Lecture Notes for

Machine Learning in Python

Professor Eric Larson

Introduction, Syllabus, Data Types

Class Logistics and Agenda

- Agenda:
  - Introductions
  - Syllabus and Course Overview
  - What is Machine Learning?
  - Types of Data
  - Numpy Demo

- My approach to this course:
  - Programming
  - Math
  - **Applications** and **Analytics**

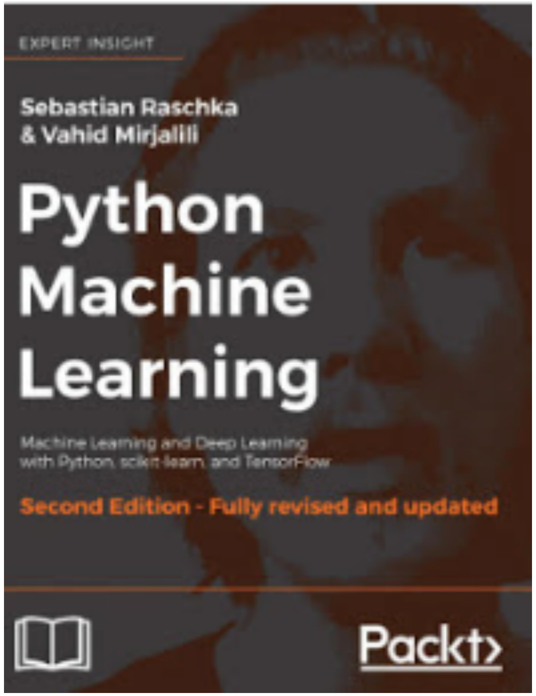
Introductions & Course Syllabus

Introductions

- Me
  - Eric 👍
  - Dr. Larson 👍
  - Prof. Larson 👍
  - Hey man 👎
- You
  - Name
  - Where you grew up
  - Department
  - Grad/Undergrad
  - Something true or false

FAQ

- Text:
  - **Recommended**: Python Machine Learning, Raschka & Mirjalili, Second edition



- Use Canvas for posted course material
- Prerequisites:
  - Linear Algebra, Calculus (Multivariate)
  - Basic statistics and probability
  - Python programming
- Version of python: 3.X
  - Install through Anaconda
  - Use conda environments
  - JupyterLab (or notebooks)
- Most Used Libraries: Numpy, Pandas, Scikit-Learn, Matplotlib, Keras with Tensorflow

How will participation be graded?

- Participation will be graded in the course:
  - **Distance students** will answer these questions via **canvas upload**
  - upload over the last submission
    - must upload the questions throughout semester for full credit
- In Class Students:
  - Choose to respond to the question:
    - Do you think this will work?
      - A: **Yes** this is going to work
      - B: This is **not** going to work
      - C: I cannot use this card
      - D: I do not have a name on my card

