TECHNOLOGY FUNDAMENTALS FOR ANALYTICS

Jason Kuruzovich

Introduction to Models 2

Agenda

- Midterm Example Questions
- Kaggle Clarification
- Supervised Segmentation Models
- Regression Models

Sample Questions

1. Below is an example of what type of encoding?

```
{ "year": 1997,
  "make": "Ford",
  "model": "E350" }
```

a) Delimited file
b) XML
c) JSON
d) ODBC
e) SQL

Sample Questions

2.. CRISP describes a ____________________?

a) Data warehouse

b) R package

c) NOSQL Database

d) R Package

e) Process for Data Mining

Sample Questions

3. The Titanic model provided a context in which ________________ was the objective of the model.

a) Classification
b) Regression
c) Clustering
d) Filtering
e) Aggregating

In addition, all questions like those from lab are possible [select columns, select rows, create features, etc.]

Kaggle

Supervised Segmentation

Review: What does it mean that a model is supervised?

Three Types of Categorical Models

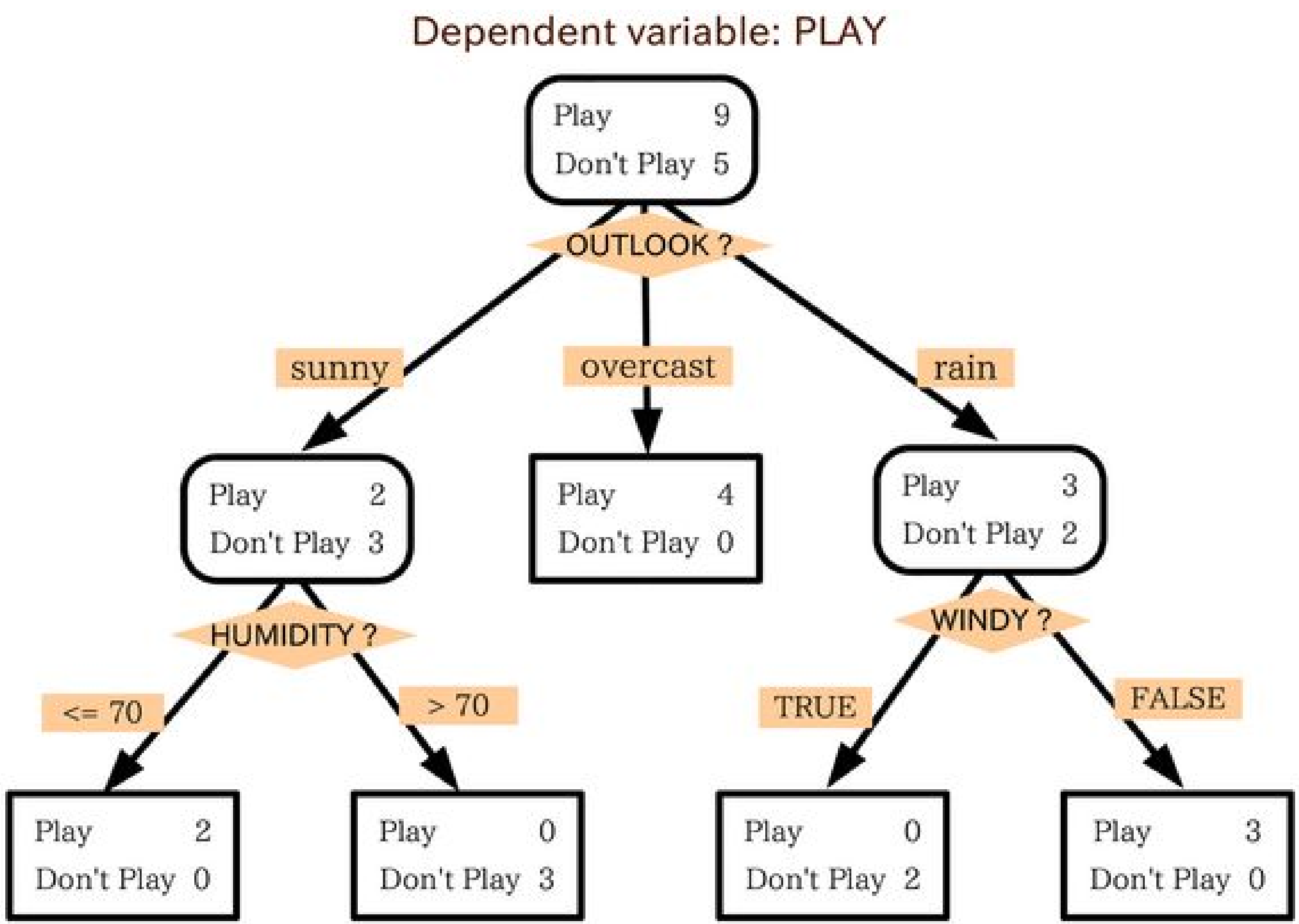
- Tree Based Models
  - Decision Tree
  - Random Forest (Ensemble Method)
- Linear Functions
  - Logistic Regression
  - Support Vector Machine
- Nonlinear Functions
  - Support Vector Machines
  - Neural Networks

Decision Tree

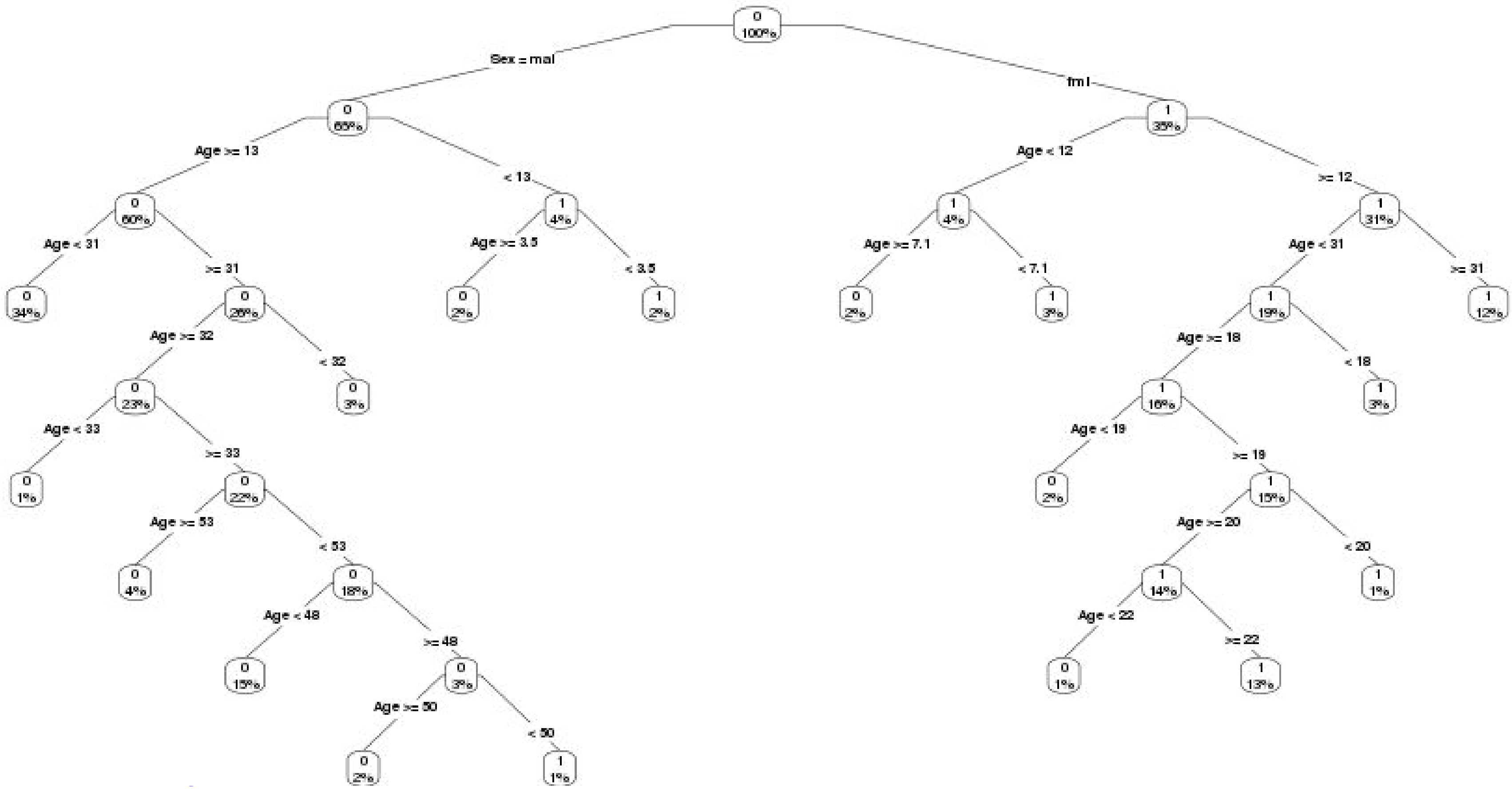
“A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes).” –Wikipedia

https://en.wikipedia.org/wiki/Decision_tree

Decision Tree - Golf



Tree Models



0
100%
Sex = mal
fml
0
65%
1
35%
Age >= 13
< 13
Age < 12
>= 12
0
60%
1
4%
1
4%
1
31%
Age < 31
>= 31
Age >= 3.5
< 3.5
Age >= 7.1
< 7.1
Age < 31
>= 31
0
34%
0
26%
0
2%
1
2%
0
2%
1
3%
1
19%
1
12%
Age >= 32
< 32
Age >= 18
< 18
0
23%
0
3%
1
16%
1
3%
Age < 33
>= 33
Age < 19
>= 19
0
1%
0
22%
0
2%
1
15%
Age >= 53
< 53
Age >= 20
< 20
0
4%
0
18%
1
14%
1
1%
Age < 48
>= 48
Age < 22
>= 22
0
15%
0
3%
0
1%
1
13%
Age >= 50
< 50
0
2%
1
1%

Supervised Segmentation

- Outcome with two or more categories
- Multiple information attributes that may be relevant to outcome
- In titanic, we used our knowledge of shipwrecks ("women & children first!"), but how might we systematize selection of appropriate attributes?

