Applied Data Science with Python

Course goal

- Train students to be full-fledged data science **practitioners** who could execute **end-to-end** data science projects to achieve **business results**

What you achieve by taking this course

- Understand the concepts and life cycle of Data Science
- Develop proficiency to use Python for all stages of analytics
- Learn Data Engineering tools and techniques
- Acquire knowledge of different machine learning techniques and know when and how to use them.
- Become a full-fledged Data Science Practitioner who can immediately contribute to real-life Data Science projects

Theory vs Practice

- Data Science principles, tools and techniques emerge from different science and engineering disciplines
- Theoretical study focuses on scientific foundations and reasoning
  - Gets into equations, formulae, derivations, reasoning etc.
- Practice (applied) on the other hand focuses on how to apply these principles, tools and techniques to business problems.
  - Focus on purpose, usage, advantages with adequate understanding of concepts
  - Available tools and libraries
- This course is focused on practice

Inclination

- Data Science is trans-disciplinary subject and complex. Mainly it covers three technical areas
  - Math and Statistics
  - Machine Learning foundations
  - Programming
- The course is oriented towards existing software professionals
  - Heavily focused on programming and solution building
  - Limited, as-required exposure to math and statistics
  - Overview of ML concepts, with focus on using existing tools to develop solutions
- Keeping things simple and easy to understand

Course Structure

- Concepts of Data Science
- Data Science Life Cycle
- Statistics for Data Science
- Data Engineering
- Modeling and Predictive Analytics
  - Use cases
- Advanced Topics
- Resource Bundle

Guidelines to students

- Data Science is a complex subject. Needs significant efforts to understand it.
  - Review and re-review videos and exercises
  - Seek out other help – books, online documentations, support forums
- If you have queries, doubts or concerns, please send a private message or post a discussion question
  - We would be happy to address them as soon as possible
- We are constantly improving our courses so all feedback is welcome
  - Feedback through private messages / emails.
- At the end of the course, if you like it, please leave a review
- Expect maximum discounts for future courses

Relationship with other V2 Maestros courses

- Our courses are focused on Data Science related topics
  - Technologies
  - Processes
  - Tools and Techniques
- We focus on making our courses self sufficient
- If you are an existing V2 Maestros student, you will see some content and examples repeated across courses

We hope this course helps you to advance your career.
Best of luck !

V2 Maestros
The Data Science Experts

What is Data Science

Understanding the domain

V2 Maestros
The Data Science Experts

Definitions

Across the web

Data Science

- Skill of extracting of knowledge from data
- Using knowledge to predict the unknown
- Improve business outcomes with the power of data
- Employ techniques and theories drawn from broad areas of mathematics, statistics and information technology

Data Scientist

- A practitioner of data science
- Expertise in data engineering, analytics, statistics and business domain
- Investigate complex business problems and use data to provide solutions

V2 Maestros
The Data Science Experts

Data

The foundation of Data Science

Entity

- A thing that exists about which we research and predict in data science.
- Entity has a business context.
- Customer of a business
- Patient at a hospital. The same person can be a patient and a customer, but the business context is different.
- Car. Entities can be non living things

Characteristics

- Every entity has a set of characteristics. These are unique properties
- Properties too have a business context
- Customer : Age, income group, gender, education
- Patient: Age, Blood Pressure, Weight, Family history.
- Car: Make, Model, Year, Engine, VIN

Environment

- Environment points to the eco-system in which the entity exists or functions.
- Environment is shared among entities. Multiple entities belong to the same environment
- Environment affects an entity's behavior
- Customer : Country, City, Work Place
- Patient: City, Climate .
- Car: Use (City/highway), Climate

Event

- A significant business activity in which an entity participates.
- Events happen in a said environment.
- Customer : Browsing, store visit, sales call
- Patient: Doctor visit, blood test
- Car: Smog test, comparison test

Behavior

- What an entity does during an event.
- Entities may have different behaviors in different environments
- Customer : Phone Call vs email, Clickstream, response to offers
- Patient: Nausea, light-headed, cramps
- Car: Skid, acceleration, stopping distances

Outcome

- The result of an activity deemed significant by the business.
- Outcome values can be
  - Boolean (Yes/No, Pass/Fail)
  - Continuous (a numeric value)
  - Class (identification of type)
- Customer : Sale (Boolean), sale value (continuous)
- Patient: Blood Pressure value (continuous). Diabetes type (class)
- Car: Smog levels (class), stopping distances (continuous), smog passed (Boolean), car type (class)

Observation

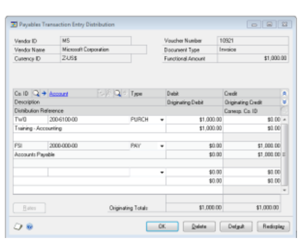
- A measurement of an event deemed significant by the business.
- Captures information about
  - Entities involved
  - Characteristics of the entities
  - Behavior
  - Environment in which the behavior happens
  - outcomes
- An observation is also called a system of record
- Customer : A phone call record, a buying transaction, an email offer
- Patient: A doctor visit record, a test result, a data capture from a monitoring device
- Car: Service record, smog test result

Dataset

- A collection of observations
- Each observation is typically called a record
- Each record has a set of <u>attributes</u> that point to characteristics, behavior or outcomes.
- A dataset can be
  - Structured (database records, spreadsheet)
  - Unstructured (twitter feeds, news paper articles)
  - Semi-structured (email)
- Data scientists collect and work on datasets to learn about entities and predict their future behavior/ outcomes.

Structured Data

- Attributes are labeled and distinctly visible.
- Easily searchable and query able.
- Stored easily in tables



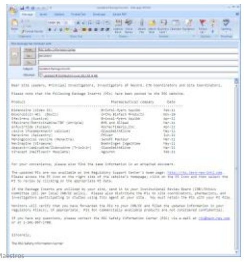
Unstructured Data

- Data is continuous text
- Attributes are not distinctly labeled. They are present within the data.
- Querying is not easy.

The Mazda3 is on a very short list of compact cars that are available as a hatchback or a sedan. It also comes with two 6-speed transmissions -- manual or automatic -- and choice of two 4-cylinder engines -- a 155-horsepower 2.0-liter or a 184-horsepower 2.5-liter -- and all of those variations are available with either body style. Its best fuel economy is an EPA-rated 41 mpg on the highway, which is near the top of the class for gasoline-powered cars (tying the Honda Civic, yet another trait they share). That rating applies to the 2.0-liter engine, whether it's backed by a manual or automatic transmission.

