A Practical Introduction to Machine Learning

O’Reilly Media live online training
August 26th, 2019
9:00am-12:00pm PST
Presented by Matthew Kirk

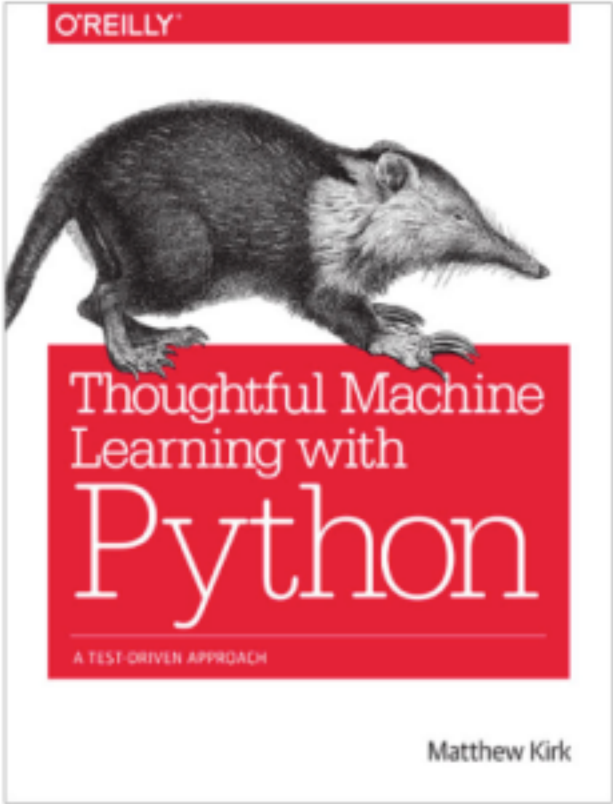
INTRODUCTION

`whoami`

Background

Thoughtful Machine Learning with Python

YourChiefScientist.com



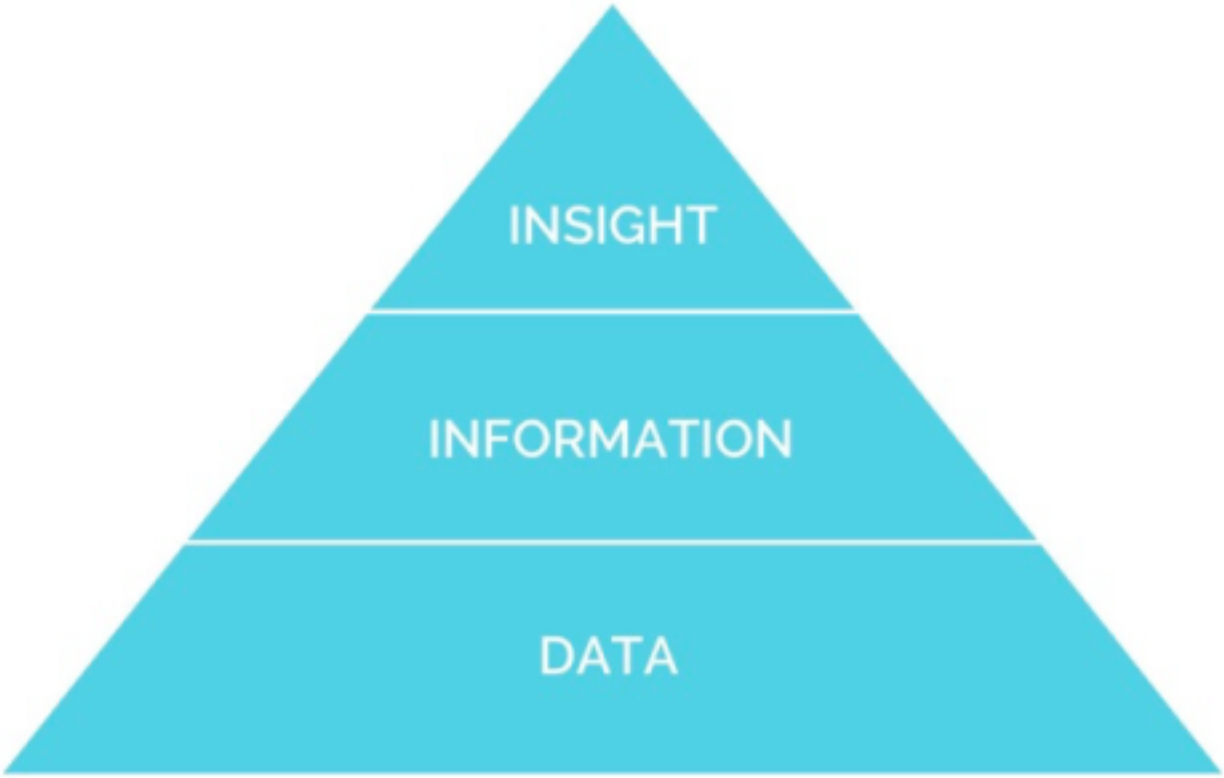
WHAT IS MACHINE LEARNING?

A toolkit of algorithms that finds insight from data.

WHAT IS MACHINE LEARNING?

WHAT IS MACHINE LEARNING?

WHAT IS MACHINE LEARNING?



Before we get started

Download Anaconda

https://www.anaconda.com/download

GENERAL HOUSEKEEPING

This course follows the format:

- Lecture
- Quiz
- Demo
- Lab

LECTURE

- It won’t be academic
- It will be about introducing a mindset behind the theory
- As well as practical applications of it
- Write your questions down to ask during Demo & Lab time.

QUIZ

- Simple 5 minute quizzes of 3 questions multiple choice.
- These serve as a way for you to remember key insights about what we’re talking about.

DEMO

- The purpose of the demo is to introduce and guide you on the Lab sections.
- I will introduce the data and general purpose of the lab sections.
- Also I will give you helpful guidance and things to experiment with

LAB

- This is the section where you get to learn and try out what we’ve been working with.
- This is meant for you to experiment.
- Try things, fail, and if you don’t finish it all in 20 minutes that’s ok.