Entropy! High Entropy, High Information

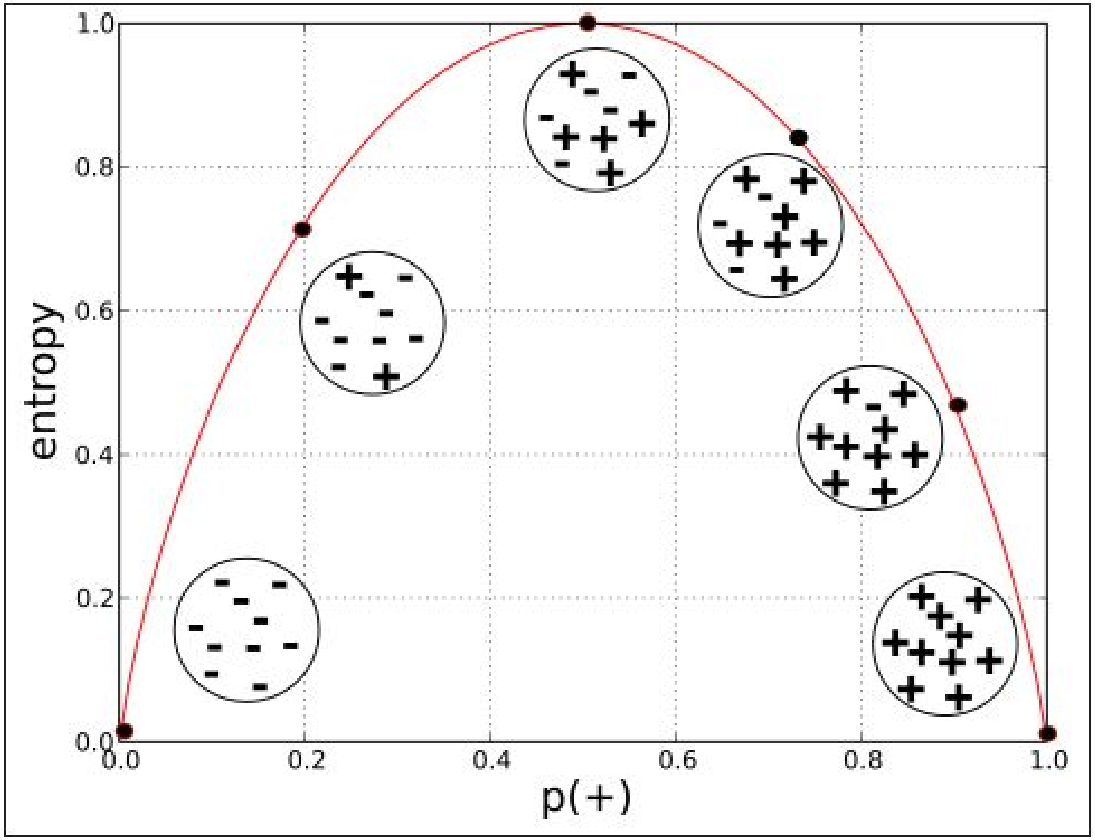


Figure 3-3. Entropy of a two-class set as a function of p(+).

Entropy = -p1(log(p1)) – p2(log(p2)) -…

InformationGain =entropy(parent) – [average entropy(children)]

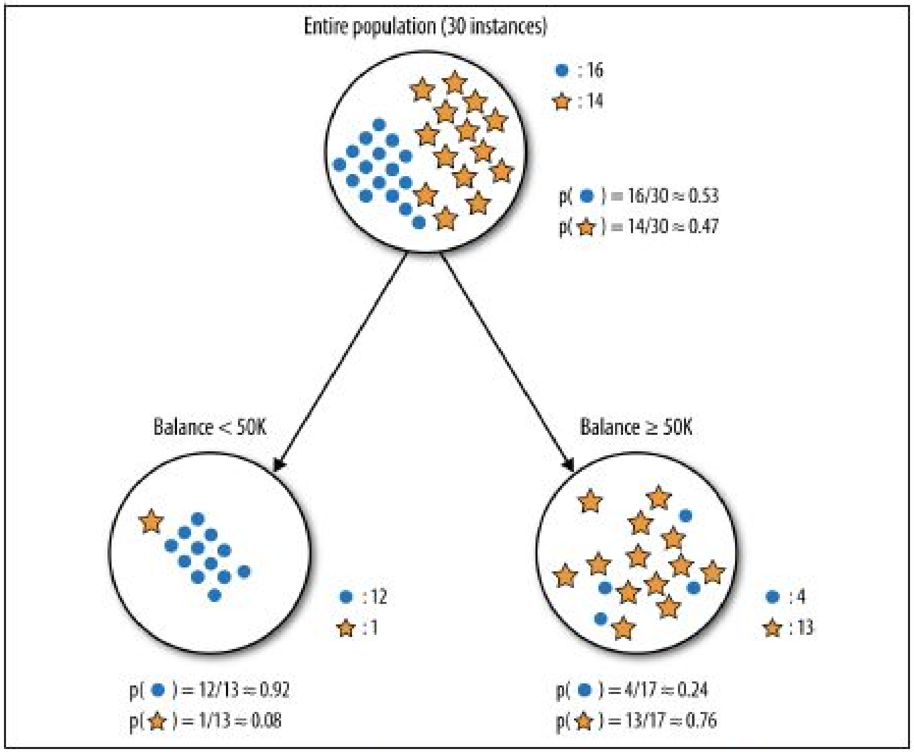


Figure 3-4. Splitting the "write-off" sample into two segments, based on splitting the Balance attribute (account balance) at 50K.

Building Models with RPART

Language: R   Library: rpart

rpart(*formula*, **data=**, **method=,control=)**

Example

my_tree <- rpart(Survived ~ Pclass + Sex + SibSp + Parch + Fare + Embarked + Title + family_size, data = train_new, method = "class", control=rpart.control(cp=0.0001))

https://cran.r-project.org/web/packages/rpart/rpart.pdf

Building Models with RPART

Formula

rpart(*formula***, data=, method=,control=)**

Survived ~ Pclass + Sex + SibSp + Parch + Fare + Embarked + Title + family_size

Survived = f(Pclass, Sex, ...)

Building Models with RPART

Method

rpart(*formula*, **data=**, **method=,control=)**

- RPART will attempt to select the right method of "anova", "poisson", "class" or "exp"
  - Factor Dependent Variable (like titanic$Survived) -> *"class" -> Generates a classification tree*
  - *Continuous Dependent Variable (like titanic$Age) -> "anova" -> Generates a regression tree*
  - Survival Analysis (model time to event) ->"exp"
  - Multiple DVs -> "poisson"

Building Models with RPART

Control (Provides Tuning to Prevent Overfit)

rpart(*formula*, **data=**, **method=,control=)**

- *control=rpart.control(minsplit=30, cp=0.001)*
- Minsplit is the minimum number of observations in a node [DEFAULT = 20]
- Cp must decrease the overall lack of fit by a factor of 0.001 (cost complexity factor) [DEFAULT = 0.01]
- Each of these prevent overfitting

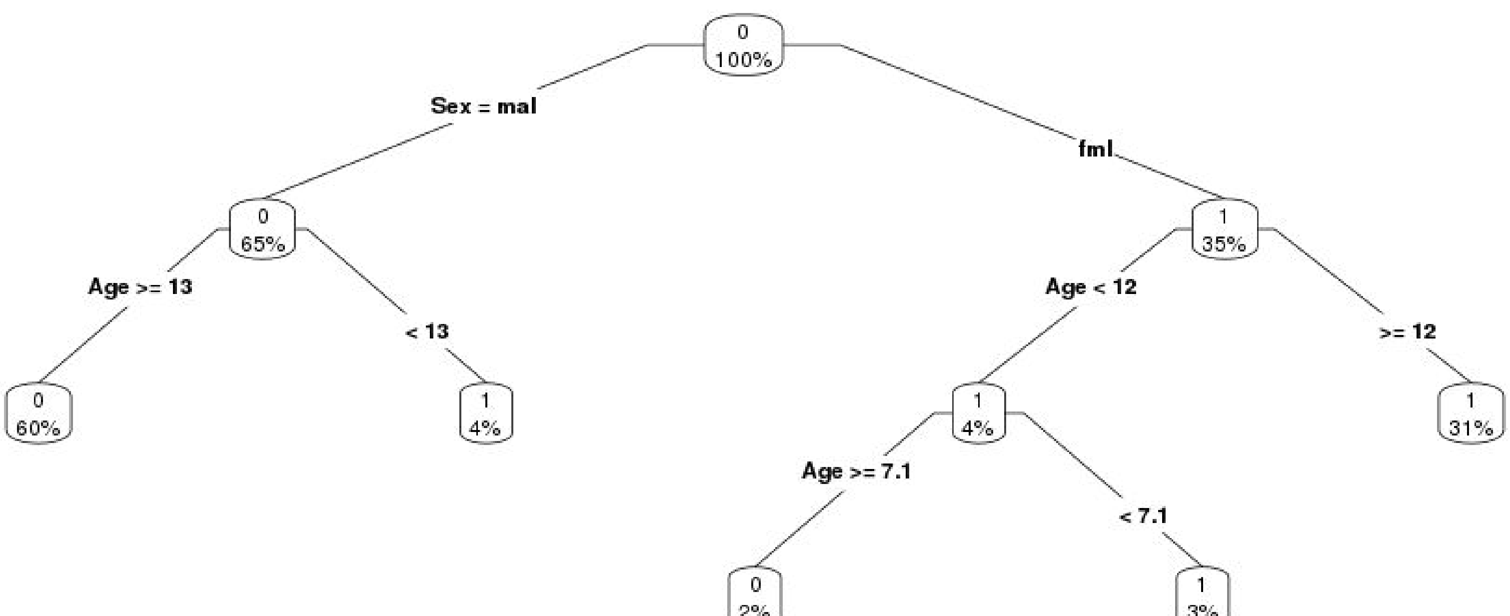
http://www.statmethods.net/advstats/cart.html

