Classification Comparisons

Math 3220 Data Mining Methods
Angelo Parker

Overview

- Classification
- C5.0
- Rpart
- SVM
- The example datasets
- Classification comparisons

Classification

- The method of taking data and breaking it down into classes to interpret certain trends and information that can be used to make predictions on future data.
- The are various methods for classifying data. The three that will be discussed are C5.0, Rpart, and Support Vector Machines.

C5.0

- C5.0 is an improved classification algorithm based on the earlier ID3's entropy and information gain's formula's:

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x) \qquad IG(A,S) = H(S) - \sum_{t \in T} p(t) H(t)$$

  - Entropy is a measure of uncertainty in the data.
  - Information Gain is the difference of different Entropies as more attributes get applied to the data.

- The goal is to shrink the amount of Entropy and increase the Information Gain.

- C5.0 will create a set of inequality rules that are determined to "best" split the data depending on the attributes of the greatest influence at that particular split.

- C4.5 algorithm created by Ross Quinlan in 1992

An example of C5.0 on Iris:

```
C5.0.default(x = IrisSet[1:4], y = IrisSet[, 5])
C5.0 [Release 2.07 GPL Edition]            Sun Oct 01 20:45:00 2017
-------------------------------
Class specified by attribute `outcome'
Read 150 cases (5 attributes) from undefined.data
Decision tree:
PL <= 1.9: Setosa (50)
PL > 1.9:
:...PW > 1.7: Virginica (46/1)
    PW <= 1.7:
    :...PL <= 4.9: Versicolor (48/1)
        PL > 4.9: Virginica (6/2)
Evaluation on training data (150 cases):
               Decision Tree
             ----------------
             Size      Errors
                4    4( 2.7%)   <<
              (a)   (b)   (c)    <-classified as
             ----  ----  ----
               50                (a): class Setosa
                     47     3    (b): class Versicolor
                      1    49    (c): class Virginica
           Attribute usage:
           100.00%    PL
            66.67%    PW
```

CART (Rpart)

- Rpart, the R version of CART, works similarly to C5.0 but utilizes a formula to minimize Gini Impurity and variance reduction shown below.
  - Gini Impurity is the chance that a random instance will be misclassed.
  - Variance is a description used to convey whether the characteristics of an instance or data set is significantly unique to another instance or data set.

$$I_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i (1 - p_i) = \sum_{i=1}^{J} (p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

$$I_V(N) = \frac{1}{|S|^2} \sum_{i \in S} \sum_{j \in S} \frac{1}{2}(x_i - x_j)^2 - \left(\frac{1}{|S_t|^2} \sum_{i \in S_t} \sum_{j \in S_t} \frac{1}{2}(x_i - x_j)^2 + \frac{1}{|S_f|^2} \sum_{i \in S_f} \sum_{j \in S_f} \frac{1}{2}(x_i - x_j)^2 \right)$$

- Cart was developed by four authors Breiman, Friedman, Olshen, and Stone in 1984 (Brieman, 2017)

Rpart example on Iris:

```
rpart(formula = IrisPred, method = "class")  n= 150
CP nsplit rel error xerror       xstd
1 0.50    0       1.00   1.20 0.048989792
0.44      1       0.50   0.75 0.061237243
0.01      2       0.06   0.08 0.02751969
Variable importance
IrisSet$PW IrisSet$PL IrisSet$SL IrisSet$SW
34         31         21         13
Node number 1: 150 observations,    complexity param=0.5  predicted class=Setosa      expected
loss=0.6666667  P(node) =1    class counts:    50    50    50   probabilities: 0.333 0.333 0.333   left
son=2 (50 obs) right son=3 (100 obs)  Primary splits:      IrisSet$PL < 2.45 to the left,
improve=50.00000, (0 missing)      IrisSet$PW < 0.8  to the left,  improve=50.00000, (0 missing)
IrisSet$SL < 5.45 to the left,  improve=34.16405, (0 missing)      IrisSet$SW < 3.35 to the right,
improve=18.05556, (0 missing)  Surrogate splits:      IrisSet$PW < 0.8  to the left,  agree=1.000,
adj=1.00, (0 split)      IrisSet$SL < 5.45 to the left,  agree=0.920, adj=0.76, (0 split)      IrisSet$SW
< 3.35 to the right, agree=0.827, adj=0.48, (0 split)
Node number 2: 50 observations  predicted class=Setosa      expected loss=0  P(node) =0.3333333    class
counts:    50     0     0   probabilities: 1.000 0.000 0.000
Node number 3: 100 observations,    complexity param=0.44  predicted class=Versicolor  expected loss=0.5
P(node) =0.6666667    class counts:     0    50    50   probabilities: 0.000 0.500 0.500   left son=6 (54
obs) right son=7 (46 obs)  Primary splits:      IrisSet$PW < 1.75 to the left,  improve=38.969400, (0
missing)      IrisSet$PL < 4.75 to the left,  improve=37.353540, (0 missing)      IrisSet$SL < 6.15 to
the left,  improve=10.686870, (0 missing)      IrisSet$SW < 2.45 to the left,  improve= 3.555556, (0
missing)  Surrogate splits:      IrisSet$PL < 4.75 to the left,  agree=0.91, adj=0.804, (0 split)
IrisSet$SL < 6.15 to the left,  agree=0.73, adj=0.413, (0 split)      IrisSet$SW < 2.95 to the left,
agree=0.67, adj=0.283, (0 split)
Node number 6: 54 observations  predicted class=Versicolor  expected loss=0.09259259  P(node) =0.36
class counts:     0    49     5   probabilities: 0.000 0.907 0.093
Node number 7: 46 observations  predicted class=Virginica   expected loss=0.02173913  P(node) =0.3066667
class counts:     0     1    45   probabilities: 0.000 0.022 0.978
```