HOW WE REASON

DEDUCTIVE REASONING

- Philosophical Logic
  - Modus Ponens
  - Syllogism
  - Contrapositives
- Economics
- Political Science
- Theory of Rationality

PROBLEMS WITH DEDUCTION

- Predictable irrationality
- Logical fallacies
- Local minima

ARTIFICIAL INTELLIGENCE

- Planning
- Expert Systems
- Ontologies
- Perception
- Mostly deductive reasoning
- Heuristics
- Learning, Machine Learning

INDUCTIVE REASONING

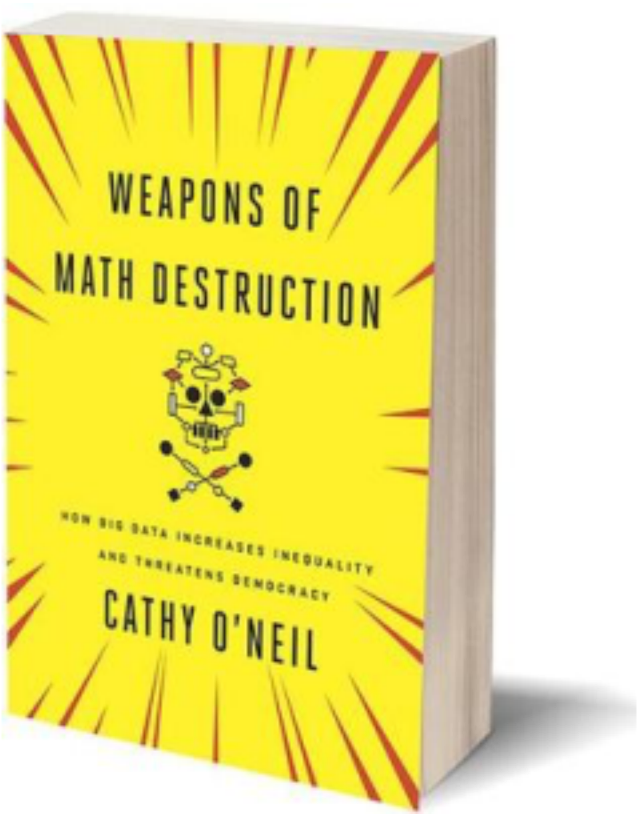
- Evidence based
- Statistics
- Generalizations
- Statistical Syllogisms
- Proof by Induction
- Prediction
- Analogies
- Causal inference

THE PROBLEMS WITH INDUCTIVE REASONING

- Biases
    - Confirmation bias
    - Attribution bias
    - Favoritism
    - Inductive bias
    - Racism
    - Sexism
- Black Swans
- Not all variables are available

WEAPONS OF MATH DESTRUCTION

- Author Cathy O'Neil
- Creditworthiness → Confirmation Bias
- Racism → Attribution Bias
- Sexism → Generalizations



3 CLASSES OF MACHINE LEARNING ALGORITHMS

3 CLASSES OF MACHINE LEARNING ALGORITHMS

SUPERVISED LEARNING

Finding a function that maps data to values based on previous observations

Examples:

Naïve Bayesian Classifier
K-Nearest Neighbors
Support Vector Machines

3 CLASSES OF MACHINE LEARNING ALGORITHMS

SUPERVISED LEARNING

Finding a function that maps data to values based on previous observations

Examples:

Naïve Bayesian Classifier
K-Nearest Neighbors
Support Vector Machines

UNSUPERVISED LEARNING

Algorithm looks for patterns in the data without any guidance of values

Examples:

Auto encoders
Clustering
Matrix Factorization

3 CLASSES OF MACHINE LEARNING ALGORITHMS

SUPERVISED LEARNING

Finding a function that maps data to values based on previous observations

Examples:

Naïve Bayesian Classifier
K-Nearest Neighbors
Support Vector Machines

UNSUPERVISED LEARNING

Algorithm looks for patterns in the data without any guidance of values

Examples:

Auto encoders
Clustering
Matrix Factorization

REINFORCEMENT LEARNING

Algorithm looks to maximize rewards over a time period, given previous observations

Examples:

Q-Learning
TD-Lambda
Multi-Armed Bandits

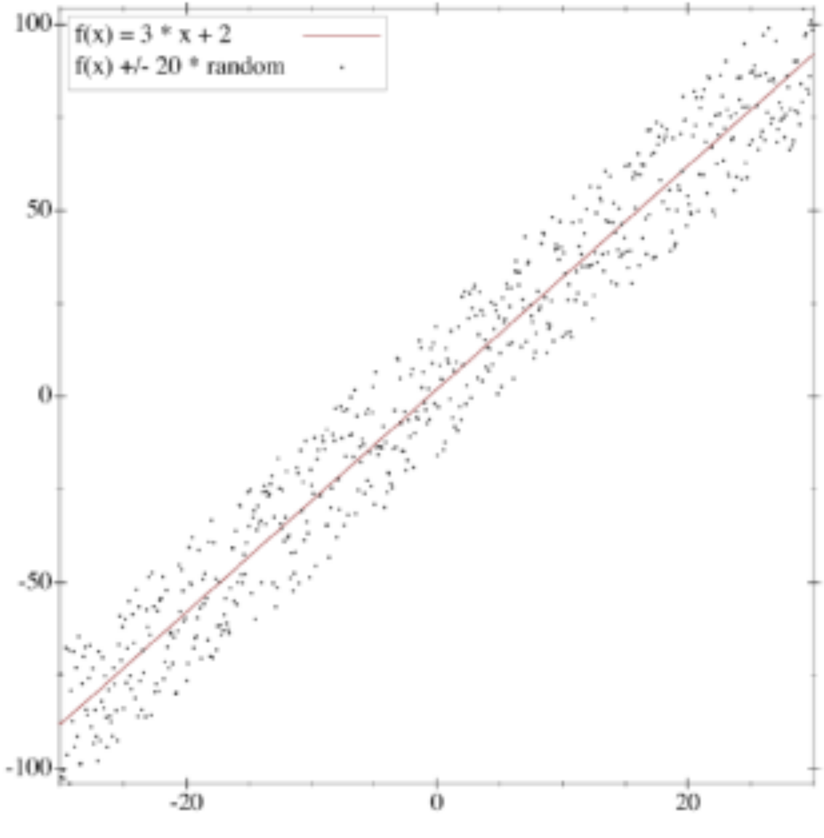
MACHINE LEARNING

- Finding insight from a mountain of data
- Unsupervised Learning
  - Clustering ⟷ Group data
  - Matrix Factorization ⟷ Generalize data
  - Autoencoders ⟷ Generalize data
- Supervised Learning
  - Classify ⟷ Analogies
  - Regression ⟷ Predict
- Reinforcement Learning
  - Policy Iteration ⟷ Statistical Syllogism