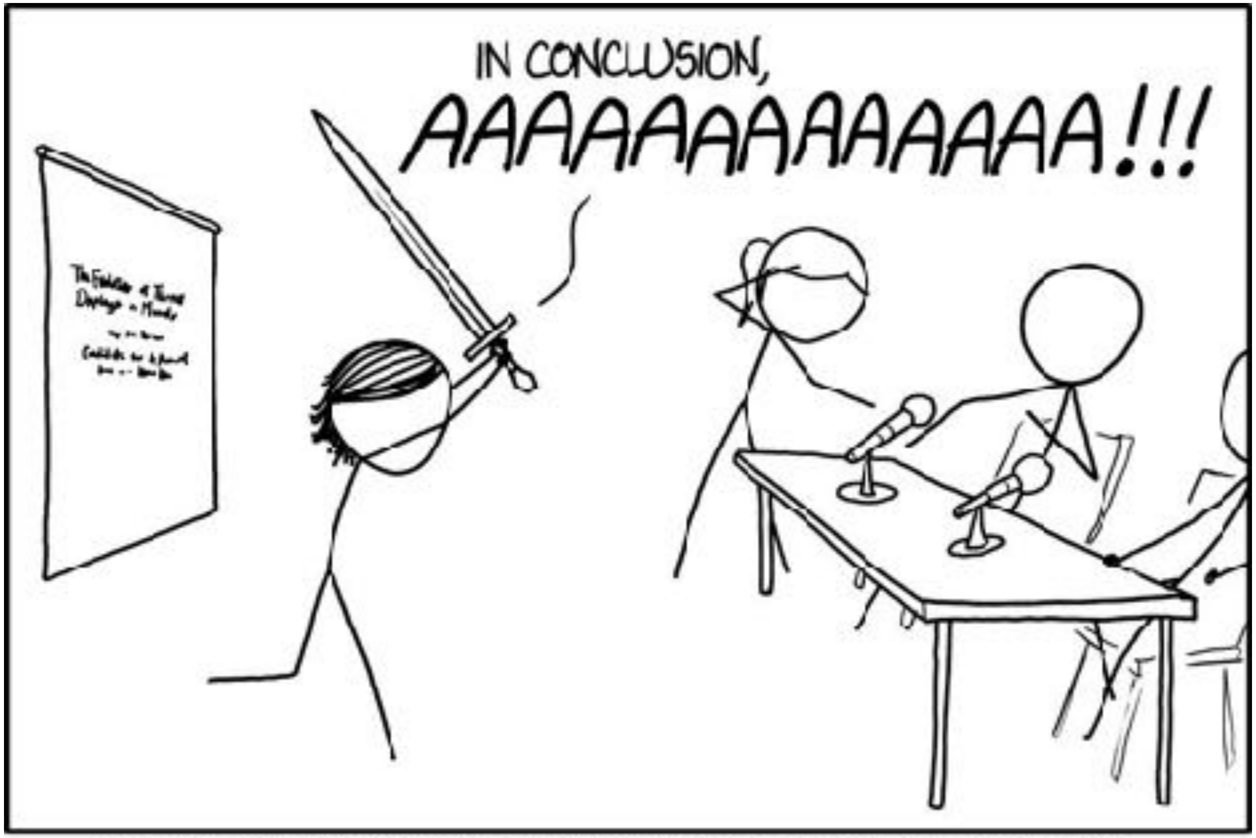
Canvas Syllabus

- Lab Assignments
- Grading Rubrics
- Participation
- Course Schedule
- In-Class Assignments
- Difference between 5000 and 7000

Is this plagiarism in this class?

- Copying code/text from another source without citing it
  - A. Yes, plagiarism!
  - B. No, its fine!
- Copying code/text from another source, citing at the end of the assignment in a blanket statment (but not making it clear which part of the assignment was from another source)?
  - A. Yes, plagiarism!
  - B. No, its fine!
- Copying code, citing the source directly next to the code, and commenting on what parts were changed?
  - A. Yes, plagiarism!
  - B. No, its fine!
- Copying text directly and citing the source with the text, but not placing the text in quotes.
  - A. Yes, plagiarism!
  - B. No, its fine!

Machine Learning Overview



THE BEST THESIS DEFENSE IS A GOOD THESIS OFFENSE.

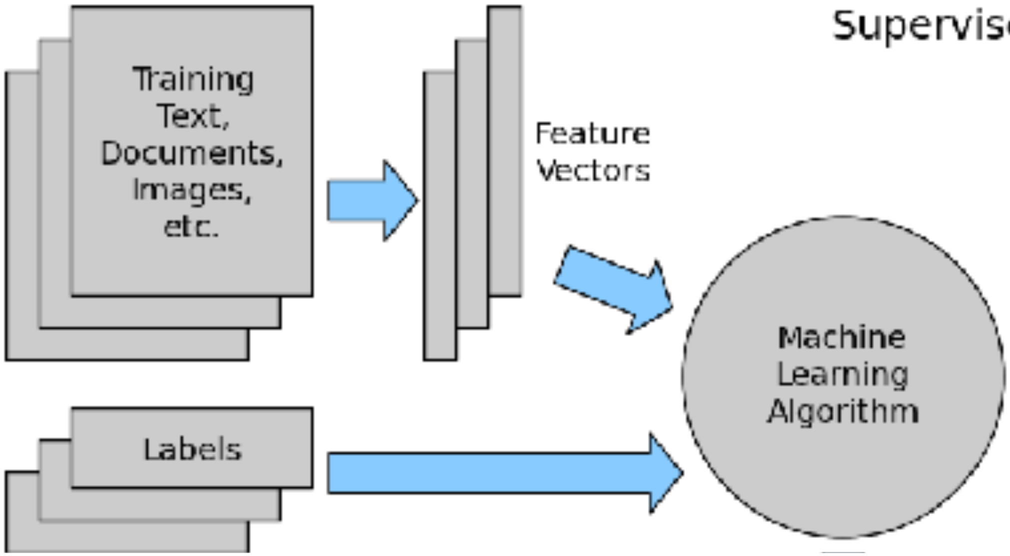
What is Machine Learning?

> **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can change when exposed to new data.
>
> What is machine learning? - Definition from WhatIs.com
> whatis.techtarget.com/**definition/machine-learning**

About this result • Feedback

- Beware:
  - full of imprecise words
  - words that play on our understanding of “learning” and consciousness

Classification: Definition



Supervised Learning Model

- *Training* Instances: Features + Labels
- Find a *model* mapping class from values of features.
- Goal: Assign class to <u>previously unseen</u> instances

Machine Learning

One Small Piece of Artificial Intelligence

Data Mining

ML

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
- Description Methods
  - Find human-interpretable patterns that describe the data.

