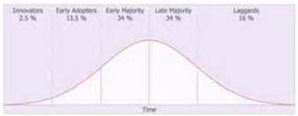
Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

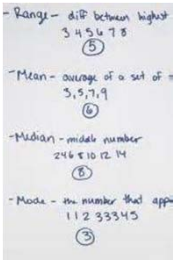
ROCK YOUR CODE TOUR · 2017 | CHICAGO

Introduction to Machine Learning with R

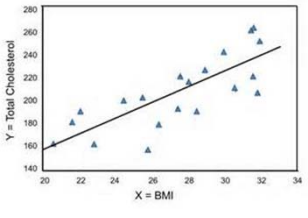
Raj Krishnan
Cloud Solution Architect
Microsoft Corporation

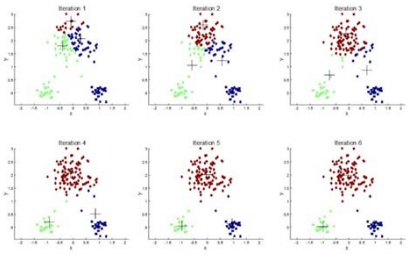
Session Objective

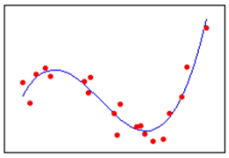




- Overview of Machine Learning
- Review relevant Statistical Concepts & Algorithms
  - Linear Regression
  - Logistic Regression
  - CART
  - Text Analytics
- Using R to solve ML problems
- R on the cloud







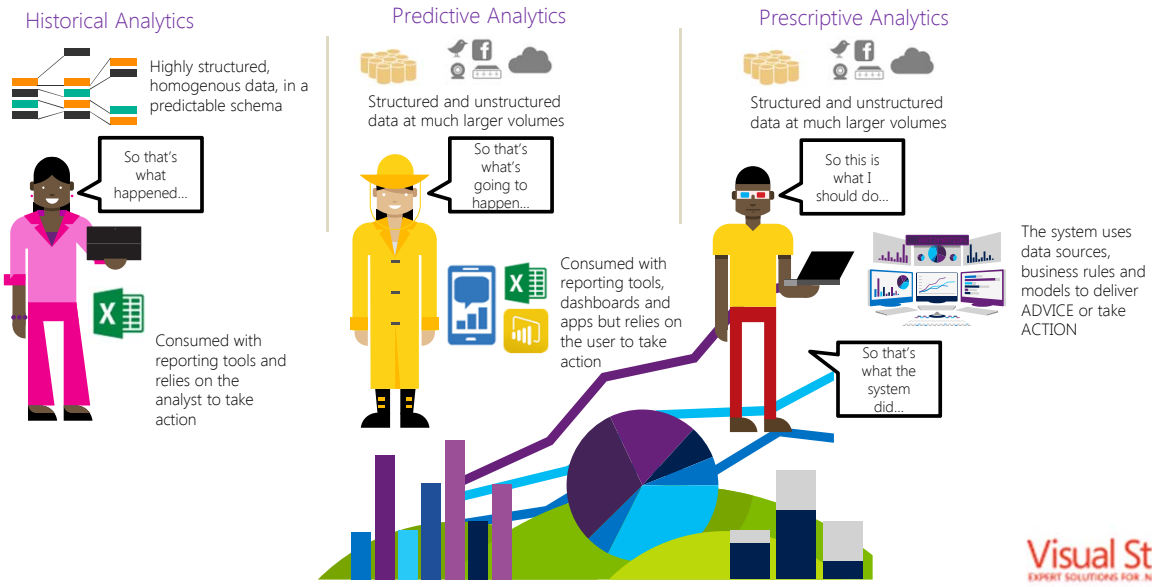
Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

What Are Data Science and Machine Learning?

- **Data science** is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. - *Chikio Hayashi*
- **Machine learning** is a “field of study that gives computers the ability to learn without being explicitly programmed” - *Arthur Samuel*
- We use data science and machine learning to enable computers to help us understand or make predictions about the world.