Ok but what’s Deep Learning?

- Learning, Machine Learning
    - Neural Networks
        - Automatic feature detection
        - Prediction from massive datasets

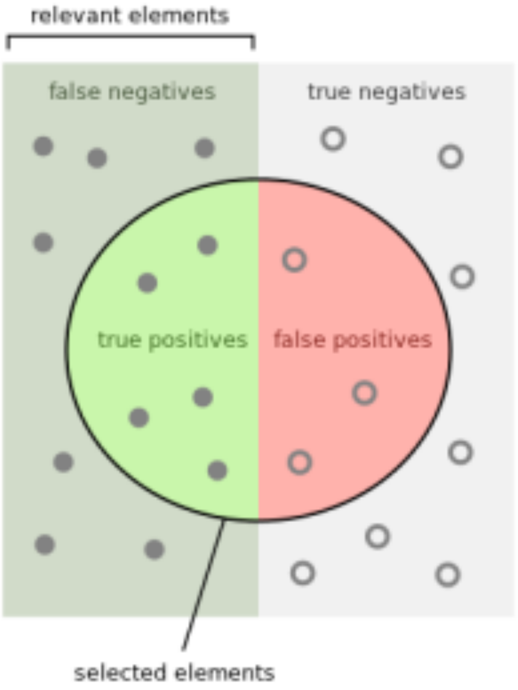
SUPERVISED LEARNING



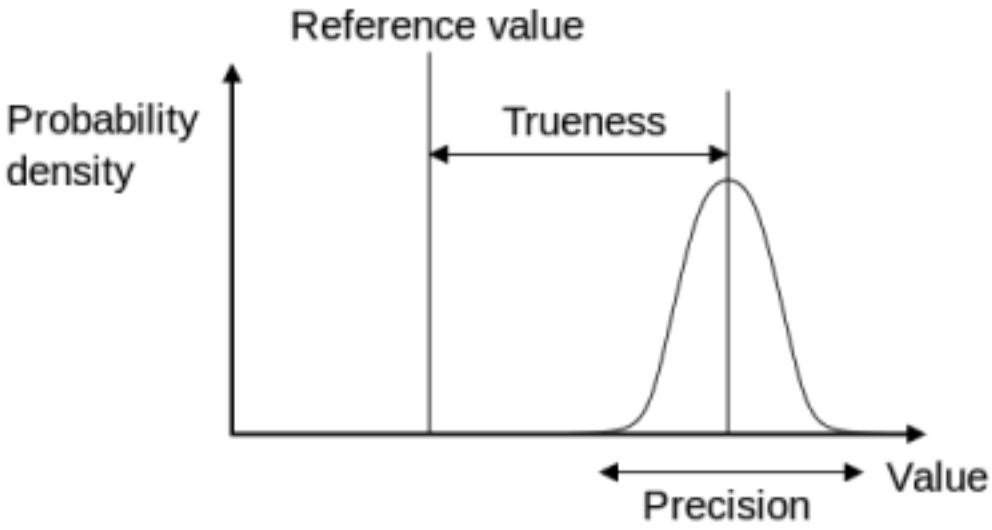
VALIDATING SUPERVISED LEARNING

- Precision
- Recall
- Accuracy
- Confusion matrix

PRECISION & RECALL



ACCURACY


Reference value
Probability
density
Trueness
Value
Precision

CONFUSION MATRIX

		Predicted		
		Cat	Dog	Rabbit
Actual	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

Precision	Cat	71.43%
	Dog	37.50%
	Rabbit	91.67%
	Average	66.87%
Recall	Cat	62.50%
	Dog	37.50%
	Rabbit	91.67%
	Average	63.89%
Accuracy	Cat	-
	Dog	-
	Rabbit	-
	Average	70.37%

CODING PRINCIPLES

- Single responsibility
- Open closed principle
- Liskov substitution
- Interface segregation
- Dependency inversion

SINGLE RESPONSIBILITY

- Machine Learning algorithms will do only one thing
  - Classify into categories
  - Regress to value
  - Group
  - Generalize
  - Strategize
- Separate out each algorithm into different pieces

OPEN CLOSED

- Machine learning models are open for configuration with parameters
- But closed for modification on the actual algorithm

LISKOV SUBSTITUTION

- You can easily use Naive Bayesian Classifier in place of Neural Nets, or KNN Classifiers. They all do the same thing differently.