Semi-structured Data

- Mix of structured and unstructured.
- Some attributes are distinctly labeled. Others are hidden within free text



Summary

- Entity
- Characteristics
- Environment
- Event
- Behavior
- Outcomes
- Observation
- Dataset

Learning

Discovering knowledge from Data

Relationships

- Attributes in a dataset exhibit relationships
- Relationships “model” the real world and have a logical “explanation”
- For attributes A and B the relationships can be
  - When A occurs, B also occurs
  - When A occurs B does not occur
  - When A increases B also increases
  - When A increases B decreases
- Relationships can involve multiple attributes too
  - When A is present and B increases, C will decrease

Relationships - Examples

- Customer
  - As age goes up, spending capacity goes up. (AGE and REVENUE)
  - Urban customers buy more internet bandwidth (LOCATION and BANDWIDTH)
- Patient
  - Older patients have more prevalence of Diabetes (AGE and DISEASE LEVEL)
  - Overweight patients typically have higher cholesterol levels (WEIGHT and HDL)
- Car
  - The more cylinders a car has, the mileage tends to be lower (CYLINDERS and MILEAGE)
  - Sports Cars have more insurance rates (TYPE and RATES)

Relationships

- Consistent vs Incidental Patterns in Data
- Correlations
- Signals and noise

V2 Maestros
The Data Science Experts

What is Learning

- Learning implies learning about relationships.
- It involves
  - Taking a domain
  - Understanding the attributes that represent the domain
  - Collecting data
  - Understanding relationships between the attributes
- Model is the outcome of learning

V2 Maestros
The Data Science Experts

Model

- A simplified, approximated representation of a real world phenomenon
- Captures key attributes and their relationships
- Mathematical model – represents relationships as an equation
- Blood Pressure

  BP = 56 + (AGE * .8) + (WEIGHT * .14) + (LDL * .009)
- Decision Tree model – represents the outcome as a decision tree
- Buying a music CD

  If AGE < 25 and GENDER=MALE, buy BEYONCE-CD = YES
- Accuracy of models depends on strength of relationships between attributes

V2 Maestros
The Data Science Experts

Prediction

- A model can be used to predict unknown attributes

  BP = 56 + (AGE * .8) + (WEIGHT * .14) + (LDL * .009)
- The above model represents the relationships between BP, AGE, WEIGHT and LDL.
- If 3 of the 4 attributes are known, the model can be used to predict the 4^{th}.
- The above equation can be considered the prediction algorithm
- Relationships can be a lot more complex, leading to complex models and prediction algorithms.

V2 Maestros
The Data Science Experts

Predictors and outcomes

- Outcomes are attributes that you want to predict
- Predictors are attributes that are used to predict outcomes.
- Learning is all about building models that can be used to predict outcomes (outputs) using the predictors (inputs)

Example	Predictors	Outcomes
Customer	Age, Income Range, Location	Buy? Yes/No
Patient	Age, Blood Pressure, Weight	Diabetic?
Car	Cylinders, acceleration	Sports vs family

V2 Maestros
The Data Science Experts

Humans vs machines

- Humans understand relationships and predict all the time.
- Build humans can only handle finite amount of data
  - One shop keeper can know preferences of 100 customers, not 10 million of them
- Machines (computers) come into play when the number of entities and data about them are large
- There in comes machine learning, predictive analytics and data science

V2 Maestros
The Data Science Experts

So what is Data Science ?

- Picking a problem in a specified domain
- Understanding the problem domain (entities and attributes)
- Collect datasets that represent the entities
- Discover relationships (Learning)
  - When computers are used for this purpose, its called machine learning.
- Build models that represent relationships
  - Uses past data where all predictors and outcomes are known
- Use models for predicting outcomes
  - Current/ future data – predictors known, outcomes unknown

Data Science Example – Website Shopper

- Problem : Predict if the shopper will buy a smartphone
- Data: Past purchase history of shoppers
  - Shopper characteristics (age, gender, income etc.)
  - Seasonal information
  - Others..
- Build Model
  - Decision model based on shopper and seasonal entities
  - Built every week
- Prediction
  - When a new shopper is browsing, predict if the shopper will buy
- Action
  - Offer Chat help

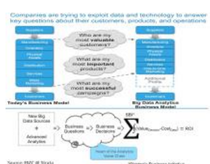
V2 Maestros
The Data Science Experts

Data Science Use Cases

How the world is benefiting

Data Science applications

- The use of data science is growing exponentially into and across multiple domains in business, science, finance and personal life
- Recent advances in computing power, open source software and predictive algorithms have made it cost effective to apply data science for commercial use



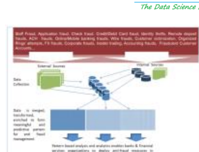
V2 Maestros
The Data Science Experts

Finance

Making money and saving money

Fraud Detection

- Credit Card Frauds exhibit patterns in transactions
- Historical Transaction Data used to identify fraudulent patterns and build models
- Each new transaction is then given a fraud score based on the model
- Action taken for high scored transactions



V2 Maestros
The Data Science Experts

Retailing

Sell more

Recommendations

- Items exhibit patterns on how they are brought together
  - Cell phones and accessories
  - Books
- Patterns used to build affinity scores between items
- When one item is brought, items with high affinity scores to that item are recommended.

Contact centers

Improving efficiency

Scoring of Callers and Agents

- Past interactions used to score callers based on their value or type
- Agents are scored based on their ability to sell or handle a specific type of problem.
- The right callers and then matched with the right agents to optimize business outcomes.
- Call recordings analyzed using machine learning to grade quality of call and outcome

Health Care

Preventive Care

Predicting Disease Outbreaks

- Dataset collected from public domains like google searches, twitter feeds etc.
- Data linked with location information and disease patterns to build outbreak forecasting models
- Model used to track potential outbreaks and take preventive actions

Data Science Life Cycle

Activities and Sequencing

Introduction

- Data Science efforts are typically executed as projects
- They are typically research projects, not build-and-operate projects
- Projects typically have a start and end state
- Projects have phases and activities
- Transitions happen between phases / activities
- In this section we talk about
  - What Data Science project phases and their activities are
  - Importance of each activity
  - How they transition between one another
  - Some of the best practices too.

Data Science Project Phases / Activities

Setup → **Data Engg.** → **Analytics** → **Production**

Setup
- Goal Setting
- Understanding the domain
- Understanding Data

Data Engg.
- Acquisition
- Cleansing
- Transformation
- Enrichment
- Persistence

Analytics
- Exploratory
- Inferential
- Modeling
- Predictions
- Recommendations

Production
- Building Data products
- Operationalized Data Feeds
- Continuous Improvement

Setup

Setup to Succeed

Goal Setting

- Every Data Science Project will / should have a goal.
- The team will focus their activities based on this goal
- Projects without goals are cars without a driver.
- Examples
  - Predict which customer(s) will churn in the next 3 months
  - Group tweets about the company based on sentiment
  - Identify patients with a possibility of having heart attack in the next 3 months

Understanding the Problem Domain

- A data science team should have solid understanding of the problem domain
  - Business basics (finance, CRM, medical)
  - Business processes and workflows
  - Key Performance metrics
- Machines only know numbers and strings, they needs humans to associate meaning to them.
- Knowledge of the domain helps the team to understand entities, relationships and patterns.
- It helps validate assumptions, identify errors and analyze if predictions will work.