Building Models with RPART

Control (Provides Tuning to Prevent Overfit)

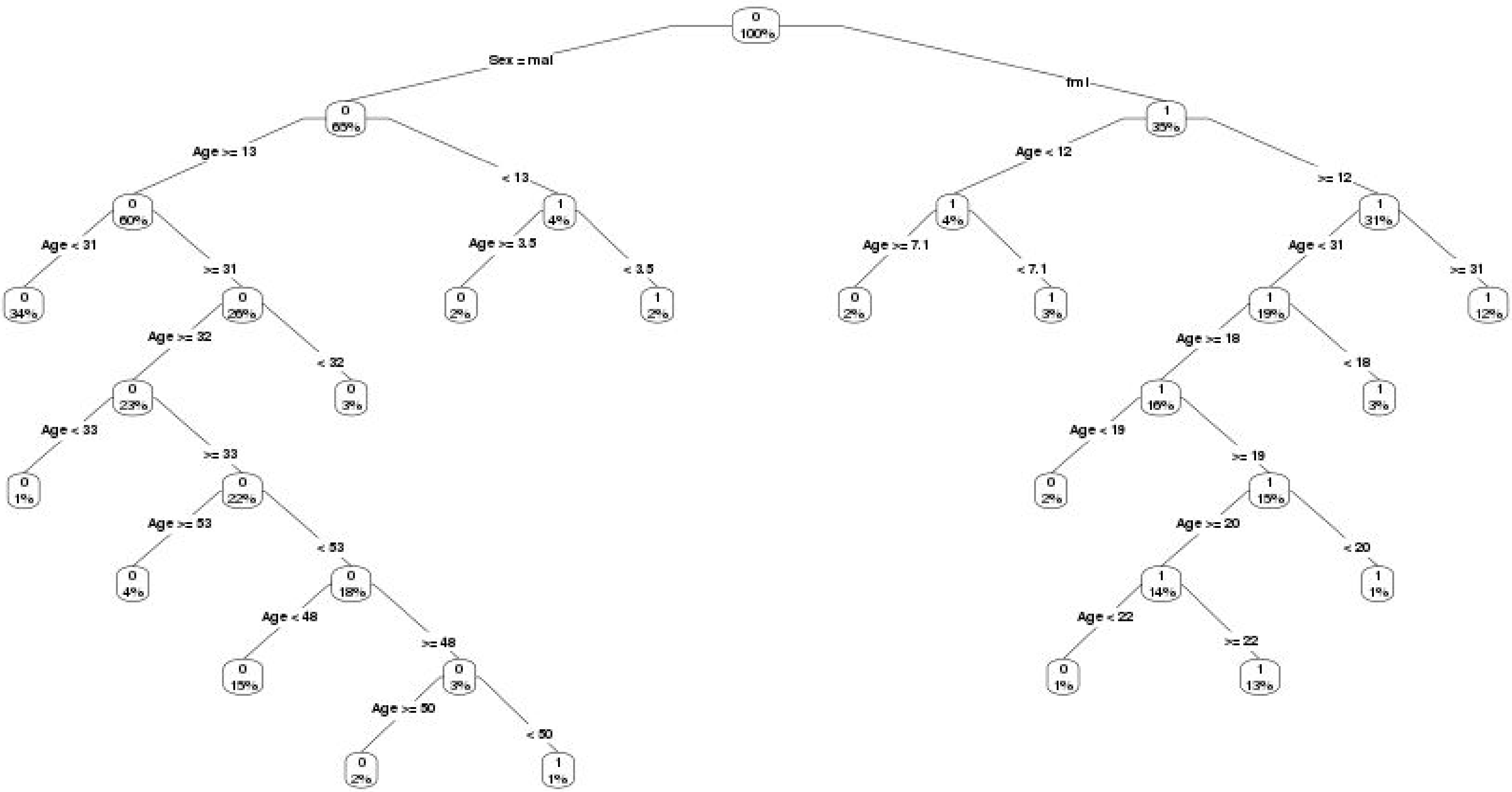
my_tree <- rpart(Survived ~ Sex + Age, data = train_new, method = "class")

my_tree <- rpart(Survived ~ Sex + Age, data = train_new, method = "class", control=rpart.control(minsplit=20, cp=0.01))



If we set *cp=0.0001, with tree grow in complexity?*

cp=0.0001



0
100%
Sex = mal
fml
0
65%
1
35%
Age >= 13
< 13
Age < 12
>= 12
0
60%
1
4%
1
4%
1
31%
Age < 31
>= 31
Age >= 3.5
< 3.5
Age >= 7.1
< 7.1
Age < 31
>= 31
0
34%
0
26%
0
2%
1
2%
0
2%
1
3%
1
19%
1
12%
Age >= 32
< 32
Age >= 18
< 18
0
23%
0
3%
1
16%
1
3%
Age < 33
>= 33
Age < 19
>= 19
0
1%
0
22%
0
2%
1
15%
Age >= 53
< 53
Age >= 20
< 20
0
4%
0
18%
1
14%
1
1%
Age < 48
>= 48
Age < 22
>= 22
0
15%
0
3%
0
1%
1
13%
Age >= 50
< 50
0
2%
1
1%

Trees as Sets of Rules

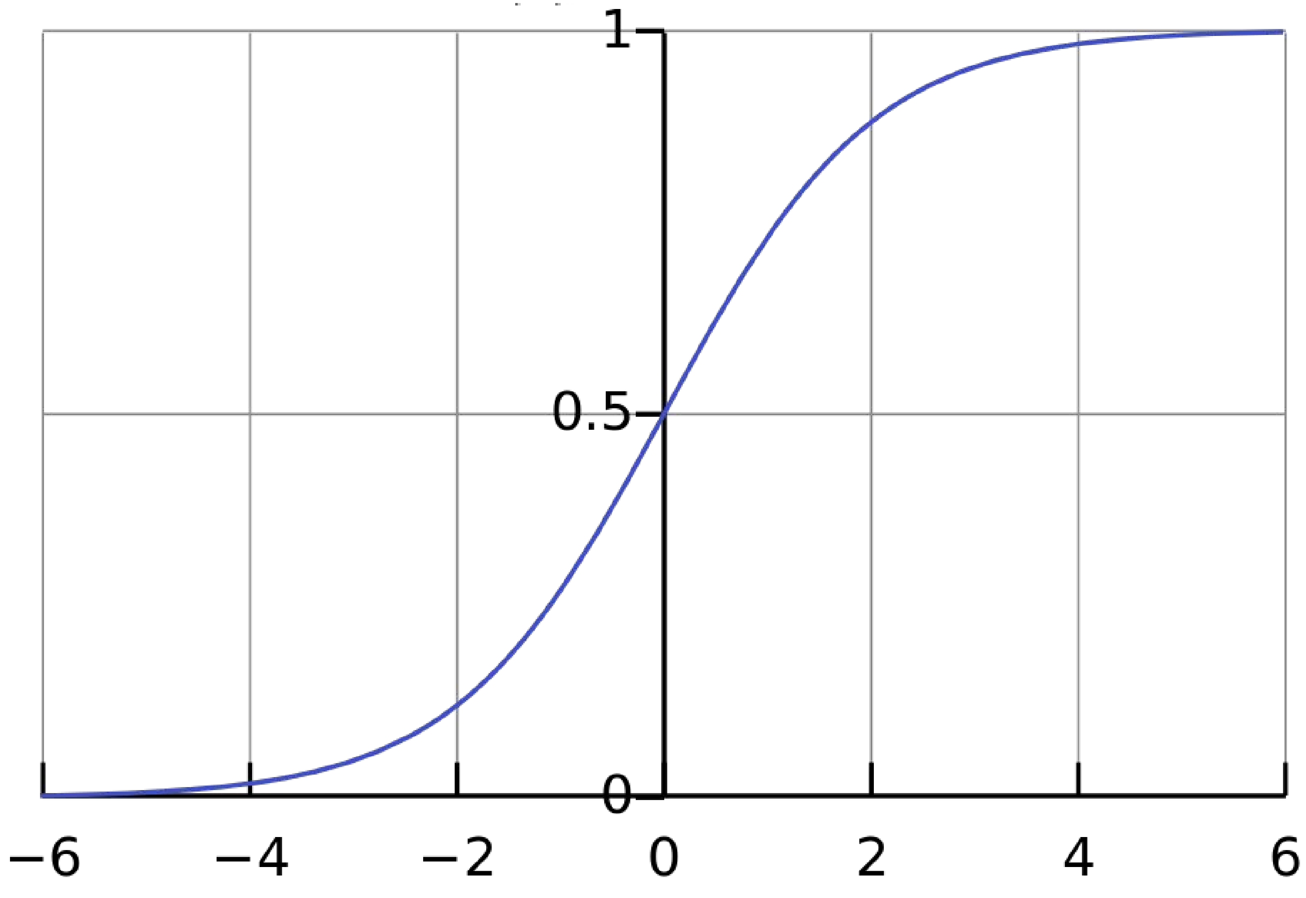
- Each Tree can be conceptualized as a set of logical if statements
- Predicting using the model runs a set of independent variables