INTERFACE SEGREGATION

- Interface to machine learning models
- Offline:
  - Function train(training_data) trains model
  - Function test(testing_data) tests model
  - Function determine(input) determines output
- Online:
  - Function add(new_data_point) adds new point to model
  - Function test(testing_data) tests model
  - Function determine(input) determines output

DEPENDENCY INVERSION PRINCIPLE

- A lot of machine learning code is plug and play now with graphlab, sklearn and others

HOW WE REASON

SECTION 1 QUIZ

What is the difference between induction and deduction?

a. Induction is irrational and deduction is rational
b. Induction starts with observation while deduction starts with a hypothesis
c. Deduction causes racism, induction is an oven type
d. Deduction is what machine learning does, induction is a type of planning

SECTION 1 QUIZ

What is the difference between induction and deduction?

a. Induction is irrational and deduction is rational
b. Induction starts with observation while deduction starts with a hypothesis
c. Deduction causes racism, induction is an oven type
d. Deduction is what machine learning does, induction is a type of planning

SECTION 1 QUIZ

What is a way to test supervised learning algorithms?

a. Propositional logic
b. Confusion Matrix
c. Fallacy disproving
d. Statistical variance testing

SECTION 1 QUIZ

What is a way to test supervised learning algorithms?

a. Propositional logic
b. Confusion Matrix
c. Fallacy disproving
d. Statistical variance testing

SECTION 1 QUIZ

Clustering is a form of:

a. Unsupervised Learning
b. Reinforcement Learning
c. Supervised Learning
d. Semi-supervised Learning

SECTION 1 QUIZ

Clustering is a form of:

a. Unsupervised Learning
b. Reinforcement Learning
c. Supervised Learning
d. Semi-supervised Learning

HOW WE REASON

DISCUSSION

- High interest credit card debt of machine learning
- Why isn’t machine learning a silver bullet?

HIGH INTEREST CREDIT CARD DEBT OF MACHINE LEARNING

http://bit.ly/1zwONap

- Entanglement
- Hidden Feedback Loops
- Undeclared Consumers
- Unstable Data dependencies
- Underutilized data dependencies
- Correction Cascade
- Glue Code
- Pipeline Jungles
- Experimental Code Paths
- Configuration Debt
- Fixed Thresholds in an ever changing system
- Correlation Changes

BREAK (1o minutes)

Check me out on Linkedin:
https://www.linkedin.com/in/matthewjkirk

Section 2: Lecture

K-NEAREST NEIGHBORS

Distance Based Classification

CALCULATING A HOUSE VALUE

- Valuation of houses
- Neighborhood centric
- Tax records
- Sales records
- Zestimate

CLASSIFICATION OR REGRESSION?

- Is the house in a good neighborhood?
  - A class to classify houses into categories
- How much is this house worth?
  - A nominal value that is attached to how much the house is worth.

CALCULATING VALUE BASED ON NEARNESS

Value is a function of how valuable the neighbors' houses are:

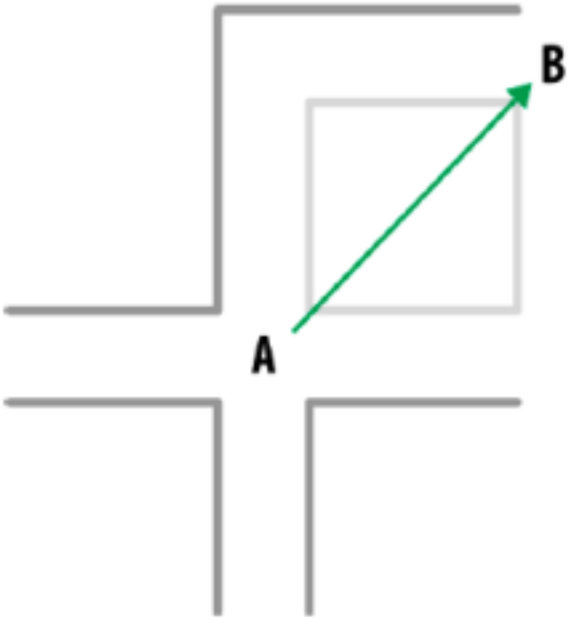
WHAT IS NEARNESS OR CLOSENESS?

- As the crow flies?
- By driving distance?
- By statistical variation?
- By angle?

