastic Net Regresion
   f. Test model fit versus number of parametes (AIC/BIC)
   g. Repeat if too many paramters from ideal AIC/BIC

Other Notes

1. Alpha
   -high alpha= more regularization ==undefitting
   -Lower alpha= more ovefitting
   - normally from 0 to 1

2. Lasso (L1 abs val_regualizaion)
   - good for achieving sparsity
   -difficult to avoid overfitting
   -good for regression feature selection

3. Ridge (L1- square parameters)
   -Normally better bias/variance tradeoff
   -Good if Normal Prior distribution
   -Good for High dimensional Data
   -Much Better for Highly correlated features
   do correlation plot
   -Avoids overfitting more

4. Elastic Net
   - shrinkage and automatic variable reduction
   -find best combo of L1 and L2 regularization

Regression Functions:

performRidgeReg(X, y, cvfolds=5, impStrategy= 'mean', aLow=0, aHigh=1, numAlphas=30)
- Determines which alpha hyperparamter makes the best RidgeRegression and prints R^2 score
-Uses Hold out validation
-alphas in range alow to aHigh
-can change impuation strategy from mean
-Standardizes Data

performLassoReg(X, y, cvfolds=5, impStrategy= 'mean', aLow=0, aHigh=1, numAlphas=30)
- Determines which alpha hyperparamter makes the best LassoRegression and prints R^2 score
-Uses Hold out validation
-alphas in range alow to aHigh
-can change impuation strategy from mean
-Standardizes Data

performElasticReg(X, y, cvfolds=5, impStrategy= 'mean', numRatios=10, aLow=0, aHigh=1, numAlphas=10)
-Perform ElasticRegression using a combo of L1 and L2 regualarization
-Uses Hold out validation
-alphas in range alow to aHigh
-Number of different ratios to produce from 0-1 in numRatios
-can change impuation strategy from mean
-Standardizes Data

Model Improve/ Visualiztion Functions:

testFitvsNumParms(X, y, impStrategy= 'mean')
- -plots a graph with AIC and BIC showing optimal number of paramters with solid line
- -can change impuation strategy from mean

showLassoParamWeights(X, y, alpha=.4, impStrategy='mean')
- -used to show which weights go to zero from Lasso
- -can change alpha: should test optimal first from performLassoReg()
- -try removing the parameters with small weights from this graph in a Ridge

getNewXfromLassoWeightThresh(X, y, alpha=.4, weightThresh=1, impStrategy='mean')
- -Get new Pandas Df of X from a paramter weight Threshold in Lasso
- - Should look at graph from showLassoParamWeights() to determine the threshold