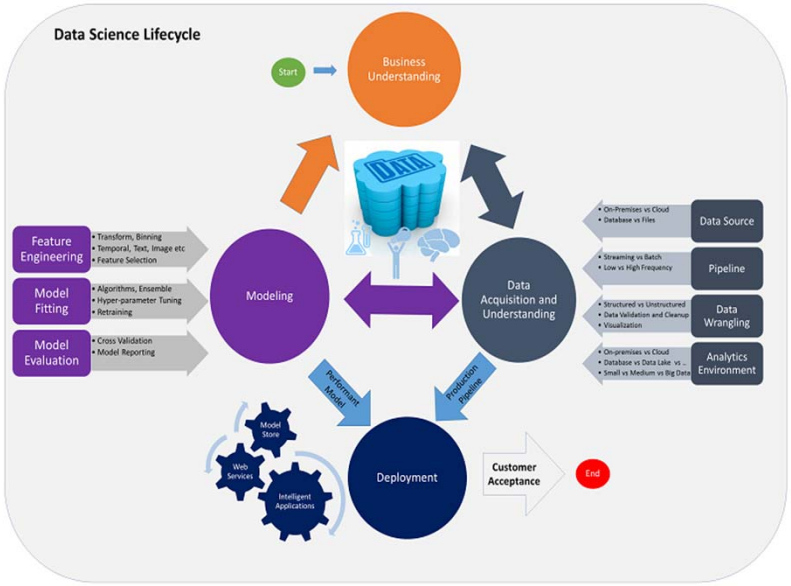
Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Data Science Goes Beyond BI



Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

A More Granular View



Microsoft Team Data Science Process

Visual Studio LIVE! EXPERT SOLUTIONS FOR .NET DEVELOPERS

Build Business Understanding

Begin with the exploration of a problem

Arrive at an analytical question that addresses the problem

Understand the data required to address the problem

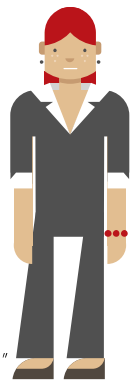
Define the consumption experience

"We spend a lot of money to acquire and retain customers. Turning over a customer, or reacquiring them is also expensive."

"It would be great if we could predict whether a customer is likely to leave us next month and target them for retention."

"We have historical data representing customer acquisition and turnover for the past n years, including all customer interactions."

How will people or systems use these conclusions, on which devices. How will they want to consume the analytic output and what will they do with it?



Define business goals with "sharp" questions that can be answered by Data Science:

- How much or how many? (regression)
- Which category? (classification)
- Which group? (clustering)
- Is this weird? (anomaly detection)
- Which option should be taken? (recommendation)

Visual Studio LIVE! EXPERT SOLUTIONS FOR .NET DEVELOPERS

The Process

- ❑ Understanding the Problem Domain
- ❑ Collecting Data
- ❑ Experimenting with Models
- ❑ Score and Evaluate Model
- ❑ Pick the best model
- ❑ Predict using the Model

A statistics refresher

- Data - facts, observations, and information that come from investigations.
- Measurement
  - quantitative data -- e.g., test score, weight
  - Categorical data also referred to as frequency or qualitative data; grouped according to some common properties)and the number of members of the group are recorded (e.g., males/females, vehicle type)

A statistics refresher

- Variable - property of an object or event that can take on different values
  - Discrete Variable - a variable with a limited number of values (e.g., gender (male/female))
  - Continuous Variable - a variable that can take on many different values, in theory, any value between the lowest and highest points on the measurement scale
  - Independent Variable - a variable that is manipulated, measured, or selected by the researcher as an antecedent condition to an observed behavior; the independent variable is the cause and the dependent variable is the outcome or effect.
  - Dependent Variable - a variable that is not under the experimenter's control -- the data. It is the variable that is observed and measured in response to the independent variable.
  - Qualitative Variable - a variable based on categorical data.

A statistics refresher

Graphs - visual display of data used to present frequency distributions so that the shape of the distribution can easily be seen.

- Bar graph – Discrete data visualization; higher the bar, higher the frequency
- Histogram - a form of a bar graph used with interval or ratio-scaled data
- Boxplot - a graphical representation of dispersions and extreme scores; box with "whiskers."
- Scatterplot - most useful techniques for gaining insight into the relationship between two variables.

A statistics refresher

- Standard deviation
  - Special kind of average - measure of how much each of the scores in the sample *differs* from the sample mean

$$S = \sqrt{\frac{\sum (x-m)^2}{n}}$$

- Confidence Interval
  - The latest poll showed 56 percent favored Clinton while 39 percent would vote for Jeb Bush. Telephone poll of 1,014 adults was conducted March 8-10 and had a margin of error of plus or minus ***3.5 percentage*** points.

