machine learning

1. weighting samples - logistic regression

2. boosting - suggested WEKA - java front end

tries to use multiple algorithms with weak predictors

MATLAB has built in machine learning/boosting stuff

3. reduce dataset & format data to what it wants

linear svm with logistic regress

Genie factor - kaggle resubmission to see how we did

PCA = reduce dimensions

ablative analysis - take out 1 column and see if it changes the algorithm by much

training data - split into positives & negatives

Gini factor

-check for false positives/false negatives – create 2x2 matrix

-ROC curve

Logistic regression

precision = 0.8678

recall = 0.2979

take out features - plot precision/recall

stragegies:

1. change output threshold
2. look at overlapping data/nonoverlapping data

-rus boost for matlab bood for weighted distribtuitons

-open file - load cvs = number auto import as numerical, strings nominal - automatically creates bins for nominal -data - classify - choose clasiffiers - - most of the boostings in meta - google what the boosts are

What our metric is going to be – evaluate success

Reduce false negatives –

PN = TN +FN – precision – definition of good/bad reversed

Precision/recall depends on the trade off between lost profit in not buying a good car vs. the cost it takes to repair/lose on a bad buy.

Precision – how well can you trust ANY prediction – trust predition a lot

Recall – how many false negatives – how good your base predictions are

-output cutoff - shift decision bound

Some of these include adaboost, logitboost, ensembleselection, and costsensitiveclassifier. - helping - best algorithms - ensemble methods/selection – j48, decision tree, naive Bayes, REptree,

-ensemble methods –

-for each of the methods, it not clear which variable to tweak for desicion boundary

-cost sensitive – albert knows how to do decision boundary

meta algorithms

-separation of data

-precision vs. recall or a different metric - success metric

-output cutoff - shift decision bound

Some of these include adaboost, logitboost, ensembleselection, and costsensitiveclassifier. - helping - best algorithms - ensemble methods/selection

-ensemble method - not sure if we can - not sure if boosted classfiers

-splitting data - variance of boosting - powerful classifier

AUC - right evaluation metric

kernalized svm

boosted classifiers - compare - maybe already overfitting - higher variance classifier

ensemble methods - wouldn't classify regressor and classifier - training on ensemble method for us - compare all algorithsm in term of ROC and AUC for training and testing

what they look for - doing the right things in feedback loop - try something - evaluate it - based on evaluation - make next step

demand 5% false positive rate - evaluate performance at 5% false positive rate - say what is true positive rate - maximize number of true positives as false positive rate

optimize for AUC - good precision and good recall

logistic regression - easy to modify it - balanced data set

meta algorithms

train on balanced - data adjust threshold

compute area under curve - provably optimum - area under curve - integral between precision/recall - AUC - area under curve - final analysis in terms of AUC - makes most sense for imbalanced binary classification tasks

numerical score - compute score - number between 0&1

ROC curve for any algorithm output real number score

feed in probability to something - compute AUC - sweeps it up to 1 - compute precion/recall at every

threshold setting - java libraries to compute AUC scores to generate ROC curves - most common

area under curve for

-separation of data

-precision vs. recall or a different metric - success metric

12-11-12

Consolidated simulation results

-66% hold out validation/percentage split

-oversampling in logistic regression – AUC calculation

-shove a balanced data sheet into weka or try preprocessing options

-over fitting already???

Poster notes

s - 2.08x1.5

seal - 2.67 all around

poster size 40x32 – vertical vs. horizontal?

text boxes 12.5 across

Sections:

Intro

Data parsing/visualization

-milestone

-balancing data

-normalizing

-balanced vs. unbalanced data plot (potentially go into algorithms)

Algorithms used/plots

-table

-roc curves

-text

Results

Discussion

Acknowledgement

matlab RUSboost - specific for unbalanced data sets

oversample logistic

compare auc

compare models

matlab svm

matlab logistic regression

weka ensemble methods/etc

we made a few submissions to calculate Gini scores/Kaggle