Logistic Regression

- Dependent variable categorical
  - Binary logistic regression (2 outcomes)
  - Multinomial logistic regression (More than 2 outcomes)
- Fits a linear model of attributes to data (not tree)
- The term regression is omewhat of a misnomer, as we said regression is for continuous variables

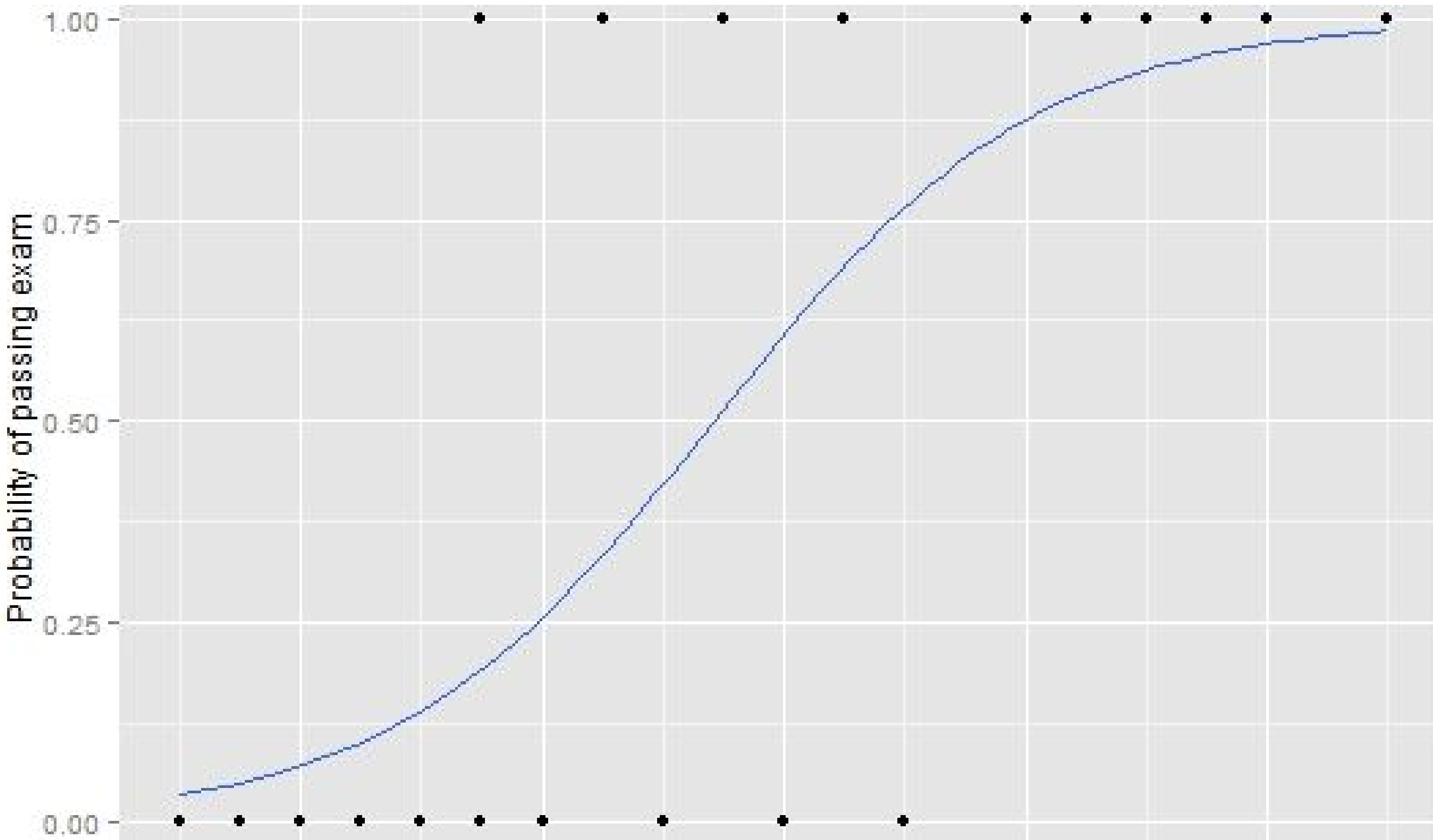
Logistic Function

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



By Qef (Created from scratch with gnuplot) [Public domain], via Wikimedia Commons

Logistic Function



Logistic Regression in R

This we will see later as the *general linear model* with specification of family=binomial for logistic regression

model<-glm(Survived ~ Sex + Pclass + Age + SibSp + Parch + Fare + Embarked, data = trainset, family = "binomial")

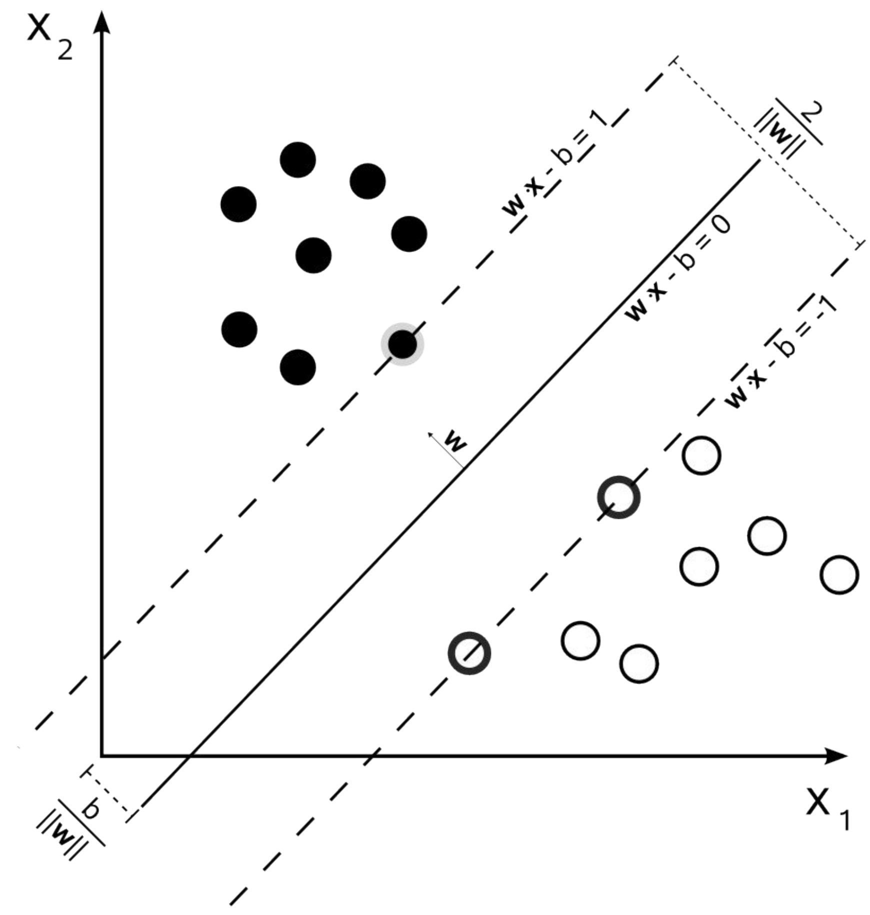
#Prediction is continuous, so this changes to 1 if >0.5 and 0 otherwise.

test$Survived <- ifelse(predict(model, test, type="response")>0.5,1,0)

https://www.kaggle.com/jasonkuruzovich/titanic/logistic-regression/edit

Support Vector Machine

- Attempts to maximize the margin between classes
- They are (like logistic regression) a discriminant function of attributes (not tree)



"Svm max sep hyperplane with margin" by Cyc - Own work. Licensed under Public Domain via Wikimedia Commons

Building Models with SVM

Language: R   Library: e1071

SVMmod <- svm(train_feature[,c(2:num_col)], train_feature[,1])

