Pruning

<http://stackoverflow.com/questions/28073351/pruning-decision-trees-in-python>

<https://github.com/sgenoud/scikit-learn/blob/4a75a4aaebd45e864e28cfca897121d1199e41d9/sklearn/tree/tree.py>

Some quick analysis of the data:

Decision Tree: Gini vs Entropy (Gini vs Information Gain)

Max\_depth

Use Pruning

Classification Problems:

Problem 1:

* Bank Notes
* <http://archive.ics.uci.edu/ml/datasets/banknote+authentication>
* 1372 instances; 762 class 0, 610 class 1
* No missing attributes
* Data Set:
  + Variance of Wavelet Transformed image (continuous)
  + Skewness of Wavelet Transformed image (continuous)
  + Curtosis of Wavelet Transformed image (continuous)
  + Entropy of image (continuous)
  + Class (integer) – Not described, assumed to be 0 = fake, 1 = real
* Give table describing data (min, max, mean, median, std dev) for class 0 vs class 1
* Problem: given attributes, predict whether a banknote is real or fake
* Citation:
  + Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
* Are the data linearly separable? Graph in 3 dimensions

Problem 2:

* “Haberman’s Survival Dataset”
* Predict Breast Cancer Survival
* Predict whether a patient will survive 5 years given the following:
  + Age of patient at the time of operation (continuous)
  + Year of the patients operation (+ 1900)
  + Number of positive axillary nodes detected
  + Survival Status
    - 1, patient survived 5 years or longer
    - 2, patient died within 5 years
* Data Set characteristics
  + 306 instances
  + 225 class 1
  + 81 class 0
* Give table of summary statistics by class
* Are the data linearly separable? Graph in 3 dimensions.

Task: Compare the results of 5 different supervised learning algorithms:

* Decision Trees + pruning
* Neural Network
* K-Nearest Neighbor
* SVM (try several kernels)
* Boost

Other stuff to note

* Overall, we are working with pretty well defined data sets here. They are pretty small (somewhat limited by the computing power of my laptop), and not very complicated. If we had different situations – i.e. missing values, etc, then this would have been a different story. I also didn’t do anything using categorical values or anything of that nature.

Conclusion

Final charts showing comparison of best output for each of the different algorithms for each

Charts/table to show:

Data Summary:

Min/max/mean/median/std dev by class for each data set

Decision Tree:

Output of each model using pruning, compare gini vs. information gain, max\_depth, pruning vs no pruning. Show running time. Can show PDF output of decision tree image as part of the appendix.

Analysis: What does this tree tell us about the data? Why aren’t we surprised at all?

SVM:

Four different kernels by different values of C (0.01, 0.1, 1, 10, 100)

Smaller values of = c = smaller margin in the hyperplane created by the SVM, larger values of C = larger margin (i.e. we would expect to see more misclassifications with smaller values of C)

Analysis: What does this value of SVM (or this kernel choice) tell us about the data?

KNN:

Show error by different values of K. Uniform vs. Weighted.

Analysis: What does this tell us about the data?

Boost:

AdaBoostClassifier defaults to DecisionTreeClassifier (No need to include that). No clue how to analyze this.

Neural Networks:

Vary on number of hidden nodes, number of hidden layers? Type of activation function. (Sigmoid vs. Tanhlayer). Bias?

Acknowledge: This is a black box, unfortunately we don’t gain much insight into the data set or the problem.

Compare/Contrast the two data sets. Why do certain algorithms work better/worse for each? What might we expect?

Structure of Paper:

1. Classification Problems/Data Set introduction
2. Analysis
3. Conclusion
4. Other stuff to think about