SVM

- SVMs are binary graphical classification models that use regression lines to separate and push data points closer to each other into more distinct groups.

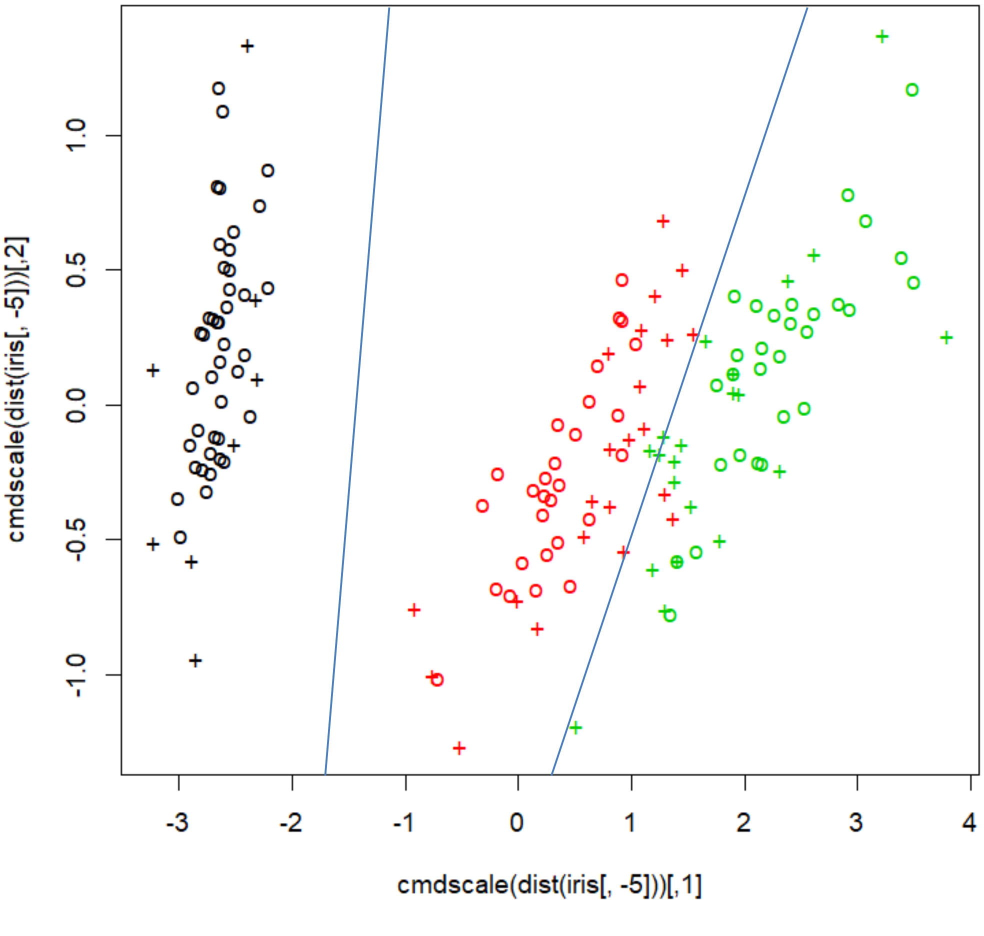
$$\vec{w} \cdot \vec{x} - b = 0,$$

$$\vec{w} \cdot \vec{x} - b = -1. \qquad \vec{w} \cdot \vec{x} - b = 1$$

  - Hard Margin SVMs
  - Soft Margin SVMs
  - Non-linear SVMs
  - Linear SVMs
  - Formulas that plot multiple SVMs

- In 1995, the most referred method, was finalized by Vapnik and Cortes.