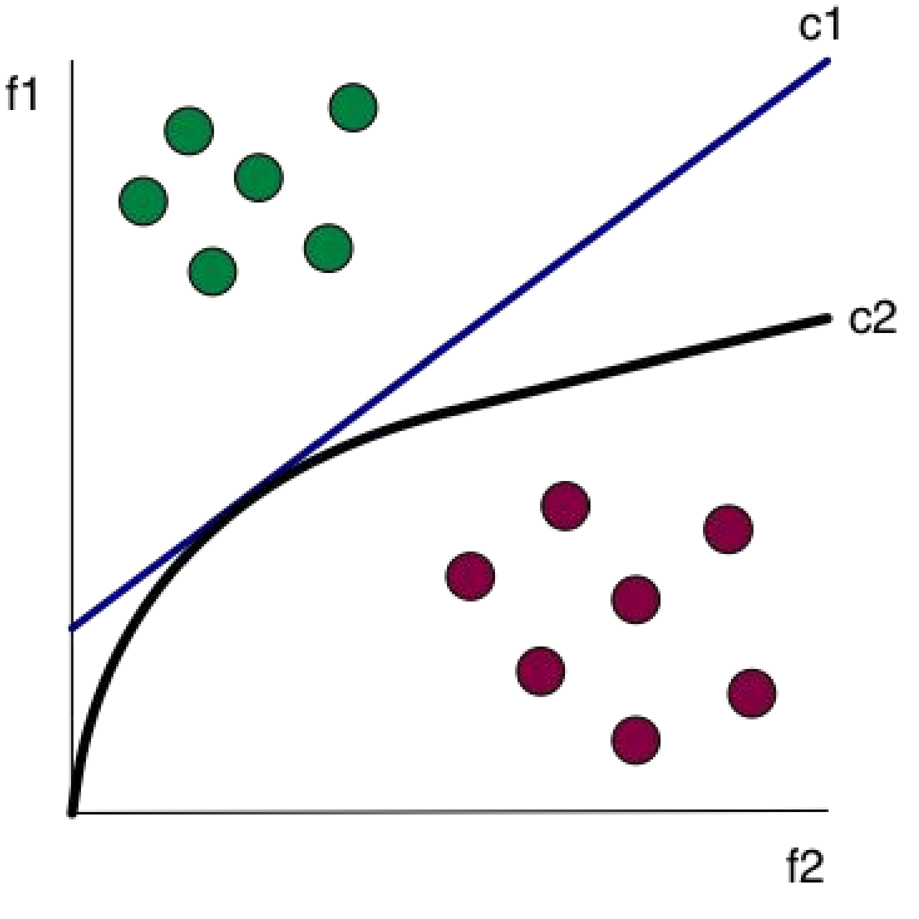
Example

SVMmod <- svm(train_feature[,c(2:num_col)], train_feature[,1])

SVMpredictions_test <- predict(SVMmod, test_feature)

https://www.kaggle.com/jasonkuruzovich/titanic/support-vector-machine-classification/edit

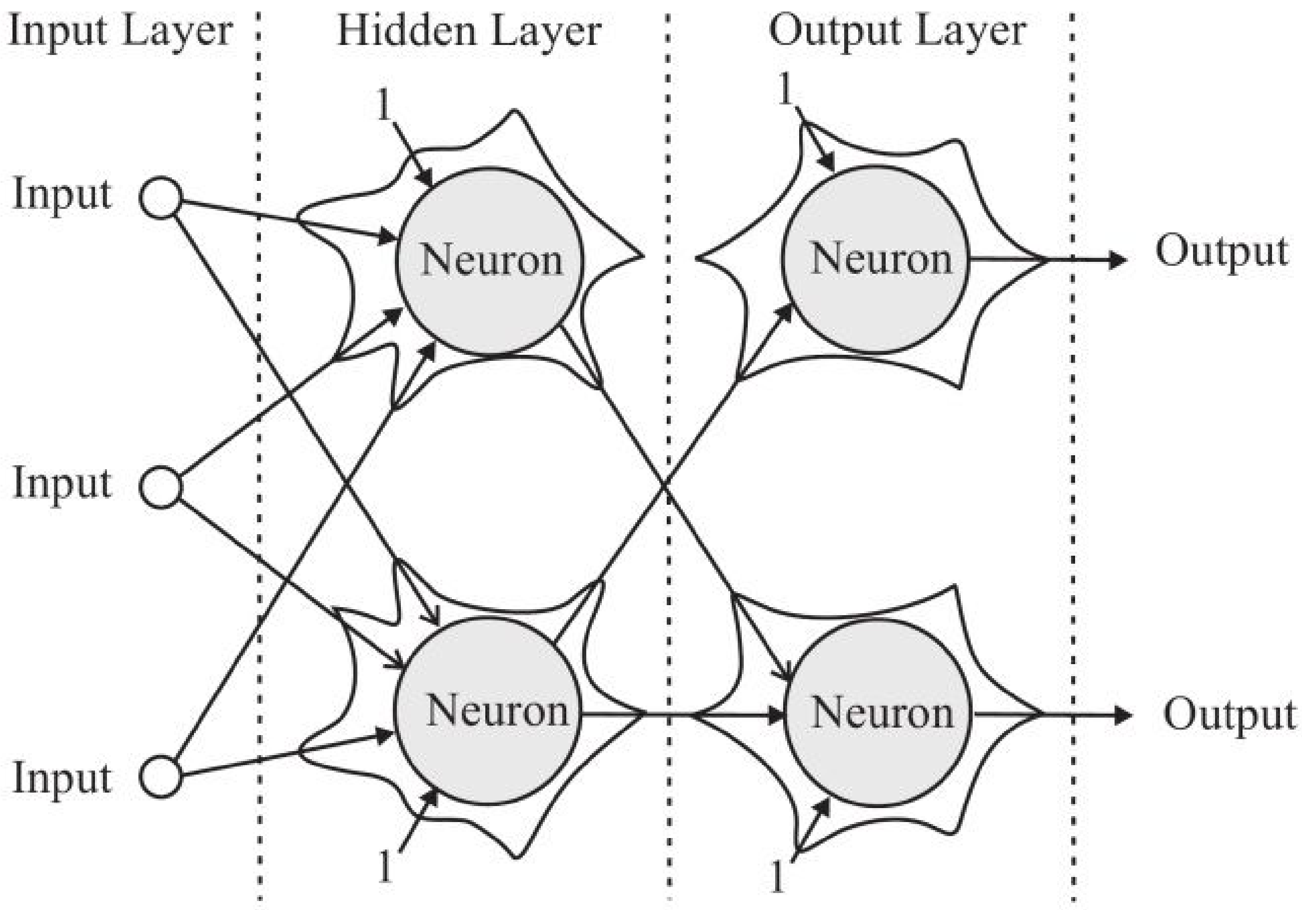
Linear (c1) and Nonlinear (c2) Classification



Neural Networks - Prediction

- Artificial neural networks are computational models inspired by animal central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition
- Include systems of interconnected "neurons" which compute values from network

Neural Networks



Building Models with Neural Networks

Language: R   Library: neuralnet

Example

nn <- nnet(Survived ~ Age + Fare + Embarked+ SibSp + Parch, train, size = 100)

nn.prediction <- predict(nn, newdata = train)

https://www.kaggle.com/jasonkuruzovich/titanic/make-neural-networks-compete/edit

Neural Networks

- Tuning Neural networks is beyond scope
  - Normalize data before training
  - How many hidden layers?

- https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf

What do you get when you put a lot of trees together?

Random Forest

- Random forest is an *ensemble* learning method that combines *feature selection* and decision trees
  - Randomly select a subset of features
  - Output the class of the mode (most frequently occurring prediction) of the trees

Building Models with Random Forests

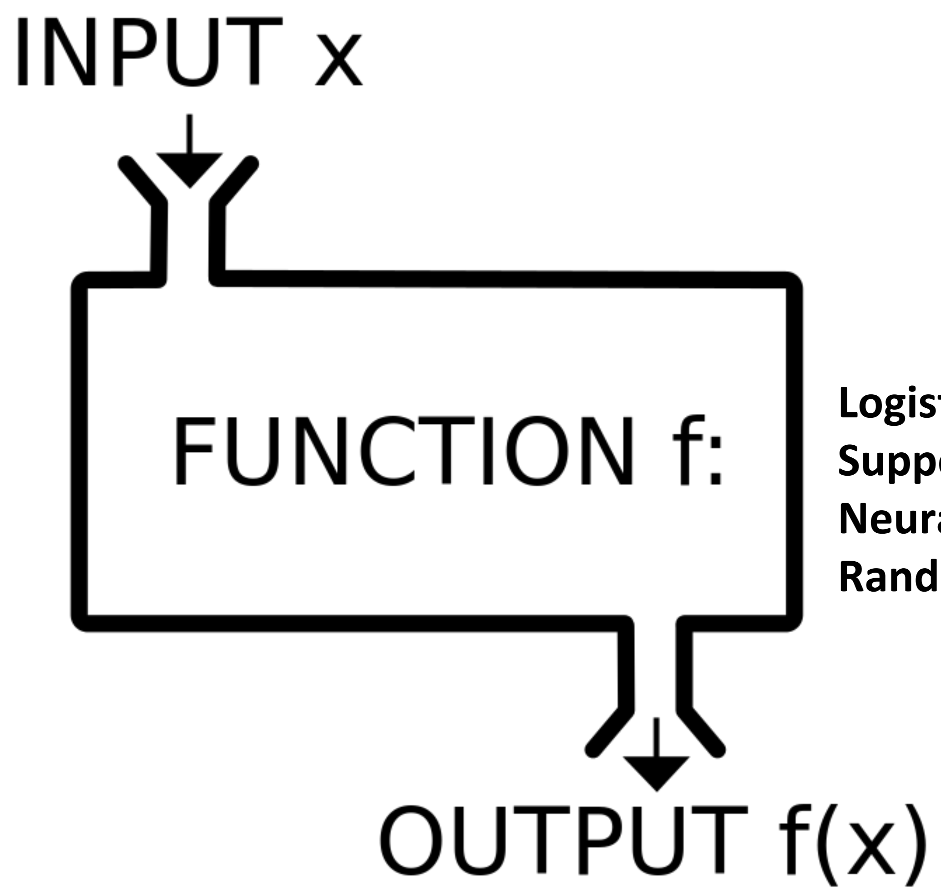
Language: R   Library: randomForest

Example

- rf <- randomForest(train, as.factor (train$Survived), ntree=100, importance=TRUE)

https://www.kaggle.com/jasonkuruzovich/titanic/random-forest-benchmark-r/edit

Independent or Explanatory Variables



Logistic Regression
Support Vector Machine
Neural Network
Random Forest

Target or Dependent Variable is Categorical

Regression

What is difference between regression and classification?