AS THE CROW FLIES

The euclidean distance

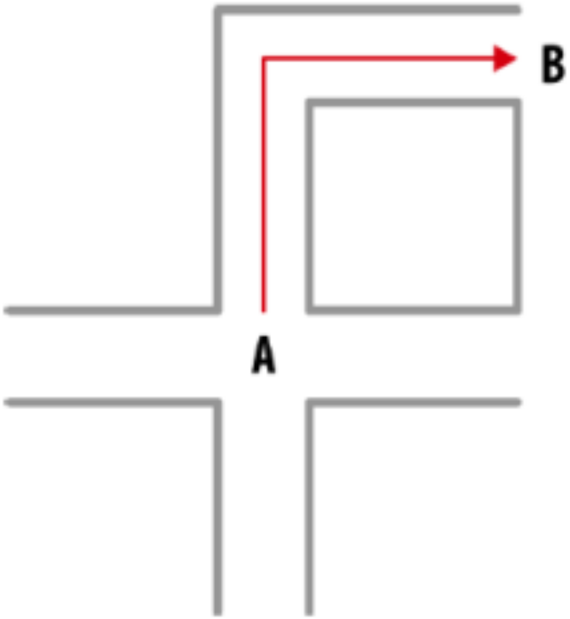
Figure 3-6, page 30



AS YOU DRIVE

Manhattan distance

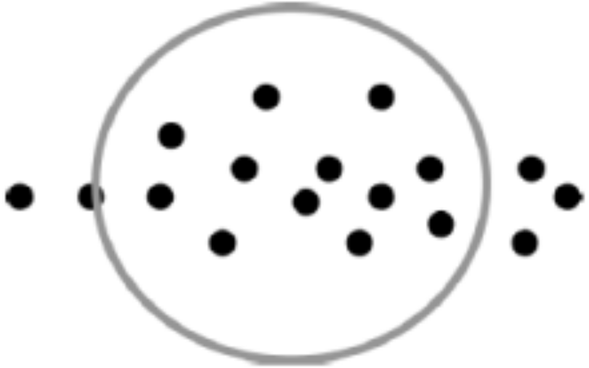
Figure 3-5, page 30



STATISTICAL VARIATION

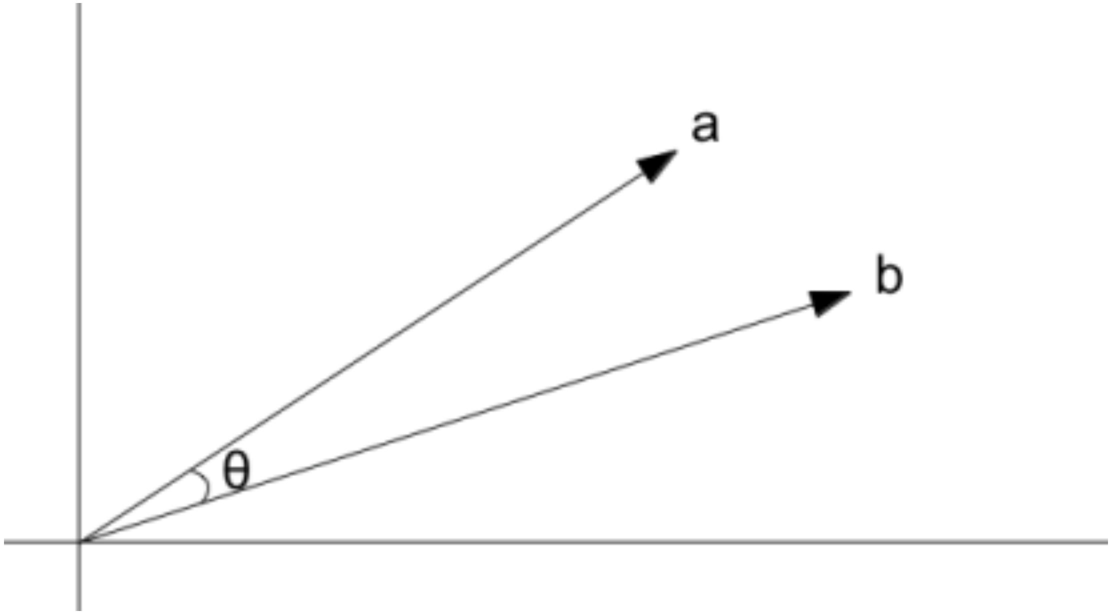
Mahalaobnis distance

Figure 3-7, page 31



DISTANCE OF THE ANGLE

Cosine Similarity



WHAT IS NEARNESS OR CLOSENESS?

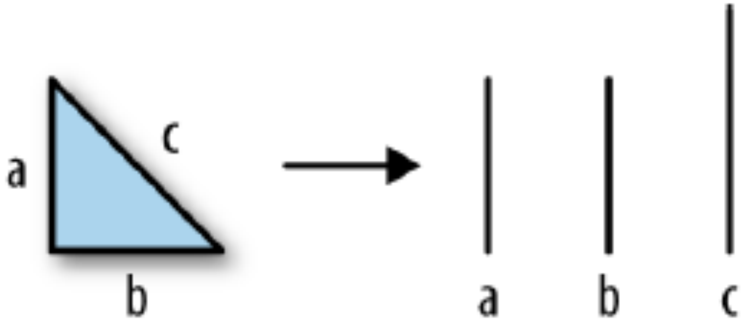
- As the crow flies? ⟷ Euclidean distance
- By driving distance? ⟷ Manhattan Distance
- By statistical variation? ⟷ Mahalanobis distance
- By angle? ⟷ Cosine Distance