Data

- Business Processes and workflows generate Data
  - Application Data Entries
  - Reports and Visualizations
  - Sensor Data feeds
  - Web clicks in a browser
  - Point-of-Sale Transactions
  - Social media feeds
- Data can be structured, unstructured or semi-structured
- Data have different origins, stored in different silos and have logical relationships

Understanding Data

- Source of the Data
- Processing /transformation steps performed
- Storage (enterprise databases, cloud, news feeds)
- Synchronization
- Relationships
- Ordering
- Understanding data helps the team identify possible sources of predictive patterns.

Data Engineering

Get the data to the form you need it.

Data Acquisition

- Acquire Data from different data sources
  - Enterprise Databases (Oracle, MySQL)
  - Cloud APIs (Sales Force)
  - Scanner / sensor feeds (Bar code scanners)
  - Social media downloads (Twitter, Facebook)
- Real time/ interval and bulk
- Sanity checking
- Most cumbersome and time-consuming to setup.
- Establishing connections to machines and humans involved can be frustrating ☹

Data Cleansing

- Data have different degrees on cleanliness and completeness
- Structured data from corporate applications are usually clean and complete
- Data from internet, social media or voice transcripts need significant cleansing.
- Handling missing data is a key decision
- Cleansing examples
  - Normalize date formats : MM/DD and DD/MM
  - Standardize decimal places
  - First Name Last Name vs Last Name, First Name

Data Transformation

- Data is transformed to extract required information while discarding un-necessary baggage
- Processing and summarization to logical activity levels
- Transformation helps cutting down data size and minimizes further processing needs
- Examples
  - Web Clicks summarized by website visit
  - Language translations
  - Medical sensor data summarized by interval

Data Enrichment

- Add additional attributes to data records that improves the quality of information
- Examples
  - Add demographics information from a customer database to the point-of-sale transaction record
  - Logical grouping of patients by past medical history – excellent health, moderate , needs consistent care

Data Persistence

- Processed data is stored in a reliable, retrievable data sink.
- All relevant information captured in a single local record as much as possible
- Example : Retail store transaction : (POS Data + customer demographics + Item characteristics + Sales associate performance)
- Scaling and query performance are important factors will choosing a data sink
  - Flat files
  - Traditional SQL databases
  - Big Data technologies.

Analytics

Learn and Predict

Exploratory Data Analysis

- Understand individual attribute patterns (range, minimum, maximum, frequency, mean etc.)
- Understand relationships between attributes (how does change in one affect another)
- Reasoning (is the behavior explainable?)
- Outliers (odd values)
- Possible errors in processing
- E.g.: Patient weights.
- Validate findings with domain experts.

Inferential analysis

- Look for signals in the data
  - Patterns
  - Correlations
  - Reasoning
- Check if patterns are consistent and reproducible
  - Month after month
  - Different use cases
- Statistical Tests
  - Can results be extrapolated for the entire population?
- Example : Patient weights vs diabetes

Modeling

- Use machine learning algorithms to build models
- Build multiple models based on different algorithms and different datasets
- Test models for accuracy
- Identify best performing models
  - Accuracy
  - Response Time
  - Resources
- As simple as an equation or a decision tree. As complex as a neural network.

Prediction

- Use models built to predict outcomes for new data
- Keep validating model accuracy to make sure accuracy levels are consistent for different variations in data
- Response time and resource usage are critical when predictions need to happen in real time
- Measure improvements made to outcome predictions using the model
- Simulations might be performed to validate prediction benefits

Recommendations

- At the end of the project, recommendations need to be provided to the project owners on the algorithms to use and expected benefits
- A Data science project might have no recommendations to make if the dataset does not exhibit any exploitable patterns
  - Does not mean it's a failure
- Sometimes unexpected patterns are discovered that might lead to other benefits
- A final presentation is made to stake holders.

Iterations

- Based on intermediate or at-the-end analysis and feedback, the analysis phase might be repeated with different objectives
- The project team "responds" to findings in the data, which might lead to multiple analysis paths.

Production

Implement continuous processes

Building Data Products

- Once the modeling and prediction algorithms are "firmed up", data products are built that would use the algorithms for production level modeling and predictions
- Have quality software rigor in development and testing
- Deployed in enterprise or cloud models.

Operationalized Data feeds

- Continuous data feeds into data products
  - Instantaneous
  - Every day
  - Periodic
- Data products perform cleansing, transformation and error reporting
- Pruning of old data might be necessary

Continuous improvement

- Changes in business environment might affect learning and prediction
- The learning and prediction steps need to be re-validated at appropriate intervals to make sure they continue to work as desired.
- Revalidation needs to happen when business processes change.
- Efforts to generate better models should be ongoing.

Summary

- Data Science projects follow a life cycle
- Data Science projects are research type projects – there is a lot of experimentation and sometimes no end result
- Signals in data drives results, not the algorithms
- Multiple iterations might be necessary before reasonable results are achieved.

V2 Maestros
The Data Science Experts

Statistics for Data Science

-

Goals

- Describe basic statistics for Data Science
- Explain the concepts
- Avoid formulae and mathematical representations as much as possible

V2 Maestros
The Data Science Experts

Types of Data

What they are and what you can do with them.

Overview

- There are 4 types of data that you would deal with
- They differ in meaning and what operations you can do on them
- Types
  - Categorical or nominal
  - Ordinal
  - Interval
  - Ratio

Categorical

- Represents categories or types
- Fixed list of values
- No implicit ordering or sequencing
- Examples :
  - Fruits : apples, oranges, grapes
  - Players : defender, mid-fielder, forward
  - Cars : sedan, coupe, SUV

Ordinal

- Represents categories
- But there is ordering among the values
- Represents a scale.
- Comparison possible (greater than, less than)
- Examples
  - Review Rating : Outstanding, Very Good, Good, Fair, Bad
  - Pain Levels: 1 – 10 (10 being the highest)
  - Student Grades : A, B, C, D, F

Interval

- Numeric Data
- Measurement where the difference is meaningful
- Represents time, distance, temperature etc.
- Addition, Subtraction possible
- Multiplication, division not possible
- Examples
  - Time of Day
  - Dates
  - Distance between two points
  - Temperature

Ratio

- Numeric Data
- All arithmetic operations possible
- True Zero possible
- Examples
  - Weight
  - Speed
  - Amount

Comparison

Operations	Nominal	Ordinal	Interval	Ratio
Discrete Values	Yes	Yes	Yes	Yes
Continuous Values	No	No	Yes	Yes
Frequency Distribution	Yes	Yes	Yes	Yes
Median and Percentiles	No	Yes	Yes	Yes
Add / Subtract	No	No	Yes	Yes
Multiply / Divide	No	No	No	Yes
Mean, Std. Deviation	No	No	Yes	Yes
Ratios	No	No	No	Yes
True Zero	No	No	No	Yes

Summary Statistics

Describe data

Overview

- Describe a set of observations
- Observations have a number of data points; Summary statistics are used to characterize them
- Describe
  - Central Tendency
    - Mean, Median, Mode
  - Variation
    - Variance, Standard Deviation
  - Skew
    - Quartiles