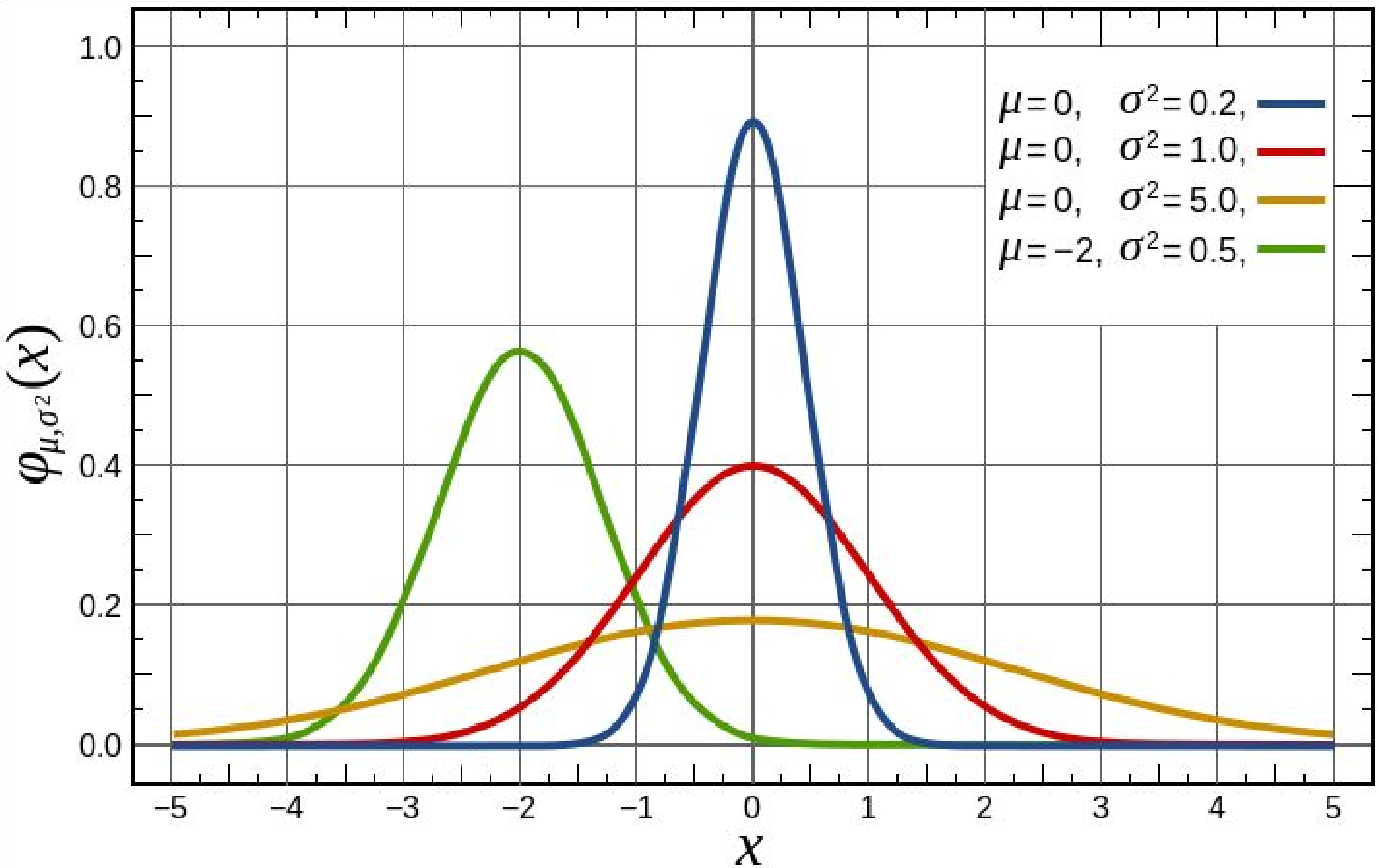
Regression vs. Classification

- Regression
  - Continuous Dependent Variable

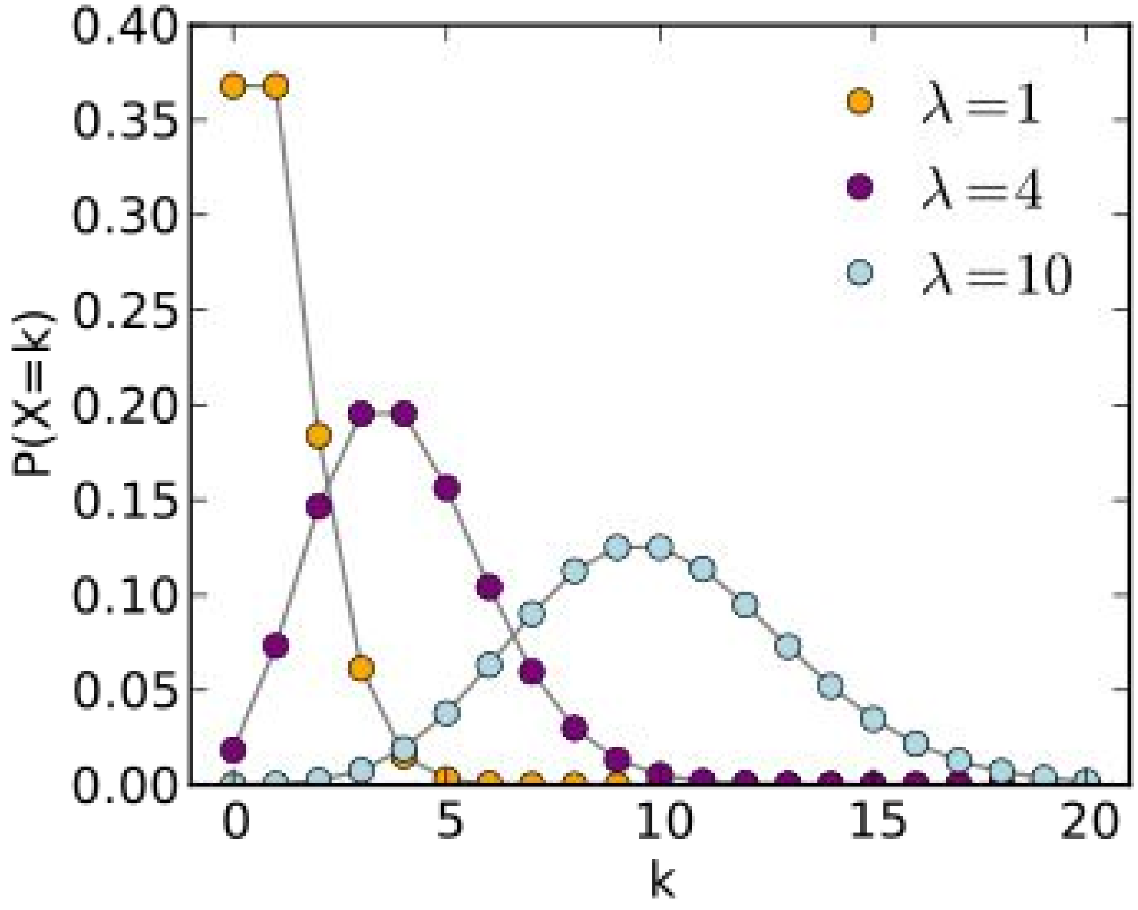
- Classification
  - Category Dependent Variable

Are there different distributions of continuous variables?