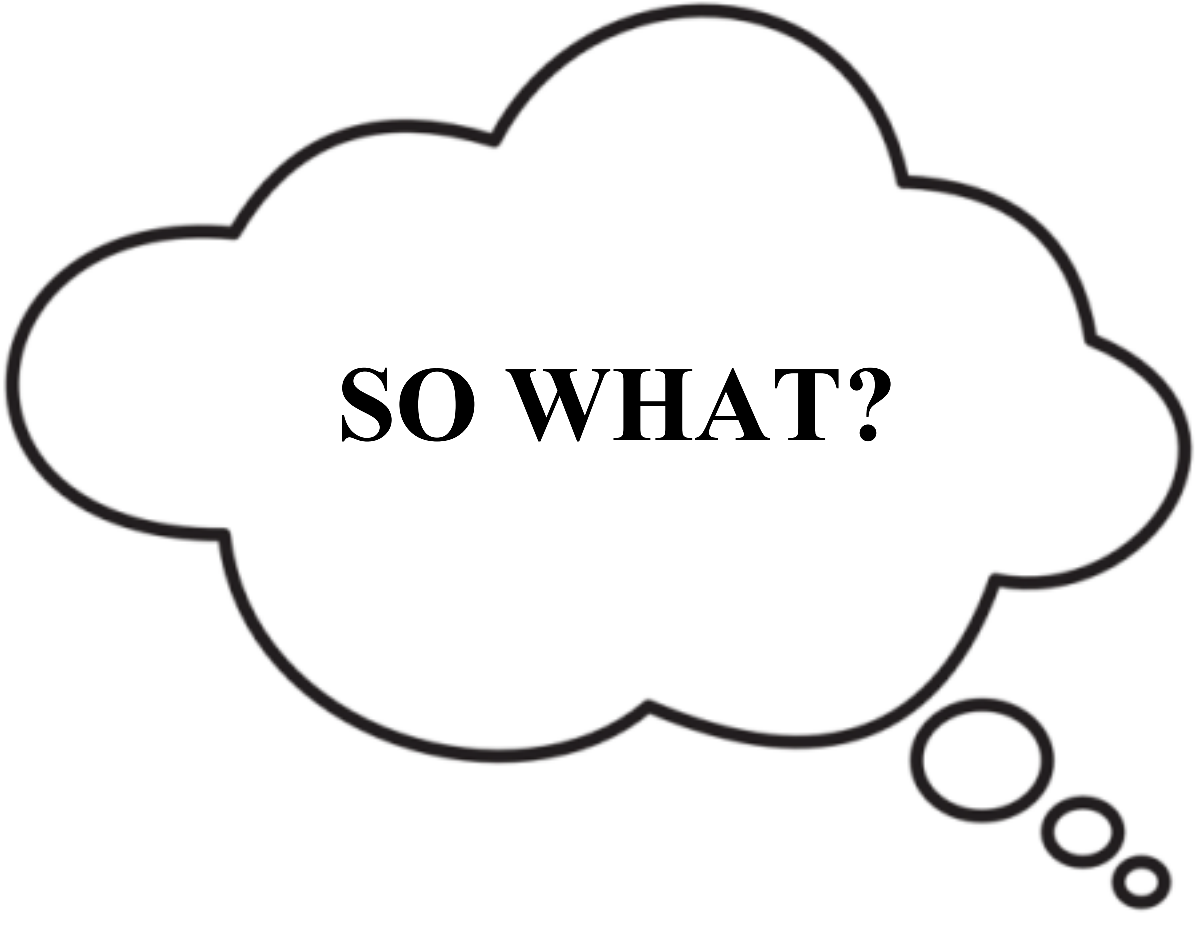
THE TRIANGLE INEQUALITY

$||x + y|| <= ||x|| + ||y||$

$\text{dist}(x + y) <= \text{dist}(x) + \text{dist}(y)$

Figure 3-2, page 25




SO WHAT?

CLASSIFYING HOUSES BY USING K-NEAREST NEIGHBORS

Algorithm:

Pick K > 1

Pick distance measure

Find K nearest points

Aggregate:

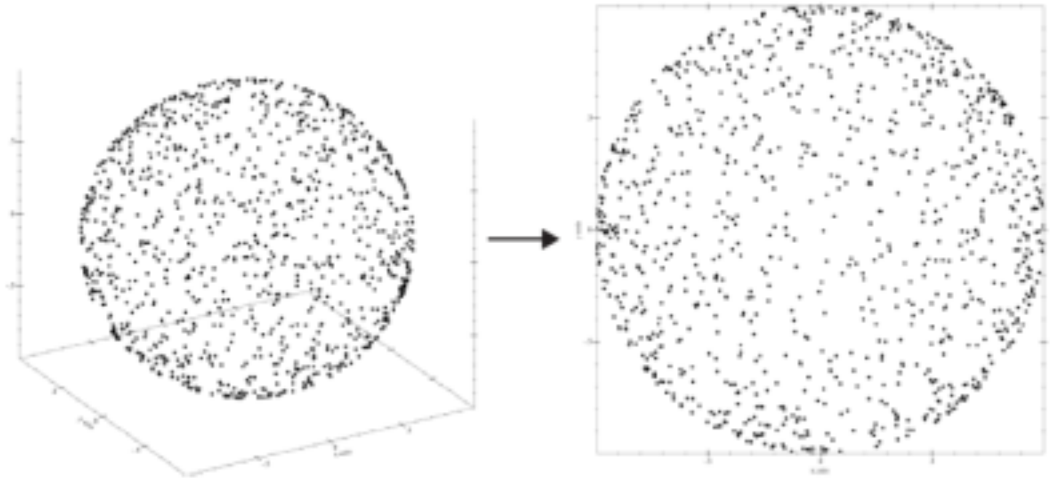
- Classify most common class
- Average/Median/Mode value

CLASSIFY BY K-NEAREST NEIGHBORS

- K=5
- Average Value = 470,400
- Only off by 19k!

THE TRADEOFF

- KNN works better with lower dimensions
  - If in higher dimensions try matrix factorization like PCA, or feature selection algorithms
- Curse of Dimensionality



K-NEAREST NEIGHBORS

Distance Based Classification

SECTION 2 QUIZ

What is an example of a distance function?

a. Kilometers
b. Triangle Inequality
c. Manhattan Distance
d. A Cake song

SECTION 2 QUIZ

What is an example of a distance function?

a. Kilometers
b. Triangle Inequality
c. Manhattan Distance
d. A Cake song

SECTION 2 QUIZ

What is the curse of dimensionality?

a. As dimensions increase distances become more difficult to measure.
b. A voodoo curse put upon the machine learning community with a gris gris.
c. Only happens on supervised learning methods.
d. A problem with all machine learning algorithms.

SECTION 2 QUIZ

What is the curse of dimensionality?

a. As dimensions increase distances become more difficult to measure.
b. A voodoo curse put upon the machine learning community with a gris gris.
c. Only happens on supervised learning methods.
d. A problem with all machine learning algorithms.

SECTION 2 QUIZ

Why would you use Euclidean vs Manhattan distances?

a. Euclidean is easier to compute than Manhattan distance
b. Euclidean is mathematically more sound
c. Manhattan distances can be used everywhere vs euclidean distances
d. Manhattan distances take into consideration constraints of movement while euclidean doesn’t.

SECTION 2 QUIZ

Why would you use Euclidean vs Manhattan distances?

a. Euclidean is easier to compute than Manhattan distance
b. Euclidean is mathematically more sound
c. Manhattan distances can be used everywhere vs euclidean distances
d. Manhattan distances take into consideration constraints of movement while euclidean doesn't.

DEMO

SECTION 2 DEMO

- Where does the data come from?
- How do we go about picking K?
- How can we test this?

THE DATA

- The data we are using today is from King County tax records.
- It includes lots of various factors like whether the house has a view of Mount Rainier or not.
- Mostly what we'll use is location data. But you could easily try something else.

PICKING K for KNN

- Picking K is undetermined. Pick K that maximizes your objective.
- Objectives could be: accuracy in house value, or possibly variance in house value.
- Today let’s go with maximizing accuracy in house value.