Central Tendency

- Mean : The average
  - Add all number and divide by their count
- Median: The middle value
  - Order the numbers and find the middle value
  - If the count is even, find average of the two middle values
- Mode: The most occurring value
  - The value that occurs most
- Usage depends on situation

Central Tendency : Example

- Observations : 1, 3, 4, 5, 5, 7, 8, 9, 9, 9
- Count: 10
- Sum: 60
- Mean : Sum / Count = 60/10 = 6 μ
- Median : Middle Value = (5 + 7) / 2 = 6
- Mode: 9

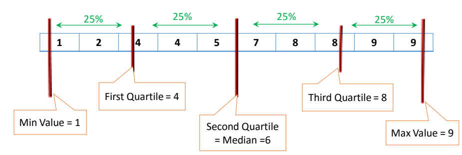
Variance

- Describes how values are distributed around the mean
  - If most values are closer to mean, low variance
  - If significant differences in values, then high variance
- To compute
  - Find the mean
  - Square the differences from the mean
  - Sum of Squares
  - Divide by count
- Standard Deviation is Square Root of variance

Values	Mean – Value	Square
4	0	0
6	-2	4
3	1	1
5	-1	1
2	2	4
Mean = 4		Sum=10
Variance = 2		
σ	Std. Dev = 1.41	

Quartiles

- Describes the central tendency, distribution, range and skew in one set of measures
- Given a set of observations, we divide them into 4 equal sets.
- The boundaries form the quartiles



Reading Quartiles

Min	1st	Median	3rd	Max	Comments
1	3	5	8	10	Evenly distributed
1	4	5	6	10	Most values closer to center
1	2	3	7	10	Skewed to the left
1	6	7	9	10	Skewed to the right

Outliers

- An Odd value occurring in a dataset
- Typically towards the min end or max end of the list
- Outliers tend to distort the summary statistics of a dataset
- Example
  - Observations : 1,2,4,5,20
  - Outlier: 20
  - With outlier, mean= 6.4, Std. Dev=7.76
  - Without outlier, mean= 3, Std. Dev=1.82

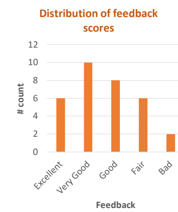
V2 Maestros
The Data Science Experts

Distributions

Summarizing trends

Overview

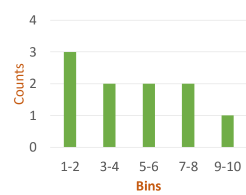
- Distributions show how data values are spread in a given observation set
- Distributions contain a set of bins
- Data is grouped in bins based on
  - Values (categorical, ordinal)
  - Value ranges (interval, ratio)



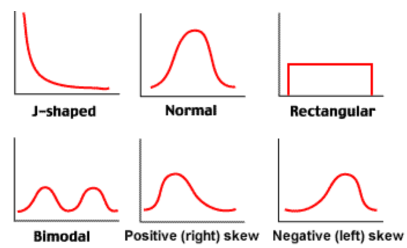
Building a Distribution

4	7	3	2	6	9	8	2	5	2

Bin	Values	Count
1-2	2,2,2	3
3-4	4, 3	2
5-6	6,5	2
7-8	7,8	2
9-10	9	1

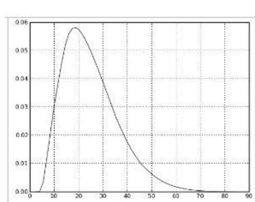


Distribution Shapes



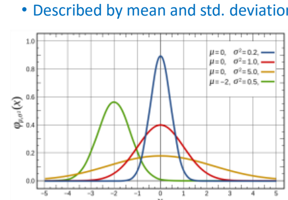
Probability Distributions

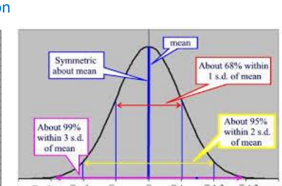
- Assigns a probability to each measurable subset of the possible outcomes of an experiment
- Each possible outcome (or range) plotted on the x-axis
- Probability (0 – 1) plotted on the y-axis
- Discrete or continuous



Normal Distribution (Gaussian)

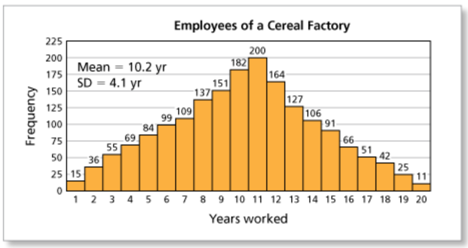
- Very commonly occurring distribution
- Described by mean and std. deviation





V2 Maestros
The Data Science Experts

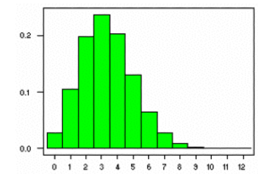
Normal distribution Example



V2 Maestros
The Data Science Experts

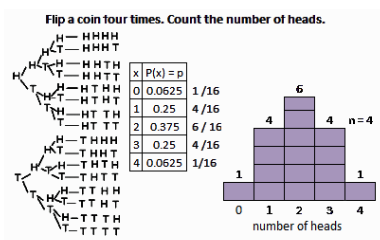
Binomial Distribution

- Describes the probability of a Boolean outcome (Yes/ No)
- If
  - n is the number of trials
  - p is the probability of success
  - k is the number of successes
- Plots the probabilities of all values of k.



V2 Maestros
The Data Science Experts

Binomial Distribution Example



V2 Maestros
The Data Science Experts

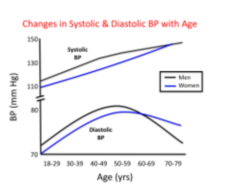
Correlation

Relationships

V2 Maestros
The Data Science Experts

Overview

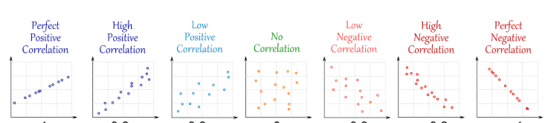
- Correlation : a mutual relationship or connection between two or more things
- Interdependence
- Correlation between 2 sets of data – how much does one change when the other changes
- The basis of data science
- Example : Age and Blood Pressure



V2 Maestros
The Data Science Experts

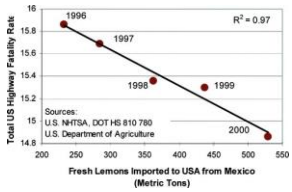
Measuring Correlation

- Pearson's Correlation co-efficient
- Values range from -1 to +1



Correlation and Causation

- Causation : The reason for a change in value
- Correlation does not imply causation
- Correlation might be due to
  - Causation
  - Common cause
  - Incidental
- An analysis needed to establish why you see what you see



Data Engineering

Data Sources

Overview

- Data Sources play a vital role in determining data and process architectures and workflows
  - Data Quality and Reliability
  - Network Planning
  - Fault tolerance
  - Security
  - Organizational boundaries

Enterprise Data Sources

- Typically RDBMS
- Within the organization's boundaries
- Easy accessibility and no rate limits
- Excellent Quality and reliability
- Data Guardians might be insecure of your use.
- Might need to work through organizational bureaucracy to get access