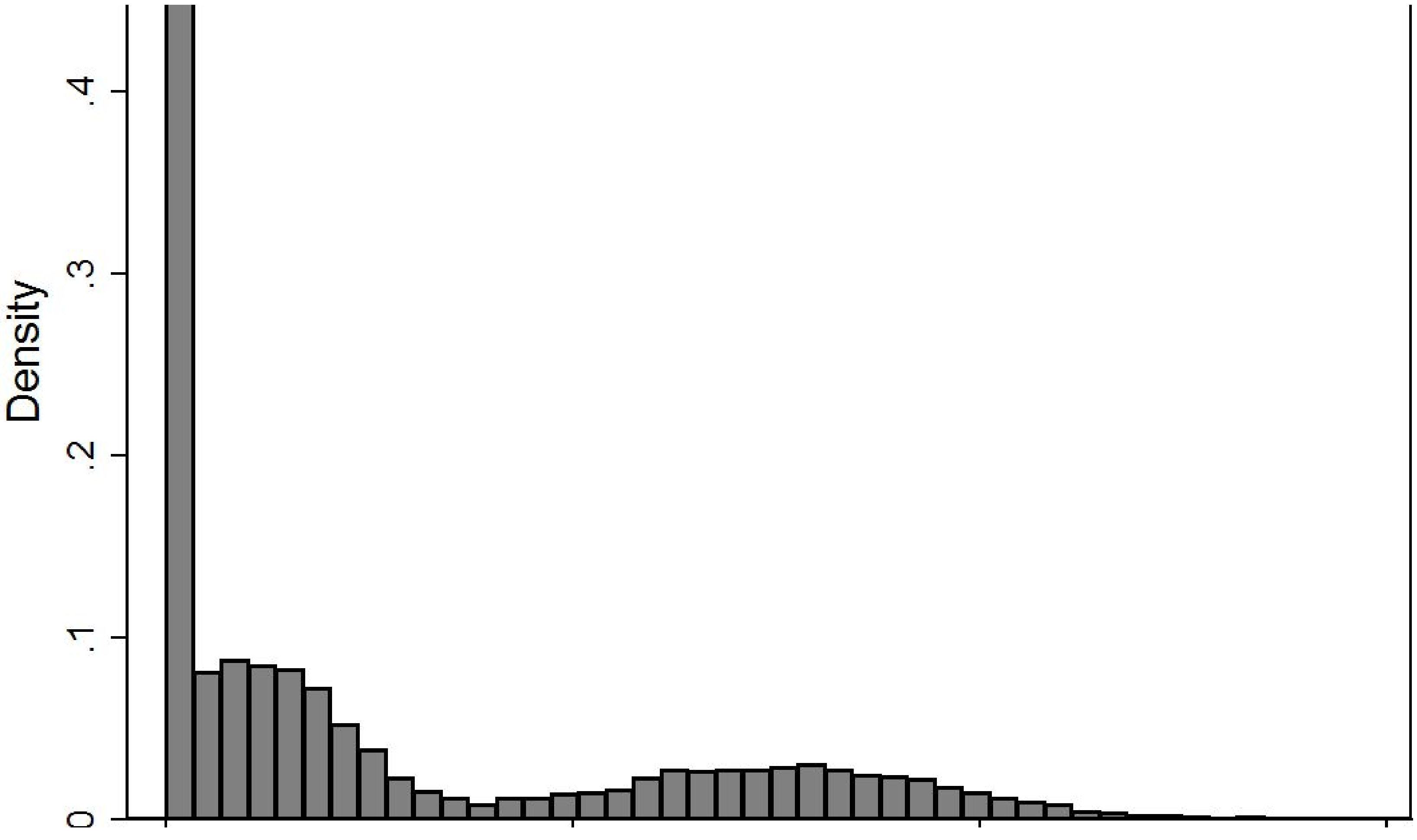
Normal Distribution



Poisson Distribution



Zero Inflated



FROM: http://home.uchicago.edu/~mcoca/docs/mixture.png

Distributions and Regression

Regression Models					
Robust Regression	Stata	SAS			R
Models for Binary and Categorical Outcomes					
Logistic Regression	Stata	SAS	SPSS	Mplus	R
Exact Logistic Regression	Stata	SAS			R
Multinomial Logistic Regression	Stata	SAS	SPSS	Mplus	R
Ordinal Logistic Regression	Stata	SAS	SPSS	Mplus	R
Probit Regression	Stata	SAS	SPSS	Mplus	R
Count Models					
Poisson Regression	Stata	SAS	SPSS	Mplus	R
Negative Binomial Regression	Stata	SAS	SPSS	Mplus	R
Zero-inflated Poisson Regression	Stata	SAS		Mplus	R
Zero-inflated Negative Binomial Regression	Stata	SAS		Mplus	R
Zero-truncated Poisson	Stata	SAS			R
Zero-truncated Negative Binomial	Stata	SAS		Mplus	R
Censored and Truncated Regression					
Tobit Regression	Stata	SAS		Mplus	R
Truncated Regression	Stata	SAS			R
Interval Regression	Stata	SAS			R

Different regression models for different dependent variable distributions

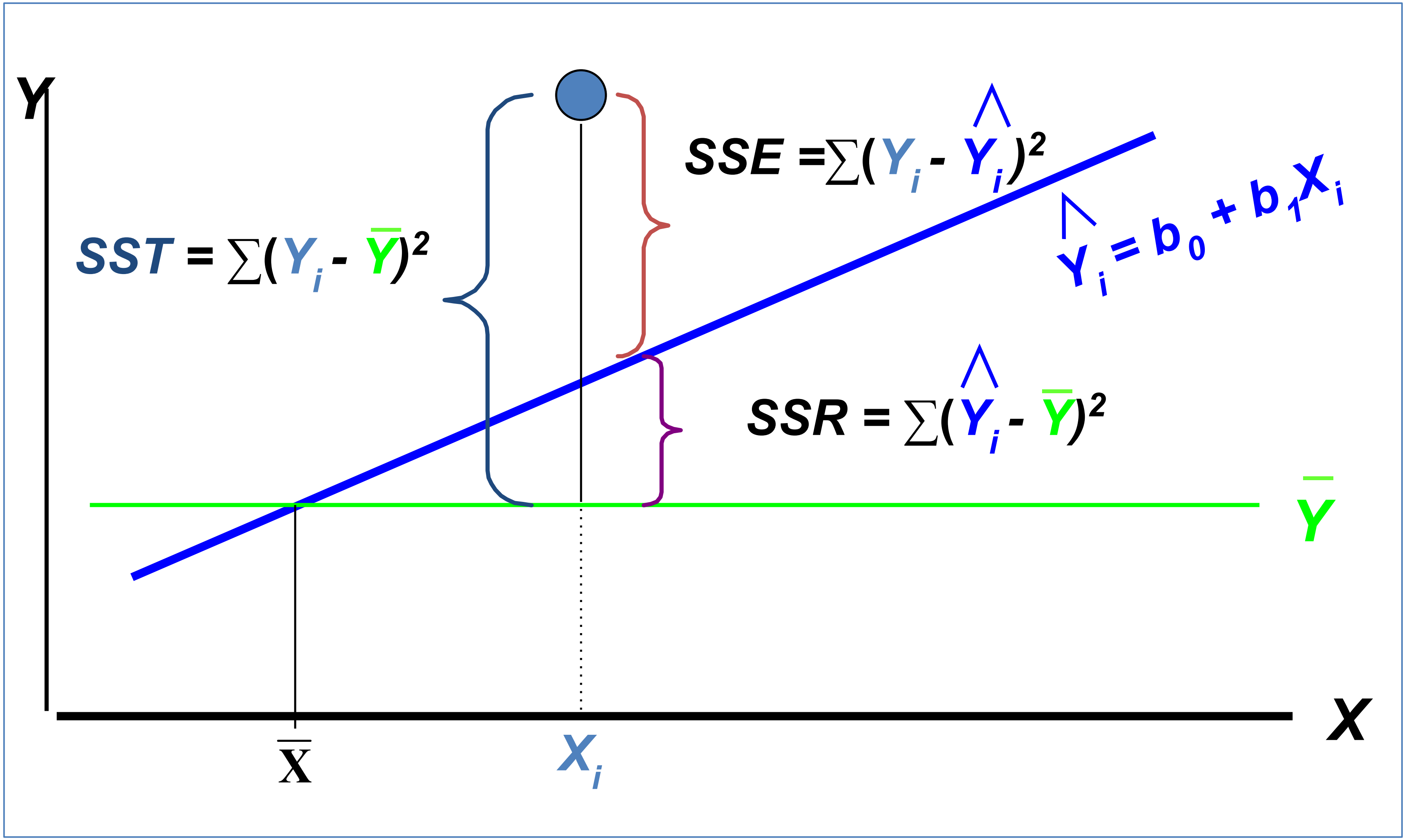
http://www.ats.ucla.edu/stat/AnnotatedOutput/

Regression, Different DVs

- Normally distributed DV [Regression]
- Binary outcome – [Logistic Regression]
  – Like Titanic
- Count model – [Poisson Regression]
  – Number of likes on a Facebook post
- Count model, with lots of 0s [Zero-inflated Poisson Regression]
  – Number of shares on a Facebook post

Regression Intuition: Draw a line that minimizes the error between the predicted and actual value