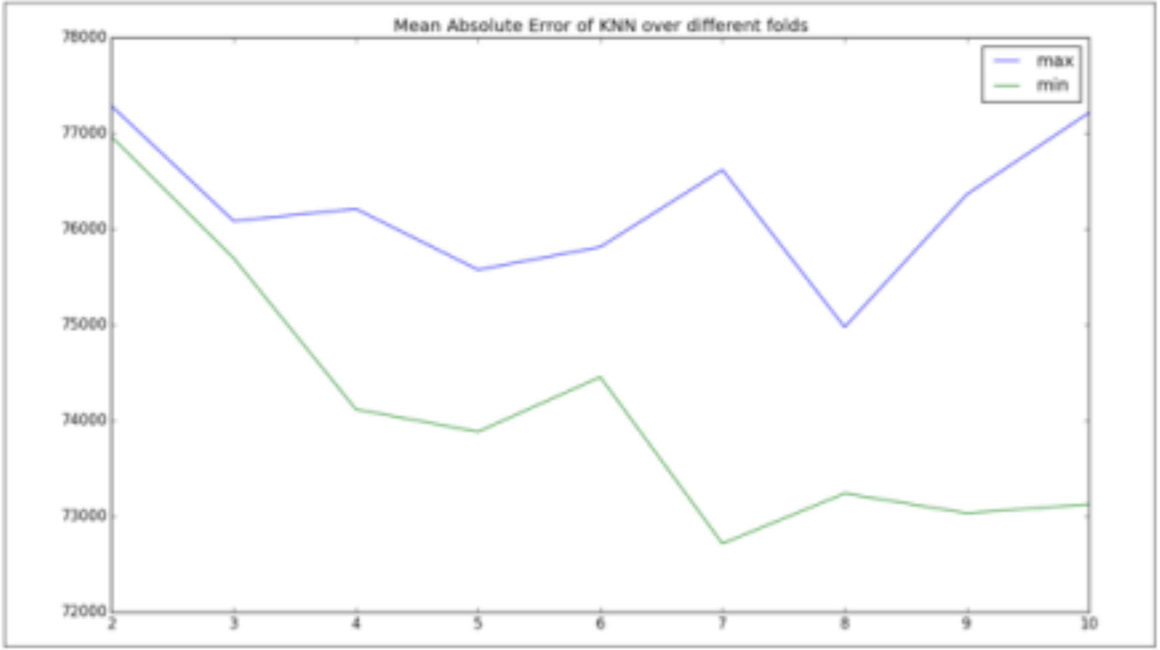
TESTING THIS VIA MEAN SQUARED ERROR

Mean squared error would be the average squared error (x - xhat)^2

This will give you a good enough answer.

Can also use mean absolute error or other metrics that measure _distance_ from original.

MY RESULTS



Mean Absolute Error of KNN over different folds
max
min
78000
77000
76000
75000
74000
73000
72000
2
3
4
5
6
7
8
9
10

LAB COAT TIME

Go to this repo and download the git repository:

https://www.github.com/thoughtfulml/course-1

FILL IN THE BLANKS AND GET SOME HELP

- Get the KNN classifier working and run cross validations
- Try out subsets of features
- Try different training subsets

BREAK (1o minutes)

Battling churn with data science: https://bit.ly/2Pz8kVe

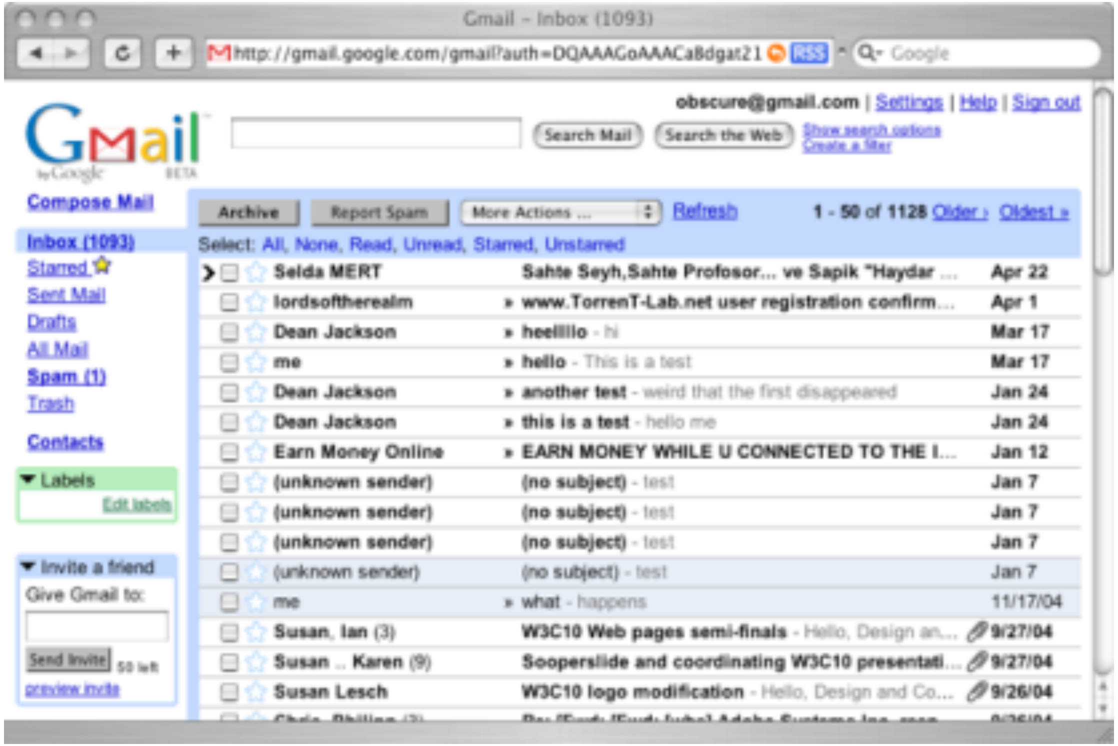
NAIVE BAYESIAN CLASSIFIERS

Probabilistic Based Classification

EMAIL IN THE 1990s

GMAIL IN THE EARLY 2000's

- No more spam why?



LIKELIHOOD OF WORDS AND SPAMMINESS

- Cloud of terms from https://bit.ly/2SJGVSr

HOW CAN WE EXPLOIT THIS?

- Let’s say we have information on emails that are spam, and those that aren’t.
- That would give us a probability of spam vs not, as well as the probability that a word is spam in a given spam category (conditional probability).

WHAT IS A CONDITIONAL PROBABILITY?