Cloud Data Sources

- Data hosted on the web like Sales Force, Marketo
- Access typically through SOAP or REST APIs
- Security is a pre-dominant factor
- Rate limits might apply
- Quality of Data would be excellent

Social Media Data Sources

- Facebook, twitter, LinkedIn, google
- Similar to Cloud Data Sources in most aspects
- Accessing public data about people and companies might involve privacy issues
- Rate limits are pretty restrictive
- Data mostly profile based then transaction based

Web Scraping

- Scraping web sites is a raw but last way to get data that is otherwise not available
- Data is very dirty and would require a lot of cleaning
- Mostly text and require significant processing resources
- Security, privacy and intellectual property concerns

Data Formats

Overview

- Tables – from RDBMS
- CSV – most common data exchange format
- XML – configuration / meta data
- JSON – new age data exchange format
- Raw Text
- Binary – images, voice streams

Data Acquisition Trends

Acquisition Intervals

- Types
  - Batch
  - Real time (push triggers)
  - Interval (e.g.: every 30 minutes)
  - Hybrid
- Determined by
  - Analytics needs
  - Availability
  - Rate limits
  - Reliability

Data Cleansing

Issues with Data Quality

- Invalid values
- formats
- Attribute dependencies
- Uniqueness
- Referential integrity
- Missing values
- Misspellings
- Misfielded values (value in the wrong field)
- Wrong references

Finding Data Quality Issues

- Sample Visual Inspection
- Automated validation code
  - Schema validation
- Outlier Analysis
- Exploratory Data Analysis

Fixing Data Quality Issues

- Fixing Data Quality issues is regular boilerplate coding in any language the team is comfortable with
- Fix the source if possible
- Find possible loopholes in data processing streams
- Analyze batches coming in and automate fixing
- Libraries and tools available

Data Imputation

- Any "value" present in a dataset is used by machine learning algorithms as valid values – including null, N/A, blank etc.
- This makes populating missing data a key step that might affect prediction results
- Techniques for populating missing data
  - Mean, median, mode
  - Multiple imputation
  - Use regression to find values based on other values
  - hybrid

Data Transformations

Code Examples

- Going forward, most examples will be covered as part of case studies

Overview

- Different sources of data follow different formats and hence standardization is required
- Having data in the same format and scale makes comparison and summarization activities easier

Data Standardization

- Numbers
  - Decimal places
  - Log
- Date and Time
  - Time Zone
  - POSIX
  - Epoch
- Text Data
  - Name formatting (First Last vs Last, First)
  - Lower case/ Upper case/ Init Case

Binning

- Convert numeric to categorical data
- Pre-defined ranges are used to create bins and individual data records are classified based on this.
- New columns typically added to hold the binned data
- Binning is usually done when the continuous variable is used for classification.

Age	Age Range
35	20 - 40
23	20 - 40
11	01 - 20
65	60 - 80
40	40 - 60
51	40 - 60
20	20 - 40

Indicator variables

- Categorical variables are converted into Boolean data by creating indicator variables
- If the variable has n classes, then n-1 new variables are created to indicate the presence or absence of a value
- Each variable has 1/0 values
- The nth value is indicated by a 0 in all the indicator columns
- Indicator variables sometimes work better in predictions that their categorical counterparts

Pressure	Is High?	Is Medium?
High	1	0
Low	0	0
High	1	0
Medium	0	1
Medium	0	1
Low	0	0
High	1	0

Centering and Scaling

- Standardizes the range of values of a variables while maintaining their signal characteristics
- Makes comparison of two variables easier
- The values are “centered” by subtracting them from the mean value
- The values are “scaled” by dividing the above by the Standard Deviation.
- ML algorithms give far better results with standardized values

	Age	Height	Cent. Age	Cent. Height
	35	150	0.00	-1.66
	23	195	-0.74	1.92
	11	161	-1.47	-0.78
	65	165	1.84	-0.47
	40	180	0.31	0.73
	51	169	0.98	-0.15
	20	176	-0.92	0.41
Mean	35.00	170.86		
Std. Dev	16.30	12.56		

Text Pre-Processing

Understanding how ML algorithms work

- ML Algorithms work with
  - numbers (continuous data)
  - classes (discrete/ categorical data)
- ML algorithms don't work with text.
- All textual data need to be converted into numbers or classes
- This is one of the main responsibilities of data pre-processing

Text Cleansing

- Remove punctuation
- Remove white space
- Convert to lower case
- Remove numbers
- Remove stop words
- Stemming
- Remove other commonly used words

TF-IDF Overview

- Text Documents are becoming inputs to ML more and more.
  - News items for classification
  - Email messages for spam detection
  - Text search
- Text need to be converted to equivalent numeric representation before ML can be used
- The most popular technique used is Term Frequency – Inverse Document Frequency (TF-IDF)
- TF-IDF output is table where rows represent documents and columns represent words
- Each cell provides a count / value that indicate the "strength" of the word with respect to the document

TF-IDF formulae

Text Frequency (given a word w1 and Document d1)

= (# of times w1 occurs in d1) / (# of words in d1)

Inverse Document Frequency (given a word w1)

= log e (Total # of docs / Total docs with w1)

TF-IDF = TF * IDF

TF-IDF steps

1. Original documents

Doc 1 = " This is a sampling of good words"
Doc 2 = " He said again and again the same word after word"
Doc 3 = " words can really hurt"

2. After cleansing

Doc 1 = "sample good word"
Doc 2 = "again again same word word"
Doc 3 = " word real hurt"

TF-IDF (contd.)

- Creating the count table

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
Doc 1	1	1	1				
Doc 2			2	2	1		
Doc 3			1			1	1

- Finding Text Frequency

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
Doc 1	.33	.33	.33				
Doc 2			.4	.4	.2		
Doc 3			.33			.33	.33

TF-IDF (contd.)

- Finding Inverse Document Frequency
  - Log e (Total docs / docs with the word)

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
IDF	1.098	1.098	0	1.098	1.098	1.098	1.098

- Finding TF-IDF (TF * IDF)

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
Doc 1	.36	.36	0				
Doc 2			0	.44	.22		
Doc 3			0			.36	.36

Analytics and Predictions

Types of Analytics

Types of Analytics

Type of Analytics	Description
Descriptive	Understand what happened
Exploratory	Find out why something is happening
Inferential	Understand a population from a sample
Predictive	Forecast what is going to happen
Causal	What happens to one variable when you change another
Deep	Use of advanced techniques to understand large and multi-source datasets

Exploratory Data Analysis

Goals of EDA

- Understand the predictors and targets in the data set
  - Spreads
  - Correlations
- Uncover the patterns and trends
- Find key variables and eliminate unwanted variables
- Detect outliers
- Validate previous data ingestion processes for possible mistakes
- Test assumptions and hypothesis

Tools used for EDA

- Correlation matrices
- Boxplots
- Scatterplots
- Principal component Analysis
- Histograms