ML

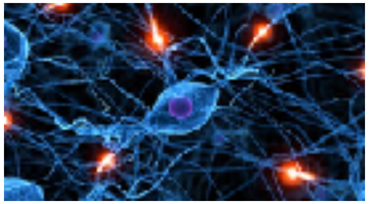
- Classification
- Regression
- Deviation Detection
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery

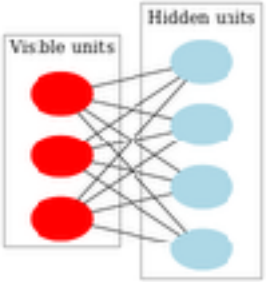
section 1, manipulated from Tan et al. Introduction to Data Mining

Abridged History of Machine Learning

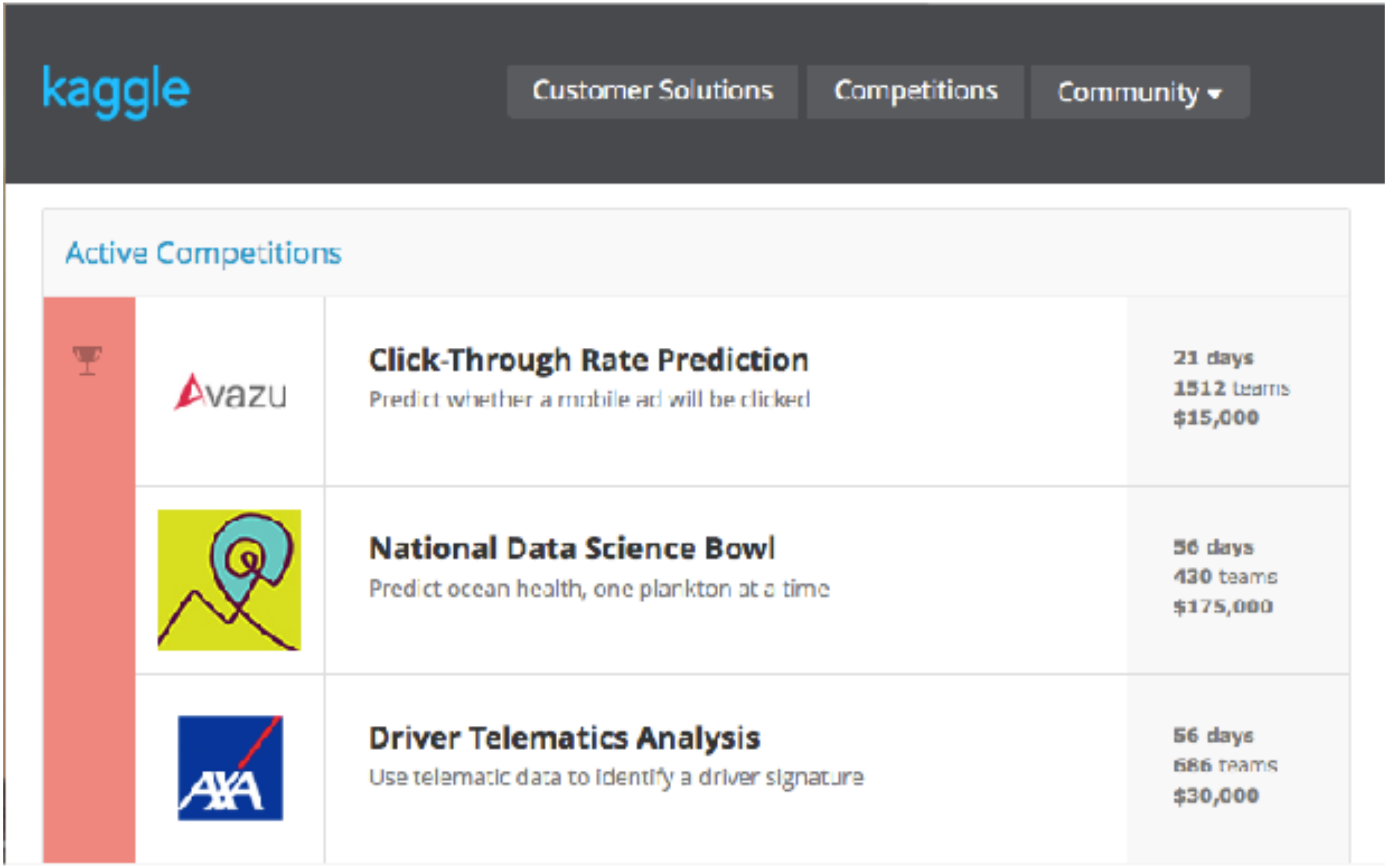
- Historically builds from disciplines statistics and computer science (algorithms)
- At present: Its really just algorithms for optimizing weights 🤠