Why Use Model

The end goal?

Predict the outcome based on the observations and historical data

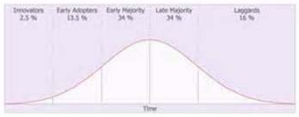
Use Statistical model to train / test data for use with real world data

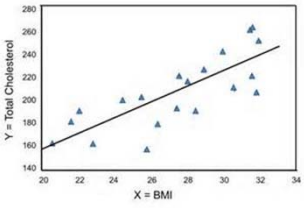
Linear Regression, Logistic Regression, CART etc. are models available for prediction

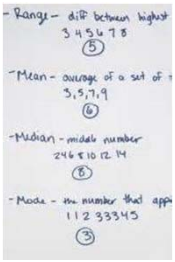
Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

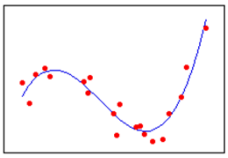
Refresh your statistics

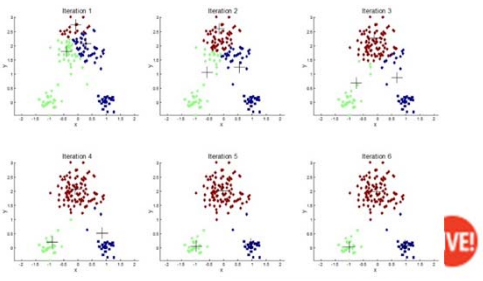
- Mean, Median & Mode
- Linear Regression
- Logistic Regression
- CART
- Text Analytics











Mean, Median & Mode

Mean – Average Value

Median – The Middle value

Mode – The value that appears the most

(median better when outliers, mean representative with large number of variables with no outliers)

variance

$\sigma^2 = \Sigma (X_i - \mu)^2 / N$

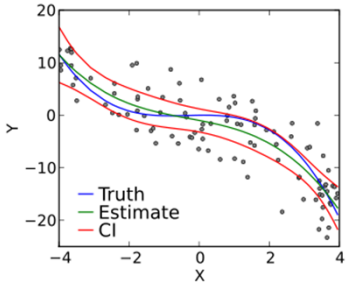
Standard deviation

$\sigma = \text{sqrt} [\sigma^2] = \text{sqrt} [\Sigma (X_i - \mu)^2 / N]$

Linear Regression

A linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of regressors x_i is linear.

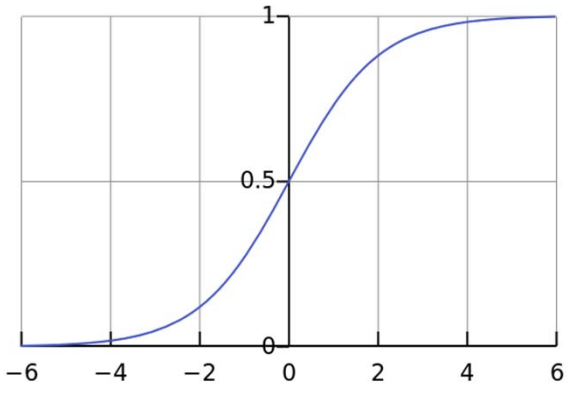
The model remains linear as long as it is linear in the parameter vector $\boldsymbol{\beta}$.



Logistic Regression

- Generalized linear models (GLMs)
- A framework for modeling a response variable *y* that is bounded or discrete
- logistic regression predicts the **probability** of the instance being positive.
  - Binomial or binary logistic regression observed outcome for a dependent variable can have only two possible types (for example, "dead" vs. "alive" or "win" vs. "loss").
  - Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., "disease A" vs. "disease B" vs. "disease C").