V2 Maestros
The Data Science Experts

Machine Learning

Overview

- Data contains attributes
- Attributes show relationships (correlation) between entities
- Learning – understanding relationships between entities
- Machine Learning – a computer analyzing the data and learning about relationships
- Machine Learning results in a model built using the data
- Models can be used for grouping and prediction

Data for machine learning

- Machines only understand numbers
- Text Data need to be converted to equivalent numerical representations for ML algorithms to work.
- Number representation
  - (Excellent, Good, Bad can be converted to 1,2,3)
- Boolean variables
  - 3 new Indicator variables called Rating-Excellent, Rating-Good, Rating-Bad with values 0/1
- Document Term matrix

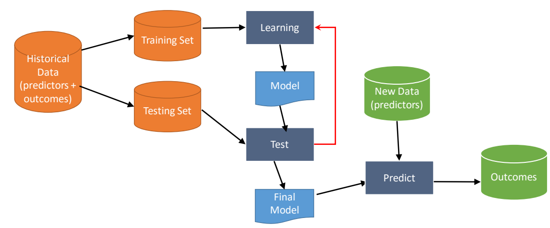
Unsupervised Learning

- Finding hidden structure / similarity / grouping in data
- Observations grouped based on similarity exhibited by entities
- Similarity between entities could be by
  - Distance between values
  - Presence / Absence
- Types
  - Clustering
  - Association Rules Mining
  - Collaborative Filtering

Supervised Learning

- Trying to predict unknown data attributes (outcomes) based on known attributes (predictors) for an entity
- Model built based on training data (past data) where outcomes and predictors are known
- Model used to predict future outcomes
- Types
  - Regression (continuous outcome values)
  - Classification (outcome classes)

Supervised Learning Process



Training and Testing Data

- Historical Data contains both predictors and outcomes
- Split as training and testing data
- Training data is used to build the model
- Testing data is used to test the model
  - Apply model on testing data
  - Predict the outcome
  - Compare the outcome with the actual value
  - Measure accuracy
- Training and Test fit best practices
  - 70-30 split
  - Random selection of records. Should maintain data spread in both datasets

V2 Maestros
The Data Science Experts

Comparing Results

Confusion Matrix

- Plots the predictions against the actuals for the test data
- Helps understand the accuracy of the predictions
- Predictions can be Boolean or classes

		Actual		
		TRUE	FALSE	Total
Prediction	TRUE	44	6	50
	FALSE	9	41	50
	Total	53	47	100

Prediction Types

- The importance of prediction types vary by the domain
- True Positive (TP) and True Negative (TN) are the correct predictions
- False Negative (FN) can be critical in medical field
- False Positive (FP) can be critical in judicial field

		Actual	
		TRUE	FALSE
Prediction	TRUE	True Positive	False Positive
	FALSE	False Negative	True Negative

Confusion Matrix metrics

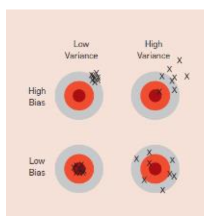
- Accuracy
  - Measures the accuracy of the prediction
  - Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Sensitivity
  - Hit rate or recall
  - Sensitivity = TP / (TP + FN)
- Specificity
  - True negative rate
  - Specificity = TN / (TN + FP)
- Precision
  - Precision = TP / (TP + FP)

		Actual	
		TRUE	FALSE
Prediction	TRUE	True Positive	False Positive
	FALSE	False Negative	True Negative

Prediction Errors

Bias and Variance

- Bias happens when the model "skews" itself to certain aspects of the predictors, while ignoring others. It is the error between prediction and actuals.
- Variance refers to the stability of a model – Keep predicting consistently for new data sets. It is the variance between predictions for different data sets.



Types of Errors

- In-Sample error is the prediction error when the model is used to predict on the training data set it is built upon.
- Out-of-sample error is the prediction error when the model is used to predict on a new data set.
- Over fitting refers to the situation where the model has very low in-sample error, but very high out-of-sample error. The model has "over fit" itself to the training data.

Linear Regression

Linear Relationships

Regression Analysis

- Method of investigating functional relationship between variables
- Estimate the value of dependent variables from the values of independent variables using a relationship equation
- Used when the dependent and independent variables are continuous and have some correlation.
- Goodness of Fit analysis is important.

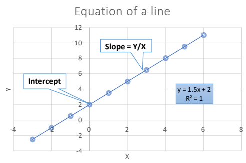
Linear Equation

- X is the independent variable
- Y is the dependent variable
- Compute Y from X using

  $Y = \alpha X + \beta$

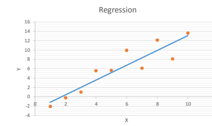
Coefficients:

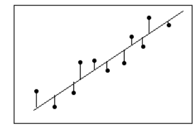
- α = Slope = Y/X
- β = Intercept = value of Y when X=0



Fitting a line

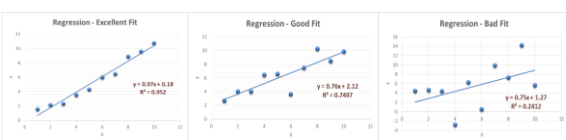
- Given a scatter plot of Y vs X, fit a straight line through the points so that the sum of square of vertical distances between the points and the line (called residuals) is minimized
- Best line = least residuals
- A line can always be fitted for any set of points
- The equation of the line becomes the predictor for Y





Goodness of Fit

- R-squared measures how close the data is to the fitted line
- R-squared varies from 0 to 1. The higher the value, the better the fit
- You can always fit a line. Use R-squared to see how good the fit is
- Higher correlation usually leads to better fit



Multiple regression

- When there are more than one independent variable that is used to predict the dependent variable.
- The equation $Y = \beta + \alpha_1 {*} X_1 + \alpha_2 {*} X_2 + \ldots + \alpha_p {*} X_p$
- Same process used for prediction as a single independent variable
- Different predictors have different levels of impact on the dependent variable

Using Linear Regression for ML

- ML Technique to predict continuous data – supervised learning
- Predictors and outcomes provided as input
- Data analyzed (training) to come up with a linear equation
  - Coefficients
  - Intercept
  - R-squared
- Linear equation represents to model.
- Model used for prediction
- Typically fast for model building and prediction

Summary – Linear Regression

Advantages

- Fast
- Low cost
- Excellent for linear relationships
- Relatively accurate Continuous variables

Shortcomings

- Only numeric/ continuous variables
- Cannot model non-linear / fuzzy relationships
- Sensitive to outliers

Used in

- Oldest predictive model used in a wide variety of applications to predict continuous values

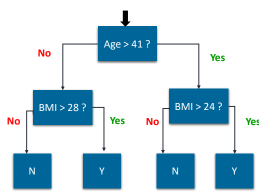
Decision Trees

Overview

- The simplest, easy to understand and easy to explain ML technique.
- Predictor variables are used to build a tree that would progressively predict the target variable
  - Trees start with a root node that start the decision making process
  - Branch nodes refine the decision process
  - Leaf nodes provide the decisions
- Training data is used to build a decision tree to predict the target
- The tree becomes the model that is used to predict on new data

Example

Age	BMI	Is Diabetic
24	22	N
33	28	N
41	36	Y
48	24	N
58	31	Y
61	35	Y



Predicting "Is Diabetic ?"