- **1952**: Arthur Samuel IBM creates checker program
- **1957**: Rosenblatt, Neural Network Perceptron
- **1967**: Nearest Neighbor Pattern Recognition
- **1970's**: AI Winter
- **1990's**: Volley of "New" Machine learning Algorithms
- **2001**: Breiman's Random Forests
- ~**2004**: Modern Support Vector Machines with Kernels
- **2005**: Second AI Winter
- ~**2010**: Deep Learning Convolutional Networks
- **2015**: Deep Learning becomes buzz word,
- **Modern Day**: you hear about it and take this course





Problem Types in Machine Learning



Example Classification: Malware

- Classify files as malware based on structure, size, and naming.
- Approach:
  - Use already classified malware files
  - Must translate name to set of features
  - ***{malware, not malware}*** decision forms the ***class attribute***
  - Collect various malware examples and a number of safe files, providing labels for each and a set of features

Training Set

TID	Name	Size	Class
1	erte.dll	916 b	not
2	fufu.bin	1M	yes
3	exe.exe	1G	not
4	ex.py	113 b	not

Unknown

TID	*Name*	*Size*
1	asdf.dll	11b

Example Regression: Housing Price

- Predict a value of a given *continuous valued* variable based on the values of other variables
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Predicting House Sales

Training Set

TI	# Rms	Sq Ft	Zip	Price
1	2	1125	74012	150K
2	2	2525	75155	200k
3	10	4678	90210	3M
4	4	2678	75154	350k