Measures of Variation: The Sum of Squares



Measures of Variation:
The Sum of Squares

SST = Total Sum of Squares

- measures the variation of the Y_i values around their mean Y

SSR = Regression Sum of Squares

- explained variation attributable to the relationship between X and Y

SSE = Error Sum of Squares

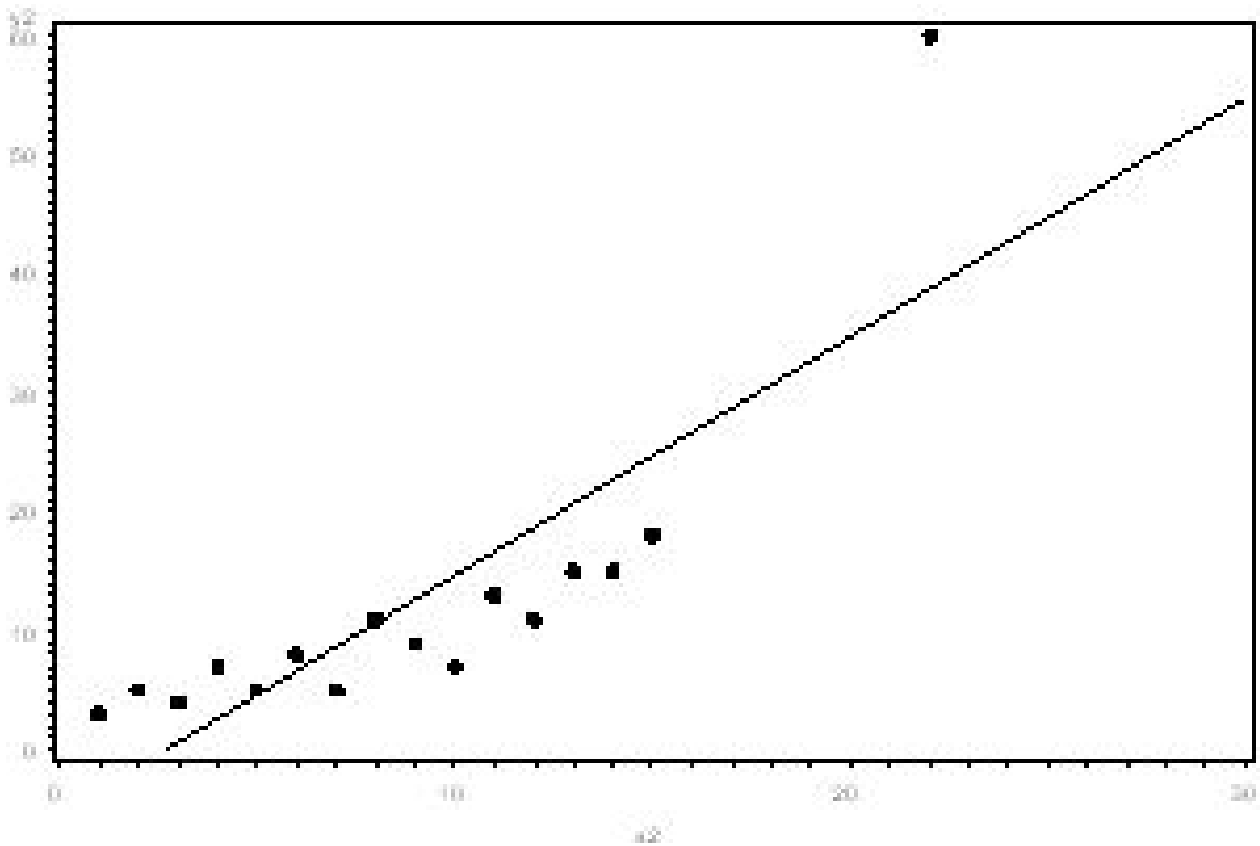
- variation attributable to factors other than the relationship between X and Y

Variance Explained

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Measures the proportion of variation that is explained by the independent variable ***X*** in the regression model

Outliers: Important to Visualize



Adjusted R2

- The use of an adjusted R2 (often written as and pronounced "R bar squared") is an attempt to take account of the phenomenon of the R2 automatically and spuriously increasing when extra explanatory variables are added to the model.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - (1 - R^2)\frac{p}{n-p-1}$$

where p is the total number of regressors in the linear model (not counting the constant term), and n is the sample size.

R2 in R

```
lm(formula = rtotal ~ rpois + rnorm, data = sample1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5031 -0.5964  0.0063  0.6415  3.3308 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.93898    0.14467   13.40   <2e-16 ***
rpois        0.60301    0.01898   31.78   <2e-16 ***
rnorm        0.71118    0.02261   31.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9525 on 497 degrees of freedom
Multiple R-squared:  0.8575, Adjusted R-squared:  0.8569 
F-statistic:  1495 on 2 and 497 DF,  p-value: < 2.2e-16
```

Statistical Testing

- “Null Hypothesis” of no effect of independent variable on dependent variable
- If the p-value is less than the required significance level (equivalently, if the observed test statistic is in the critical region), then we say the null hypothesis is rejected at the given level of significance. Rejection of the null hypothesis is a conclusion.

Statistical Testing

- $p<0.10$ “Marginally Significant”
- $P=0.05$ “Significant”
- $P<0.01/001$ “Significant, low probability of type I error”

Statistical Testing

```
lm(formula = rtotal ~ rpois + rnorm, data = sample1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5031 -0.5964  0.0063  0.6415  3.3308 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.93898    0.14467   13.40   <2e-16 ***
rpois        0.60301    0.01898   31.78   <2e-16 ***
rnorm        0.71118    0.02261   31.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9525 on 497 degrees of freedom
Multiple R-squared:  0.8575, Adjusted R-squared:  0.8569 
F-statistic:  1495 on 2 and 497 DF,  p-value: < 2.2e-16
```

This is really small p values, indicating a very small chance of type I error

β Coefficient

- Beta coefficients indicate how the change in independent variables influence the change in dependent variables
- Very difficult to interpret by themselves, dependent upon scale, relationship, significance

Factor Variables

- Factor type variables have N levels and will report N-1 beta coefficients
- Coefficients are always relative to the omitted level

- EXAMPLE: 2 Levels: Vocational/General/Academic

```
progacademic          16.749      4.943   3.388 0.000854 ***
progvocational       -15.110      5.622  -2.688 0.007828 **
---
```

Statistical Testing

```
lm(formula = rtotal ~ rpois + rnorm, data = sample1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5031 -0.5964  0.0063  0.6415  3.3308
```

How much will change in independent variable change DV

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.93898    0.14467   13.40   <2e-16 ***
rpois        0.60301    0.01898   31.78   <2e-16 ***
rnorm        0.71118    0.02261   31.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9525 on 497 degrees of freedom
Multiple R-squared:  0.8575, Adjusted R-squared:  0.8569
F-statistic:  1495 on 2 and 497 DF,  p-value: < 2.2e-16
```

Interaction Effect Regression

Thus, for a response Y and two variables x_1 and x_2 an *additive* model would be:

$$Y = c + ax_1 + bx_2 + \text{error}$$

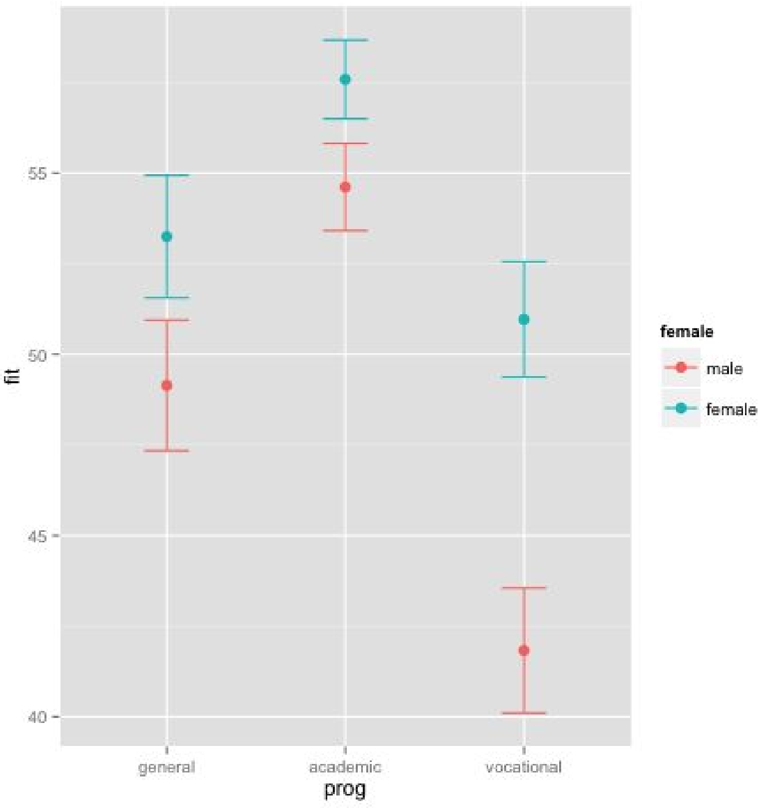
In contrast to this,

$$Y = c + ax_1 + bx_2 + d(x_1 \times x_2) + \text{error},$$

statistics, an interaction may arise when considering the relationship among three or more variables, and describes a situation in which the simultaneous influence of two variables on a third is not additive. Most commonly, interactions are considered in the context of regression analyses.

Interaction Effects

- EXAMPLE
- Prog*gender



Statistical Testing and Big Data

- Statistical Significance (the is a relationship based on rejecting the null hypothesis) and practical significance two different things
- Almost all variables will be significant in many analyses involving large datasets
- Important to understand the marginal effect

- Example: Find the effect of one standard deviation change of independent variable on a dependent variable

Associated Assumptions

- All models have some limitations, but many are very useful
  - Statistical Analysis: Close observance of assumptions is necessary in order to ensure that statistical insights are relevant
  - Predictive Analysis: A model's usefulness in making predictions and can be assessed directly, making underlying assumptions less relevant

Super Learners

- Rather than creating each algorithm separately, we can put a wrapper class around an algorithm to validate it
- http://cran.r-project.org/web/packages/SuperLearner/vignettes/SuperLearnerPresent.pdf

Super Learner

Algorithm	Description	Package
`glm`	linear model	**stats**
`randomForest`	random Forest	**randomForest**
`bagging`	bootstrap aggregation of trees	**ipred**
`gam`	generalized additive models	**gam**
`gbm`	gradient boosting	**gbm**
`nnet`	neural network	**nnet**
`polymars`	polynomial spline regr.	**polspline**
`bart`	Bayesian additive regr. trees	**BayesTree**
`glmnet`	elastic net	**glmnet**
`svm`	support vector machine	**e1071**
`bayesglm`	Bayesian glm	**arm**
`step`	stepwise glm	**stats**