Choosing the right Predictors

- The depth of trees are highly influenced by the sequence in which the predictors are chosen for decisions
- Using predictors with high selectivity gives faster results
- ML implementations automatically make decisions on the sequence /preference of predictors

Summary – Decision Trees

Advantages
- Easy to interpret and explain
- Works with missing data
- Sensitive to local variations
- Fast

Shortcomings
- Limited Accuracy
- Bias builds up pretty quickly
- Not good with large predictors

Used in
- Credit approvals
- Situations with legal needs to explain decisions
- Preliminary categorization

Naïve Bayes

Bayes' theorem (too) simplified

- Probability of an event A = P(A) is between 0 and 1
- Bayes' theorem gives the conditional probability of an event A given event B has already occurred.

  P(A/B) = P(A intersect B) * P (A) /P(B)
- Example
  - There are 100 patients
  - Probability of a patient having diabetes is P(A) = .2
  - Probability of patient having diabetes (A) given that the patient's age is > 50 (B) is P(A/B) = .4

Naïve Bayes Classification

- Application of Bayes' theorem to ML
- The target variable becomes event A
- The predictors become events B1 – Bn
- We try to find P(A / B1-Bn)

Age	BMI	Is Diabetic	
24	22	N	Probability of Is Diabetic = Y given that Age = 24 and BMI = 22
41	36	Y	Probability of Is Diabetic – Y given that Age = 41 and BMI = 36

Model building and prediction

- The model generated stores the conditional probability of the target for every possible value of the predictor.

	Overall	Age						Gender	
Salary		1 to 20	20 to 30	30 to 40	40 to 50	50 to 60	60 to 100	Female	Male
< 50K	.75	0.1	0.3	0.25	0.17	0.1	0.08	0.39	0.61
> 50K	.25	0.03	0.08	0.3	0.32	0.2	0.07	0.15	0.85
Overall		.08	.24	.26	.21	.12	.08	.33	.67

- When a new prediction needs to be done, the conditional probabilities are applied using Bayes' formula to find the probability
  - To predict for Age = 25
  - P(Salary < 50K / Age=25) = 0.3 * 0.75 / 0.24 = ~ 0.92
  - P(Salary > 50K / Age=25) = 0.08 * 0.25 / 0.24 = ~ 0.08

Summary – Naïve Bayes

Advantages
- Simple and fast
- Works well with noisy and missing data
- Provides probabilities of the result
- Very good with categorical data

Shortcomings
- Limited Accuracy
- Expects predictors to be independent
- Not good with large numeric features

Used in
- Medical diagnosis
- Spam filtering
- Document classification
- Sports predictions

Random Forests

Overview

- Random Forest is one of the most popular and accurate algorithms
- It is an Ensemble method based on decision trees
  - Builds multiple models – each model a decision tree
  - For prediction – each tree is used to predict an individual result
  - A vote is taken on all the results to find the best answer

How it works

- Lets say the dataset contains m samples (rows) and n predictors (columns)
- x trees are built, each with a subset of data
- For each tree, a subset of m rows and n columns are chosen randomly.
- For example, if the data has 1000 rows and 5 columns, each tree is built using 700 rows and 3 columns
- The data subset is used to build a tree
- For prediction, new data is passed to each of the x trees and x possible results obtained
- For example, if we are predicting buy=Y/N and there are 500 trees, we might get 350 Y and 150 N results
- The most found result is the aggregate prediction.

Summary – Random Forest

Advantages

- Highly accurate
- Efficient on large number of predictors
- Fully parallelizable
- Very good with missing data

Shortcomings

- Time and Resource consuming
- For categorical variables, bias might exist if levels are disproportionate

Used in

- Scientific Research
- Competitions
- Medical Diagnosis

V2 Maestros
The Data Science Experts

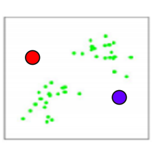
K-means Clustering

Overview

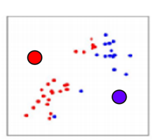
- Unsupervised Learning technique
- Popular method for grouping data into subsets based on the similarity
- Partitions n observations with m variables into k clusters where by each observation belongs to only one cluster
- How it works
  - An m dimensional space is created
  - Each observation is plotted based on this space based on the variable values
  - Clustering is done by measuring the distance between points and grouping them
- Multiple types of distance measures available like Euclidian distance and Manhattan distance

Clustering - Stages

- Dataset contains only m=2 variables. We will create k=2 clusters
- Plot observations on a two dimensional plot

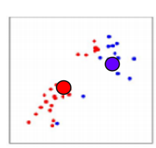


- Choose k=2 centroids at random
- Measure the distance between each observation to each centroid

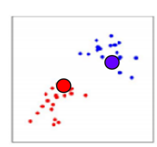


- Assign each observation to the nearest centroid
- This forms the clusters for round 1

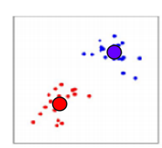
Clustering - Stages



- Find the centroid of each of the cluster
- Centroid is the point where the sum of distances between the centroid and each point is minimum



- Repeat the process of finding the distance between each observation to each centroid (the new one) and reassign each point to the nearest one



- Find the centroid for the new clusters
- Repeat the process until the centroids don't move

Summary – K-means clustering

Advantages
- Fast
- Efficient with large number of variables
- Explainable

Shortcomings
- K needs to be known
- The initial centroid position has influence on clusters formed

Used in
- Preliminary grouping of data before other classification
- General grouping
- Geographical clustering

Collaborative Filtering

Association Rules Mining

Overview

- ARM shows how frequently sets of items occur together
  - Find Items frequently brought together
  - Find fraudulent transactions.
  - Frequent Pattern Mining/ Exploratory Data Analysis
  - Finding the next word
- One of the clustering techniques
- Assumes all data are categorical, not applicable for numeric data
- Helps generate association rules that can be then used for business purposes like stocking aisles.