Unknown

TI	# Rms	Sq Ft	Zip
1	2	2200	75115

Example Classifying: Objects in Images

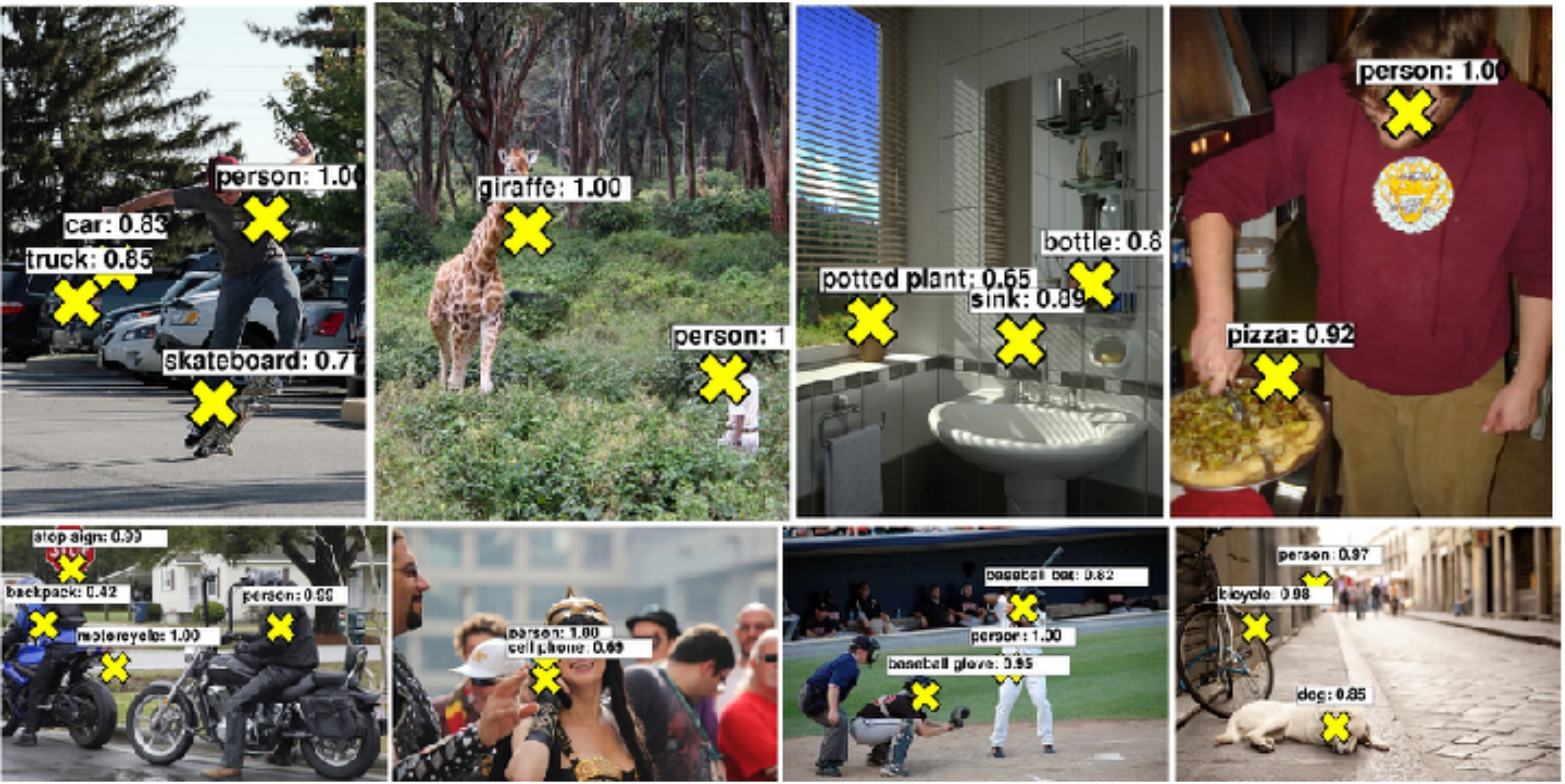


Image Net:

- 14 million images
- 200 Labeled Categories
- 1000 Location Labels

Attributes:

- Images

Self Test

- **A. Classification**
  B. Regression
  C. Not Machine Learning

- Dividing up customers by potential profitability?

- Extracting frequency of sound?

Types of Data and Categorization

Table Data

- Table Data: Collection of data **instances** and their **features**
- **Python:** Pandas Dataframe
- **R:** Data.frame
- **Matlab:** Table
- **C++**: Trick Question

Attributes, columns, variables, fields, characteristics, **Features**

Objects, records, rows points, samples, cases, entities, **Instances**

TID	*Pregnant*	*BMI*	*Age*	*Diabetes*
1	Y	33.6	41-50	positive
2	N	26.6	31-40	negative
3	Y	23.3	31-40	positive
4	N	28.1	21-30	negative
5	N	43.1	31-40	positive
6	Y	25.6	21-30	negative
7	Y	31.0	21-30	positive
8	Y	35.3	21-30	negative
9	N	30.5	51-60	positive
10	Y	37.6	51-60	positive

section 2, from Tan et al. Introduction to Data Mining

Feature Type Representation

	Attribute	Representation Transformation	Comments
Discrete	**Nominal**	Any permutation of values **one hot encoding**	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., new_value = f(old_value) where f is a monotonic function. **integer**	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Continuous	**Interval**	new_value =a * old_value + b where a and b are constants **float**	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	new_value = a * old_value **float**	Length can be measured in meters or feet.

Self Test

- Are these **A. ordinal, B. nominal, or C. binary**?
  - military rank
    - ordinal
  - coat check number
    - nominal

Before Next Lecture

- Before next class:
    - install python on your laptop
    - install anaconda distribution of python

- Look at Python primer if you need review
    - I made ~4 hours of YouTube content…
    - https://www.youtube.com/playlist?list=PL7IPdRN5E0YKCnVI-fvx8jOOCWVeGTsrV