Logistic Regression



Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

CART - Classification and Regression Tree

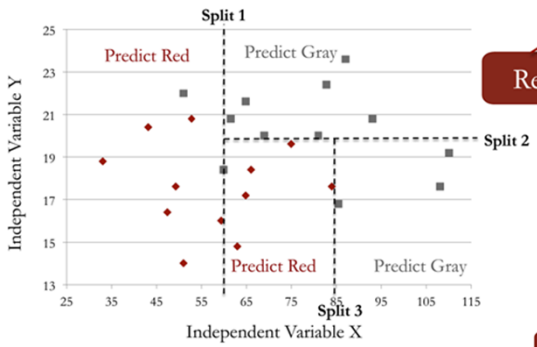
Build a tree by splitting on variables

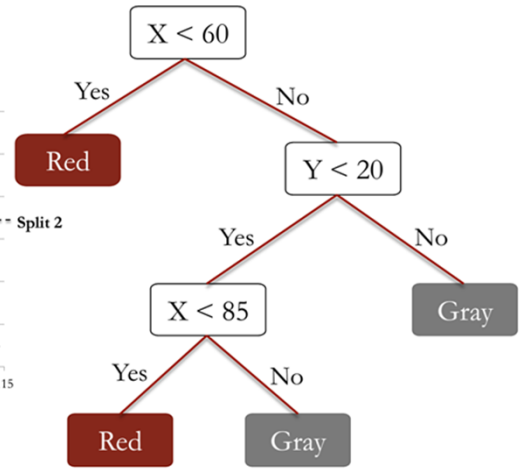
To predict the outcome for an observation, follow

the splits and at the end, predict the most frequent outcome

Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

CART





Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Text Analytics

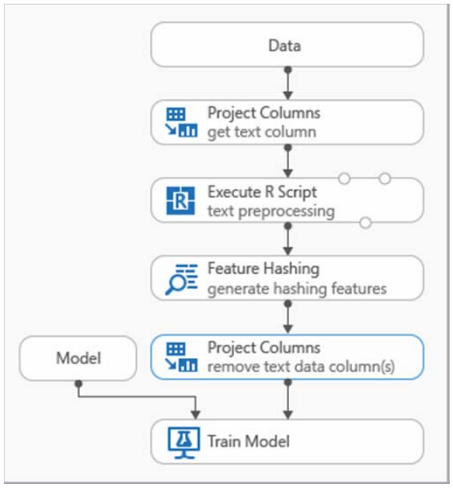
Preparing the text data (R package – tm)

Read the text – build the corpus

Clean the text – Remove Punctuation, remove stop words, remove stem) -> Ready for ML

Document Term Matrix - the frequency of terms that occur in a collection of documents

	I	like	hate	Donuts
D1	1	1	0	1
D2	1	0	1	1



Text Analytics

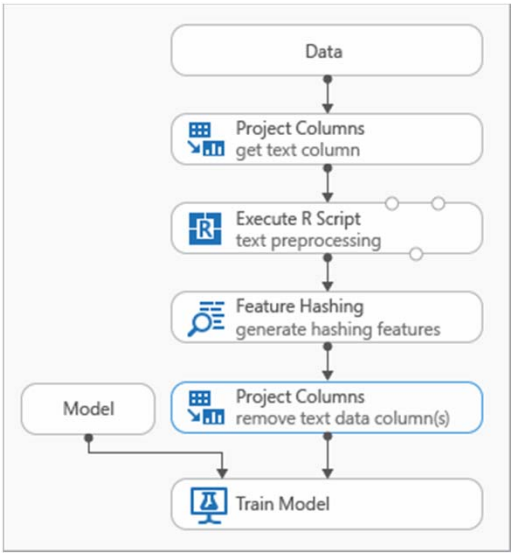
Remove sparse terms (at least appears in more than x% document)

(Feature Hashing)

Train / Test

Use the model – CART, Vowpal Wabbit model

Train and predict



What is R

- A software package for data analysis, statistical computing and visualization with its own language
- Open Source
- Over 2 million users
- Official page: http://www.r-project.org
- Download page: http://www.cran.r-project.org
  • Some helpful websites: • http://www.statmethods.net , www.rseek.org , http://www.ats.ucla.edu/stat/r/ , http://finzi.psych.upenn.edu/search.html, http://www.r-tutor.com/elementary-statistics/numerical-measures/quartile

Getting Ready for R

- ❑ Installing R – The R Studio
  - ❑ Optionally Revolution R-Enterprise 7.3.0
- ❑ R – Basic Commands
  - ❑ Reading Data – Working Directory;
  - ❑ Using R to get statistical information of the data
  - ❑ Installing Packages
- ❑ Vectors and Data frames
- ❑ Loading Data files

Basic R Exercise

```
#Create a data set
#Vectors
X = c(1,2,3,4,5,6)
y<-c(1,2,3,4,5,6,7)
# numeric vector
a <- c(1,2,5.3,6,-2,4)
# character vector
b <- c("one","two","three")
#logical vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE)
#Combining data
firstnames =c("Raj","Joe","Marc")
fullname=paste(firstnames,"Krishnan")
fullname
age=c("56","45","45")
nameage =cbind(fullname,age)
nameage
newentry=c("Fran Krishnan","32")
nameage1=rbind(nameage,newentry)
nameage1
```