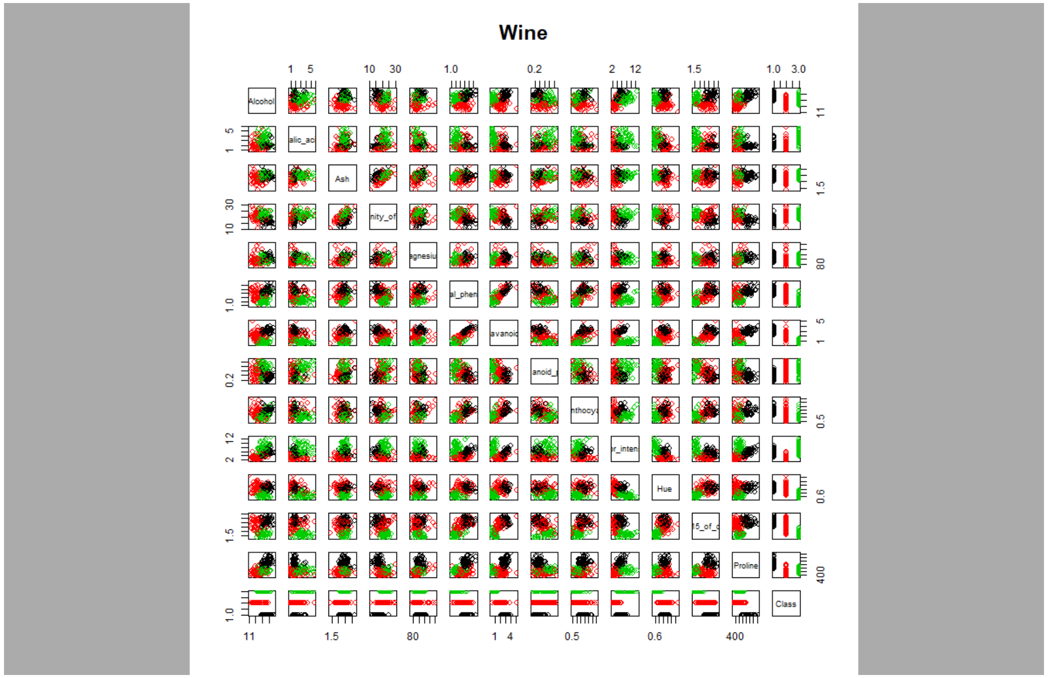
SVM example on Iris



Data Sets

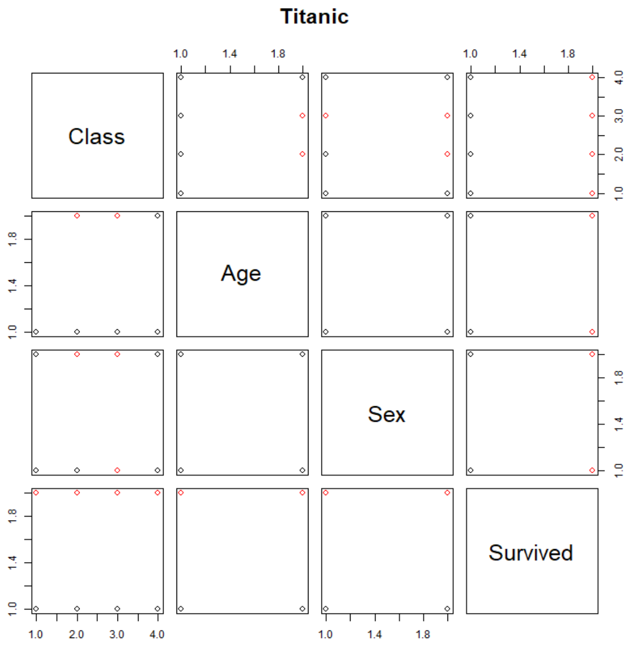
- There were three data sets used for this presentation. Each are multivariate.
  - Iris
  - Wine
  - Titanic

Wine (Data Set)