Demo

If time:
Jupyter Notebooks

`01_Numpy and Pandas Intro.ipynb`

Lecture Notes for

Machine Learning in Python

Professor Eric Larson

Numpy, Pandas, Document Features

Class Logistics and Agenda

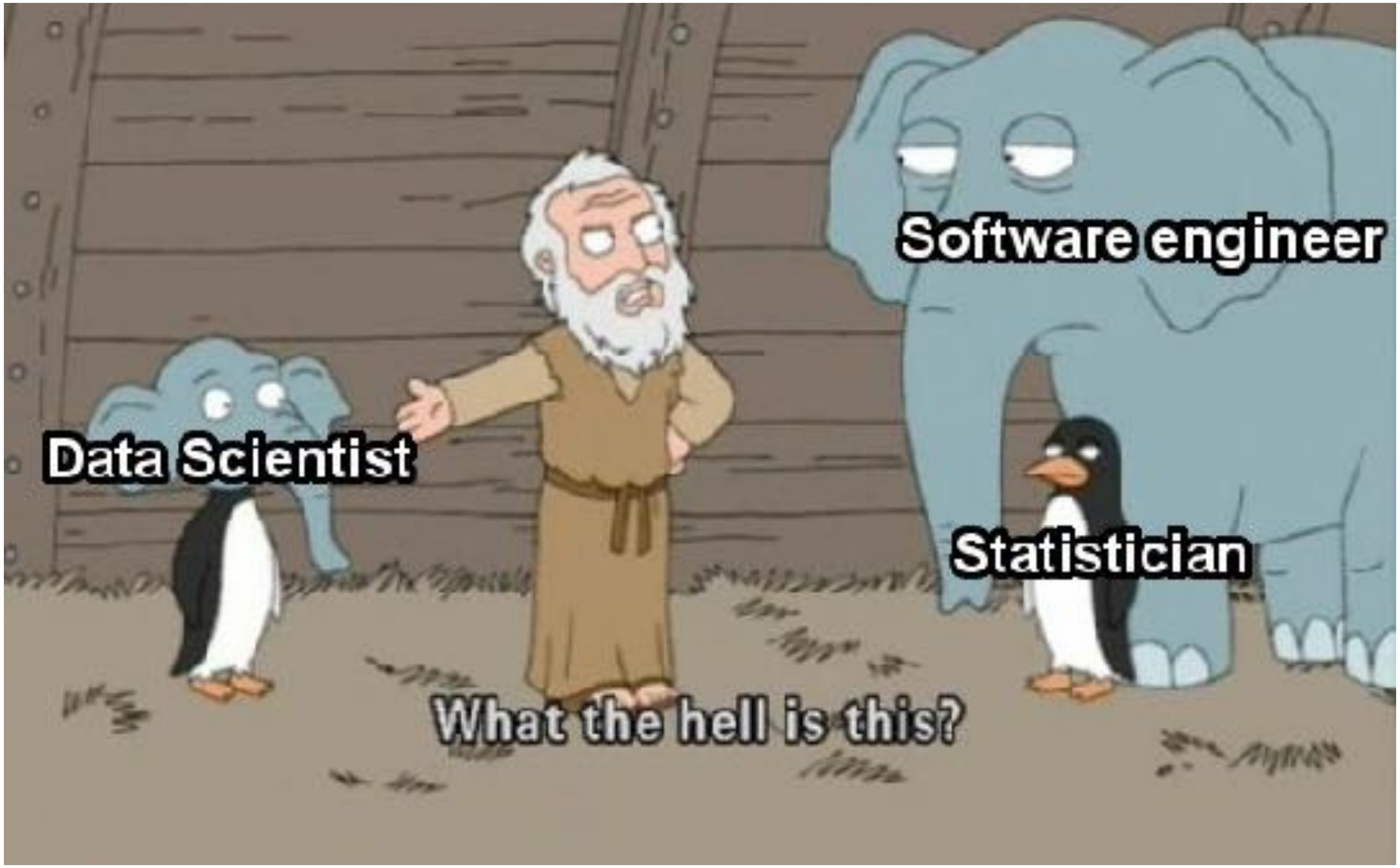
- Canvas? Anaconda Installs?
- Distance transfers?
- Agenda:
  - Numpy
  - Data Quality
  - Attributes Representation
    - documents
  - The Pandas eco-system
    - loading and manipulating attributes
- Needing some more help?
  - **fast.ai** has great, free resources