- Think of it as a likelihood estimate of whether something will be a given attribute.
- $P(A|B)$ = probability of A happening given that B happened.
- P("$$$" | Spam) = Probability of $$$ happening given we are look at spam messages.

THAT SOUNDS INTERESTING...

- Let's say that we know that P($$$|Spam) = 1% while P($$$|~Spam) = 0.01%.
- That is useful information.

BUT THIS ONLY DESCRIBES THE DATA

- P($$$|Spam) only describes what we already know (the training)
- What we’d rather have is P(Spam|$$$, …).

WARNING:

Mind bending information ahead

BAYES THEOREM

- From now on I'll use W to be the vector of features in a given e-mail (words, features, stems, etc).
- $P(A \mid B) = P(A \,\&\, B) / P(B)$
- $P(B \mid A) = P(A \,\&\, B) / P(A)$
- Therefore
- $P(A \,\&\, B) = P(B|A) * P(A) = P(A|B) * P(B)$
- Therefore
- $P(A|B) = (P(B|A) * P(A)) / P(B)$ Bayes Theorem (not Bae's theorem)

WHAT DOES BAYES THEOREM HAVE TO DO WITH IT?

- We have data on $P(W \mid Spam)$ and want to invert
- Bayes theorem gives us
- $P(Spam \mid W) = P(W \mid Spam)\, P(Spam) / P(W)$

IN SIMPLE TERMS

- Posterior = (Prior * Likelihood) / Evidence
- Posterior is the new distribution we have found as a result of using Bayes theorem
- Prior is the prior distribution is a Binomial distribution of Spam / Not Spam
- Likelihood is the measure of how likely the words are to be spam
- Evidence is really how strong our model is. How many training points do we have?

BUT THIS STILL DOESN'T REALLY HELP US

- $P(W)$ will be extremely small. The probability of you getting the exact same email is pretty slim. Making for calculating this pretty… tough
- Also $P(W \mid Spam) = P(w_1, w_2,...,w_n|Spam) = P(w_1, w_2, \dots, w_n, Spam)$
- Which equates to:
  - $P(w_1, w_2, \dots, w_n, Spam) = P(w_1 \mid w_2, \dots, w_n, Spam) * P(w_2 \mid w_3, \dots, w_n, Spam) * P(w_{n-1} \mid w_n, Spam) * P(w_n \mid Spam) * P(Spam)$

:'(

WHAT IS THERE TO DO?

- Do we need evidence? No
- Can we rewrite that big long conditional probability? Yes (with a catch)

DO WE NEED “EVIDENCE”?

- Not exactly, evidence is mainly a way to make sure that our probabilities are on the scale from 0 to 100%.
- That means we can ignore evidence because really what we care about is a scoring. So instead of focusing on probability we focus on the Spamminess Score.

CAN WE SIMPLIFY THE CONDITIONAL PROBABILITY CHAIN?

- The conditional probability chain basically states that:
- A joint probability is conditional on all of it's parts. So for instance.
- P(Spam | "prince", "$$$") = P("prince", "$$$", Spam)
- P("prince", "$$$", Spam) = P("prince" | "$$$", Spam) * P("$$$"| Spam) * P(Spam)
- We have P("$$$" | Spam) and P(Spam) but not P("prince" | "$$$", Spam)

WHAT MAKES NAIVE BAYESIAN CLASSIFIERS NAIVE?

The Big Naive Assumption:

P("prince", "$$$", Spam) = P("prince" | Spam) * P("$$$" | Spam) * P(Spam)

CAN WE DO THAT?

- Yes and No.
- Naive Bayes works exceptionally well on things like emails where each word will contribute to something like Spam.
- On the other hand it doesn’t work well for data that is dependent on each other.

NAIVE BAYESIAN CLASSIFIER

- Pick highest “score” of given classes.
- argmax c_k NBC(W) = P(W, C_k) / Z

SECTION 3 QUIZ

What is the probability of X given Y?

a. Probability of X and Y divided by Probability of Y
b. Probability of Y and X divided by the probability of X
c. Probability of Y or X divided by Probability of X
d. Probability of X or Y divided by probability of Y

SECTION 3 QUIZ

What is the probability of X given Y?

a. Probability of X and Y divided by Probability of Y
b. Probability of Y and X divided by the probability of X
c. Probability of Y or X divided by Probability of X
d. Probability of X or Y divided by probability of Y

SECTION 3 QUIZ

Naive Bayesian classifiers works on spam filters because…

a. It is fast
b. It is probabilistic
c. It assumes words are independent of each other
d. All of the above

SECTION 3 QUIZ

Naive Bayesian classifiers works on spam filters because…

a. It is fast
b. It is probabilistic
c. It assumes words are independent of each other
d. All of the above

SECTION 3 QUIZ

Why is the naive bayesian classifier naive?

a. Because it didn’t go to prep school.
b. Because it assume probabilities are independent of each other.
c. Because it assumes probabilities are dependent on each other.
d. Because it ignores evidence.

SECTION 3 QUIZ

Why is the naive bayesian classifier naive?

a. Because it didn’t go to prep school.
b. Because it assume probabilities are independent of each other.
c. Because it assumes probabilities are dependent on each other.
d. Because it ignores evidence.

DEMO

TESTING NAIVE BAYESIAN CLASSIFIERS

- Confusion Matrix
- ROC Curve
- Cross Validation

THE IMPORTANT PRINCIPLE

- Filtering out emails that are important is a bad idea.
- False positives is bad False negatives are less bad.

LAB COAT TIME (YOUR TURN)

- Go to https://github.com/thoughtfulml/course-2
- Read the README and complete the TODO’s listed in the code
- Don’t get discouraged and ask for help if needed.

CONCLUSION WRAP UP

- We’ve covered a lot today:
  - SOLID
  - Inductive vs Deductive Reasoning
  - Test Driven Development
  - Distance Based classification and regression
  - Probabilistic based classification
- But this is only the start

SOME RESOURCES

- Battling churn with Data Science:
    - https://bit.ly/2Pz8kVe
- Getting started with data science:
    - www.yourchiefscientist.com