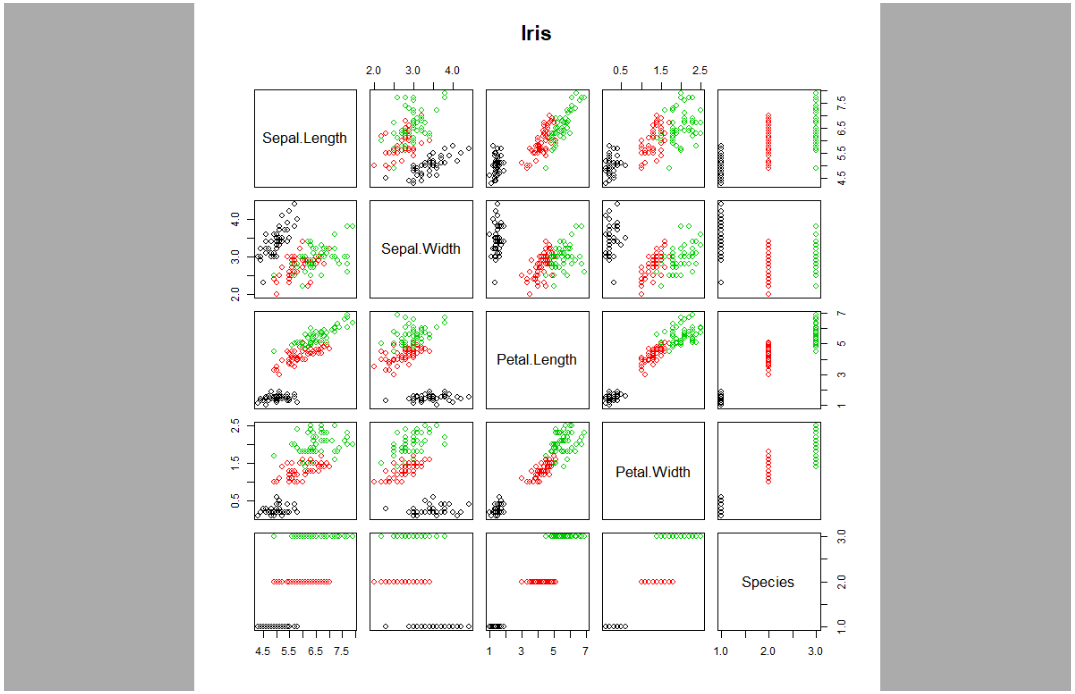
The Wine data set is a set of 153 different wines from three Italian cultivers, divided by 13 attributes: Alcohol, Malic Acid, Ash, Alkalinity of Ash, Magnesium, Number of Phenols, Proanthocyanins, Color intensity, Hue, Proline, and OD280/OD315 of diluted wines.

Titanic (Data Set)



The Titanic data set is a roster of 2201 passengers and crew aboard the Titanic. The instances are categorized by class or crew, age, sex and whether they survived or not.

Iris



Based on a paper by Sir R. A. Fisher, this is a set of three types of Iris plants Setosa, Versicolor, and Virginica, 50 each. Each instance is measured by four physical attributes. This is a classic statistic and machine learning practice data set.

Comparisons (Iris)

Iris C5.0

	(a)	(b)	(c)
	----	----	----
Setosa	50		
Versicolor		47	3
Virginica		1	49

Iris Rpart

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	49	5
virginica	0	1	45

Iris SVM

irispred	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	2	48

Percentage of Misclassification:

C5.0: 4/150 (2.67%)

Rpart: 6/150 (4%)

SVM: 4/150 (2.67%)

Comparisons (Wine)

Wine C5.0

	(a)	(b)	(c)
	----	----	----
Class_1	47		
Class_2		60	1
Class_3			45

Wine Rpart

truepred	Class_1	Class_2	Class_3
Class_1	43	0	0
Class_2	4	60	0
Class_3	0	1	45

Wine SVM

WinePred	Class_1	Class_2	Class_3
Class_1	47	0	0
Class_2	0	61	0
Class_3	0	0	45

Percentage of Misclassification:

C5.0: 1/153 (0.65%) Rpart: 5/153 (3.27%) SVM: 0/153 (0%)

Comparisons (Titanic)

Titanic C5.0

	(a)	(b)	<-classified as
	----	----	
No	1470	20	
Yes	457	254	

Titanic Rpart

truepred	No	Yes
No	1470	441
Yes	20	270

Titanic SVM

TitanicPred	No	Yes
No	1470	441
Yes	20	270

Percentage of Misclassification:

C5.0: 477/2201 (21.67%) Rpart: 461/2201 (20.95%) SVM: 461/2201 (20.95%)

Summary and Conclusion

- Understanding of Classifications.
- There are multiple Classification methods depending on the desired information.
- SVMs is becoming the more popular algorithm.
- Brief on C5.0, Rpart, and SVMs.
- Other data sets may affect the Methods differently.

References

- https://archive.ics.uci.edu/ml/datasets/wine
- https://archive.ics.uci.edu/ml/datasets/Iris
- Data Mining Methods report 4
- Data mining methods report 2
- Brieman, F. O. (2017, April 1). *Package 'rpart'*. Retrieved from rpart.pdf
- *C4.5 Algorithm*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/C4.5_algorithm
- *Classification and regression trees*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Predictive_analytics#Classification_and_regression_trees_.28CART.29
- *Decision tree learning*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Decision_tree_learning
- *ID3 Algorithm*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/ID3_algorithm
- Meyer, D. (2017, February 2). *Package 'e1071'.* Retrieved from CRAN: https://cran.r-project.org/web/packages/e1071/e1071.pdf
- Meyer, D. (2017, February 1). *Support Vector Machines.* Retrieved from CRAN: https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf
- Parker, A. (2017). *Report 2.*