Demo

“Finish”
Jupyter Notebooks
and Numpy

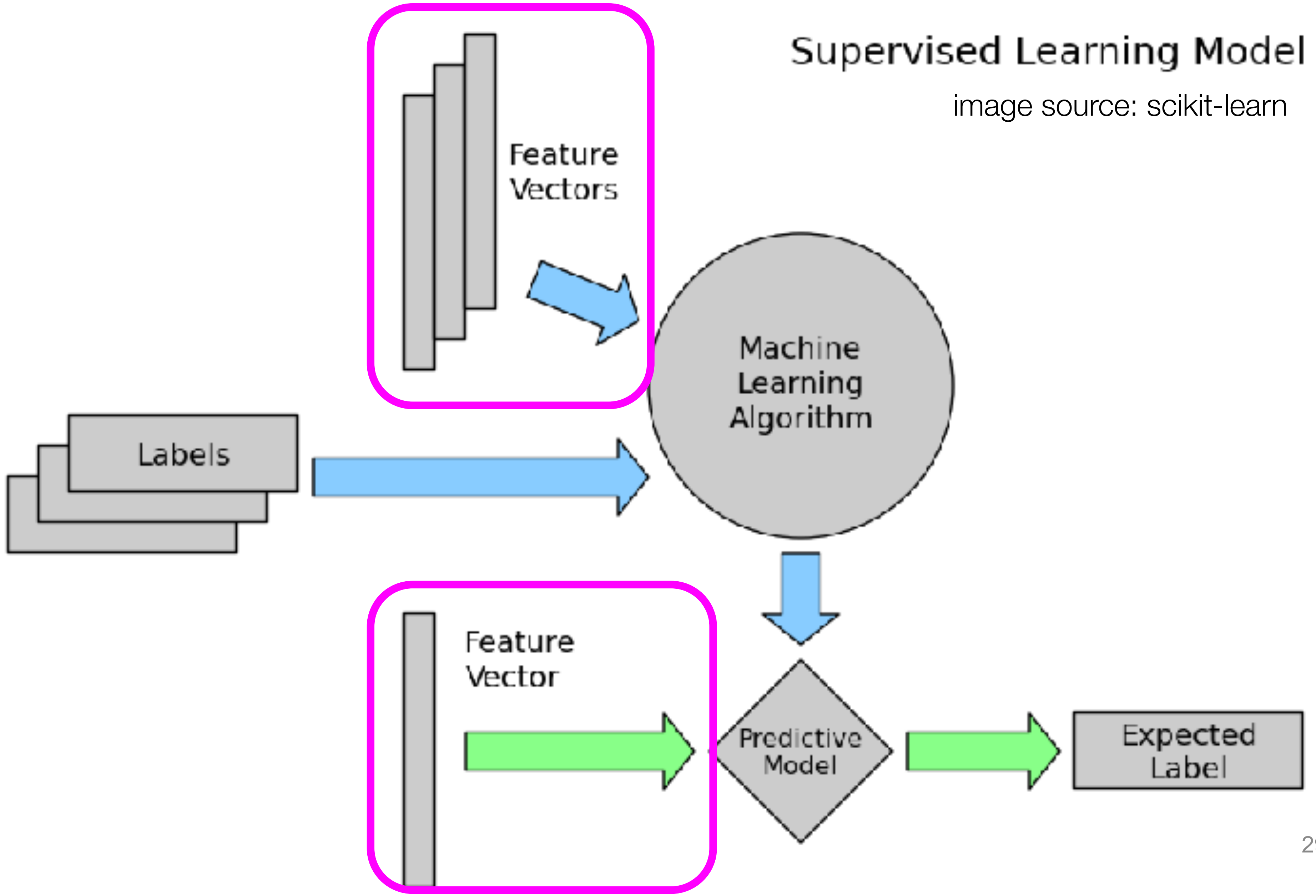
`01_Numpy and Pandas Intro.ipynb`

Data Quality



Review of Feature Data

image source: scikit-learn



Data Quality Problems

- Missing
  - Easy to find, NaNs
- Duplicated
  - Easy to find, hard to verify
- Noise or Outlier
  - Hard to define
  - Hard to catch

Information is not collected (e.g., people decline to give their age and weight)

Features **not applicable** (e.g., annual income for children)

UCI ML Repository: 90% of repositories have missing data

TID	Hair Color	Height	Age	Arrested
1	Brown	5'2"	23	no
2	Hazel	1.5m	12	no
3	Bl	5	999	no
4	Brown	5'2"	23	no

Handling Issues with Data Quality

- **Eliminate** Instance or Feature
- **Ignore** the Missing Value During Analysis Replace with all possible values (talk about later)
- **Impute** Missing Values

How?

Stats?
mean
median
mode

Imputation

- When is it probably fine to impute missing data:
  - (A) When there is not much missing data
  - (B) When the missing feature is mostly predictable from another feature
  - (C) When there is not much missing data for each subgroup of the data
  - (D) When it is the class you want to predict

Split-Impute-Combine

TID	Pregnant	BMI	Age	Diabetes
1	Y	33.6	41-50	positive
2	N	26.6	31-40	negative
3	Y	23.3	**?**	positive
4	N	28.1	21-30	negative
5	N	43.1	31-40	positive
6	Y	25.6	21-30	negative
7	Y	31.0	21-30	positive
8	Y	35.3	**?**	negative
9	N	30.5	51-60	positive
10	Y	37.6	51-60	positive

split: pregnant
split: BMI > 32

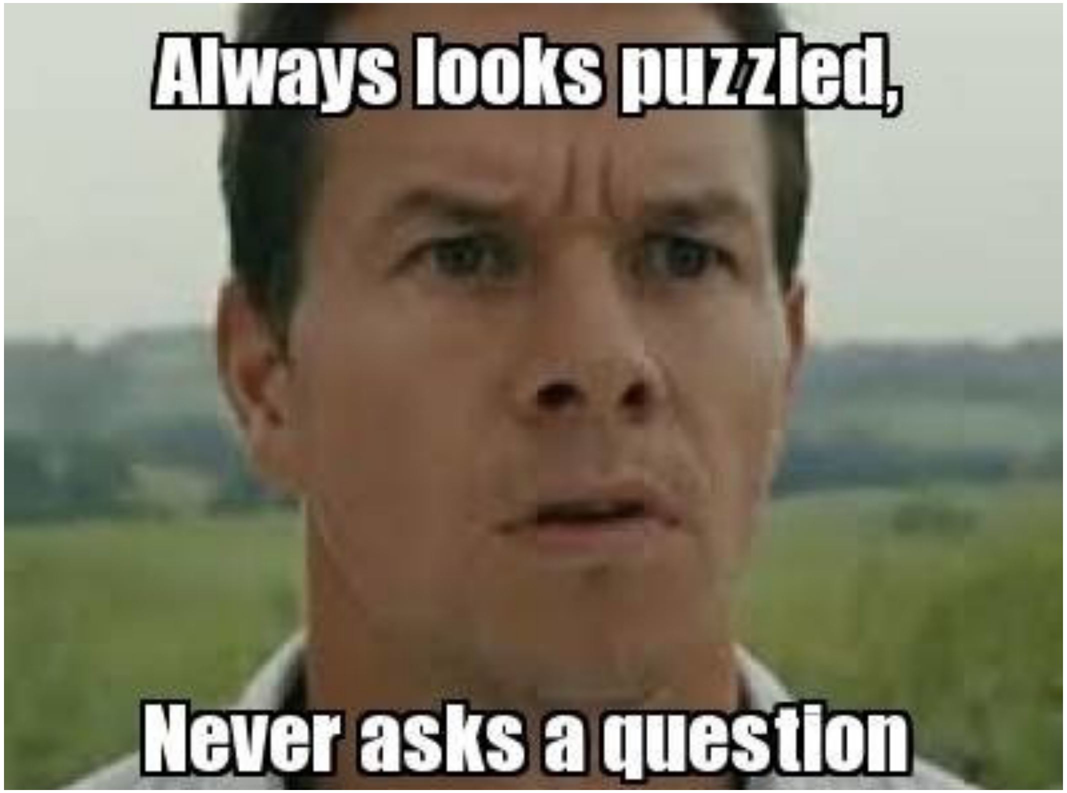
TID	Pregnant	BMI	Age	Diabetes
1	Y	>32	41-50	positive
8	Y	>32	**?**	negative
10	Y	>32	51-60	positive