Basic R Exercise

```
#Data Frame
d <- c(1,2,3,4)
e <- c("red", "white", "red", NA)
f <- c(TRUE,TRUE,TRUE,FALSE)
mydata <- data.frame(d,e,f)
# variable names
names(mydata) <- c("ID","Color","Passed")
Mydata$Color
#Categorical variable / Factor variables
#Factor variables are categorical variables that can be either numeric / string variables;
#Levels - sorted list of all the distinct values of the data vector
#Categorical variables to factor variables
#Whenever you use random number generation, using the same seed will yield
set.seed(124) the same results (e.g. Separating training and Testing data set; generating sample
data)
schtyp <- sample(0:1, 20, replace = TRUE)
Schtyp
schtyp.f <- factor(schtyp, labels = c("private", "public"))
schtyp.f
is.factor(schtyp.f)
```

Sample Data Set in R

```
#Built in data frame
mtcars
#How many rows are there in the dataset?
mtcars$mpg
#What is the difference between the following commands?
mtcars$mpg[8]
mtcars[8,]
mtcars["Merc 240D",]

#What does the following command do? Explain the output
mtcars[which.max(mtcars$mpg),]

#Plotting data
attach(mtcars)
 plot(wt, mpg, main="Scatterplot Example",
    xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
```

Optional Data exploration in R

```
#Load Crime Data mvtWeek1.csv to mvt
mvt=read.csv(mvtWeek1)
DateConvert = as.Date(strptime(mvt$Date, "%m/%d/%y %H:%M"))
mvt$Date = DateConvert
mvt$Month = months(DateConvert)
mvt$Weekday = weekdays(DateConvert)

#In which month did the fewest motor vehicle thefts occur?
sort(table(mvt$Month))

#hist(mvt$Date, breaks=100)
boxplot(mvt$Date ~ mvt$Arrest)


#What is the proportion of arrests made in year 2001
table(mvt$Arrest, mvt$Year=="2001")
```

The Money Ball Story

Read NBA Statistics data

Use R to build regression model

Oakland A

On-Base Percentage (• Percentage of time a player gets on base -including walks) was the most important

Slugging Percentage (How far a player gets around the bases on his turn –measures power) was important

Batting Average was overvalued

Money Ball

```
# Read in data
baseball = read.csv("baseball.csv");
str(baseball)
# Subset to only include moneyball years
moneyball = subset(baseball, Year < 2002);
str(moneyball)
# Compute Run Difference
moneyball$RD = moneyball$RS - moneyball$RA;
str(moneyball)
# Scatterplot to check for linear relationship
plot(baseball$W,baseball$Team,col=ifelse(baseball$Playoffs==1,"red","blue"));
plot(moneyball$RD, moneyball$W);
# Wins > 95 to make it to playoffs
```

Money Ball

```
# Regression model to predict wins
WinsReg = lm(W ~ RD, data=moneyball);
summary(WinsReg);
#Wins= 80.881375 + .105766 *RD >= 95;RD >133.4
# Regression model to predict runs scored
RunsReg = lm(RS ~ OBP + SLG + BA, data=moneyball);
summary(RunsReg);
RunsReg = lm(RS ~ OBP + SLG, data=moneyball);
summary(RunsReg) (BA is not very useful or makes no
difference)
```

Money Ball

```
# Regression model to predict runs scored
Use 2001 OBP and SLG  to predict 2002 Runs Scored
And OpponentOnBasePercentage and OpponetSlug.
Now caclculate the RD and use Wins predicted.
If it is > 135, they will go to playoff
RS = -804.63 + 2737.77(OBP) +1584.91(SLG)
RA = -837.38 +2913.6(OOBP) + 1514.29(OSLG)
Wins = 80.8814 + 0.1058(RS-RA)

2001 to predict 2002
OBP = .339, SLG=.430 RS=805
OOBP=.307 and OSLG=.373 RA= 622
RS=805, RA=622 wins=100
           R          Paul        Actual
RS         805        800-820  800
RA         622        650-670  653
W          100      93-97   103
```

Leveraging the Cloud

DEMO OF RUNNING AND CONSUMING R SERVICES IN AZURE