Datasets

- Market basket transactions
  - Tran 1 { bread, cheese, milk}
  - Tran 2 { apple, eggs, yogurt}
  - Tran 3 {bread, eggs}
- Text document data set (bag of words)
  - Doc 1 { cricket, sachin, India }
  - Doc 2 { soccer, messi, Barcelona}
  - Doc 3 { sachin, messi, superstars}

ARM measures

- Let N be the number of transactions
- Let X, Y and Z be individual items
- Support measures how frequently an combination of items occurs in the transactions
  - Support(X) = count(transactions with X)/ N
  - Support(X,Y)= count(transactions with X and Y)/N
- Confidence measures the expected probability that Y would occur when X occurs
  - Confidence(X -> Y) = support(X,Y) / support(X)
- Lift measures how many more times X and Y occurs together than expected
  - Lift(X -> Y) = confidence(X->Y) / support(Y)

Rules and goals

- A rule specifies when one item occurs the other too occurs
  - When bread is brought, milk is brought 33% of the time.
  - When India occurs in the bag of words, sachin occurs 20% of the time.
- Goal is to find all rules that satisfy the user specified minimum support and minimum confidence
- A frequent itemset is an itemset whose support is > the minimum support level specified.
- Apriori algorithm is the most popular ARM algorithm

Data Formats

- Transaction form
  - a, b, c
  - a, c, d, e
  - a, d
- Table form

Attr1,	Attr2,	Attr 3
A	B	C
A	C	D

- Table should be converted to transaction

  (Attr1 = A), (Attr2 = B), (Attr3 = C)
  (Attr1 = A), (Attr2 = C), (Attr3 = D)

Artificial Neural Networks

Overview

- Biologically inspired by how the human brain works.
- A Black box algorithm (a full explanation would require few hours and mathematical prerequisites)
- Used in artificial intelligence domain and of late for machine learning
- Helps discover complex correlations hidden in the data similar to the human brain
- Works well on noisy data and where the relationships between variables is vaguely understood.
- Fast prediction
- Very slow training and easy to over fit

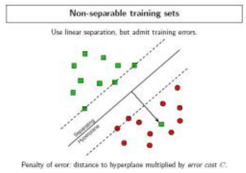
Support Vector Machines

Support Vector Machines

- A Black box method for machine learning – inner workings are complex, tricky and difficult to understand.
- One of the kernel methods.
- Algorithm based on vector geometry and statistical learning theory
- Can model highly complex relationships – very popular in pattern recognition (face recognition, text recognition etc.)
- Successful applications in many fields like bioinformatics, image recognition etc.
- Used for both classification and regression (discrete and continuous outcomes).

How it works

- Plot feature variables in a multi dimensional plot
- Try to draw a hyper plane that separates similar groups of points
- Look for the maximum margin hyper plane (one that provides the most separation between the points).
- Support vectors are data points that lie closest to the hyper plane



Bagging

Overview

- Bootstrap Aggregating /Ensemble method
- Uses a base classifier (Decision trees) to train on multiple sample sets and to generate multiple models
- Prediction is done using each model and the most occurring result across all models is selected.
- For each training round a different bootstrap replicate dataset is constructed based on the original dataset.
  - If the original dataset has m examples, then n rounds of sampling is done to get m/n examples each
  - The n samples sets are then added up to for a dataset of m size
  - Examples could be duplicated in the sample sets or might not occur at all.

How it works

- Suppose we want to run training 5 times on a dataset that has 8 records (Record ID 1:8).
- For each round, we do 2 sets of sampling with replacement to build the bootstrap aggregate.
- Training round 1:
  - sample 1 : 1,4,5,7
  - sample 2 : 2,4,6,7
  - bootstrap replicate : 1,2,4,4,5,6,7,7
- Training round 2:
  - sample 1 : 2,3,5,6
  - sample 2 : 1,2,6,8
  - bootstrap replicate : 1,2,2,3,5,6,6,8

Things to note

- May produce improved results than the base classifier if the base classifier produces unstable results.
- High resource requirements and takes longer times to build models
- Various models available – difference is the base classifier used (some examples)
  - adaBag
  - Bagged CART
  - Bagged Flexible Discriminant Analysis
  - Bagged Logic Regression
  - Model Averaged Neural Network

Boosting

Overview

- Ensemble method like bagging
- Creates multiple models.
- Prediction done on multiple models and results aggregated to deliver the final prediction
- Different samples (records /observations) get different weights during learning for each of the training rounds
- Weight determined based on misclassification during previous round.

How it works

- Multiple rounds of training. Weights of all records equal for the first round
- For each model building round
  - Build the model
  - Predict on the same training set using the model
  - Find misclassified records
  - Increase weights of misclassified records
  - Repeat model building
- Results in multiple models
- Predictions done using each model. Results aggregated.

Things to note

- High resource requirements and takes longer times to build models
- Use a set of weak learners to create a strong learner
- Reduces bias
- Different algorithms available
  - Boosted Classification Trees
  - Boosted Generalized Additive Model
  - Boosted Generalized Linear Model

Dimensionality Reduction

Principal Component Analysis

Issues with too many predictors

- Memory requirements
- CPU requirements / time taken for machine learning algorithms
- Correlation between predictors
- Over fitting
- Some ML algorithms don't work fine with too many predictors

Manual selection

- Using domain knowledge
  - Purely based on hypothesis
  - Risky – there could be unknown correlations
- Using Correlation co-efficients
  - Variables with good correlation can only be picked up.
- Using Decision Trees
  - Decision trees are fast and choose variables based on correlation
  - Variables used in the decision trees can be picked for further processing

Principal Component Analysis

- Used to reduce the number of predictors
- Based on Eigen Vectors and Eigen Values.
- Given a set of M predictors, PCA transforms this to a set of N predictors such that N < M
- The new predictors are derived predictors called PC1, PC2, PC3
- The new predictors retain similar levels of correlation and predictability like the original predictors

Recommendation Engines

What is a Recommendation Engine?

- Also called Collaborative filtering
- Analyze past data to understand user / entity behavior
- Identify "similar" items /users / entities
- Recommend based on similarity of behavior
- Example
  - Tom and Chris both like similar items. In the past 1 year, Tom has brought 42 items and Chris has brought 35 items. 28 of these are same.
  - Tom buys a new item which Chris has not bought. Recommend that to Chris.
- Used by
  - Netflix for movie recommendations
  - Amazon for product recommendations
  - YouTube for video recommendations

Recommendation Types

- User based Recommendations
  - Identify similar users and form User neighborhoods.
  - If one user buys a new product, recommend that to all users in the neighborhood.
  - "Similar customers brought…"
- Item based Recommendations
  - Identify Items that are frequently brought together.
  - If one item is brought by a user, recommend the related items to the user.
  - "Users who brought this item also brought…"

Input for Recommenders

- Recommender algorithms take as input as specific format
  - User ID
  - Item ID
  - Score
- Scores indicate relative preference of the user for the item
  - Boolean values
  - Rating scores
  - Measure of sales volume

Building a User based recommender

- Find affinity scores between users based on similarity of behavior.
  - Uses similarity measures like cosine similarity, Pearson correlation etc.
- For user neighborhoods with each neighborhood containing users with high inter-user scores.
- If one user shows a new behavior (buys an item), recommend that to other users in the neighborhood.
- Continuous processing of past data and building of neighborhoods.

Building a Item based recommender

V2 Maestros
The Data Science Experts

- Find affinity scores between items based on usage (used together).
  - Uses similarity measures like cosine similarity, Pearson correlation etc.
- If one item is brought by a user, recommend items with high similarity scores with that item.
- Continuous processing of past data and building of neighborhoods.
- Item based recommenders are superior to user based, since
  - No. of items are limited
  - More usage data available since the list of items are relatively static.

Congratulations on finishing this course !

We hope this course helps you to advance your career.

Best of luck !