Mode: none, can't impute

TID	Pregnant	BMI	Age	Diabetes
3	Y	<32	**?**	positive
6	Y	<32	21-30	negative
7	Y	<32	21-30	positive

Mode: 21-30

Data Representation and Documents



Feature Type Representation Review

	Attribute	Representation Transformation	Comments
Discrete	**Nominal**	Any permutation of values one hot encoding	If all employee ID numbers were reassigned, would it make any difference?
Discrete	**Ordinal**	An order preserving change of values, i.e., new_value = f(old_value) where f is a monotonic function. integer	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Continuous	**Interval**	new_value =a * old_value + b where a and b are constants float	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Continuous	**Ratio**	new_value = a * old_value float	Length can be measured in meters or feet.

Data Tables as Variable Representations

Table

TID	Pregnant	BMI	Age	Eye Color	Diabetes
1	Y	33.6	41-50	brown	positive
2	N	26.6	31-40	hazel	negative
3	Y	23.3	31-40	blue	positive
4	N	28.1	21-30	brown	inconclusive
5	N	43.1	31-40	blue	positive
6	Y	25.6	21-30	hazel	negative

Internal Rep.

TID
1
2
3
4
5
6

Data Tables as Variable Representations

Table

TID	Pregnant	BMI	Age	Eye Color	Diabetes
1	Y	33.6	41-50	brown	positive
2	N	26.6	31-40	hazel	negative
3	Y	23.3	31-40	blue	positive
4	N	28.1	21-30	brown	inconclusive
5	N	43.1	31-40	blue	positive
6	Y	25.6	21-30	hazel	negative

Internal Rep.

TID	Binary	Float	Ordinal	Object	Diabetes
1	1	33.6	2	hash(0)	1
2	0	26.6	1	hash(1)	0
3	1	23.3	1	hash(2)	1
4	0	28.1	0	hash(0)	2
5	0	43.1	1	hash(2)	1
6	1	25.6	0	hash(1)	0

Bag of words model

TID	Pregnant	BMI	Chart Notes	Diabetes
1	Y	33.6	Complaints of fatigue wh…	positive
2	N	26.6	Sleeplessness and some…	negative
3	Y	23.3	First saw signs of rash o…	positive
4	N	28.1	Came in to see Dr. Steve…	inconclusive
5	N	43.1	First diagnosis for hospit…	positive
6	Y	25.6	N/A	negative

Vocabulary

Bag of Words

TID	Sleep	Fatigue	Weight	Rash	First	Sight
1	0	1	0	0	2	0
2	1	1	0	0	1	1
3	1	1	0	2	1	1

number of occurrences

Feature Hashing

what happens when we get more words?

TID	Slee	Fati	Wei	Ras	First	Sigh	Why	Fox	Bro	Lazy	Dog	Etc	Stev
1	0	1	0	0	2	0	0	0	0	1	0	2	0
2	1	1	0	0	1	1	0	0	4	0	1	3	0
3	1	1	0	2	1	1	1	0	1	0	0	1	0

or we could have a hashing function, h(x) = y

TID	h(x)=1	h(x)=2	h(x)=3	h(x)=4	h(x)=5	h(x)=6
1	0	1	0	1	2	0
2	1	1	4	0	2	1
3	2	1	1	2	1	1

multiple words mapped to one feature

(want to minimize collisions or make collisions meaningful)

Term-Frequency, Inverse-Document-Frequency

TID	Slee	Etc	Stev
1	0	.86	0
2	0.1	.02	0
3	0.1	.1	0

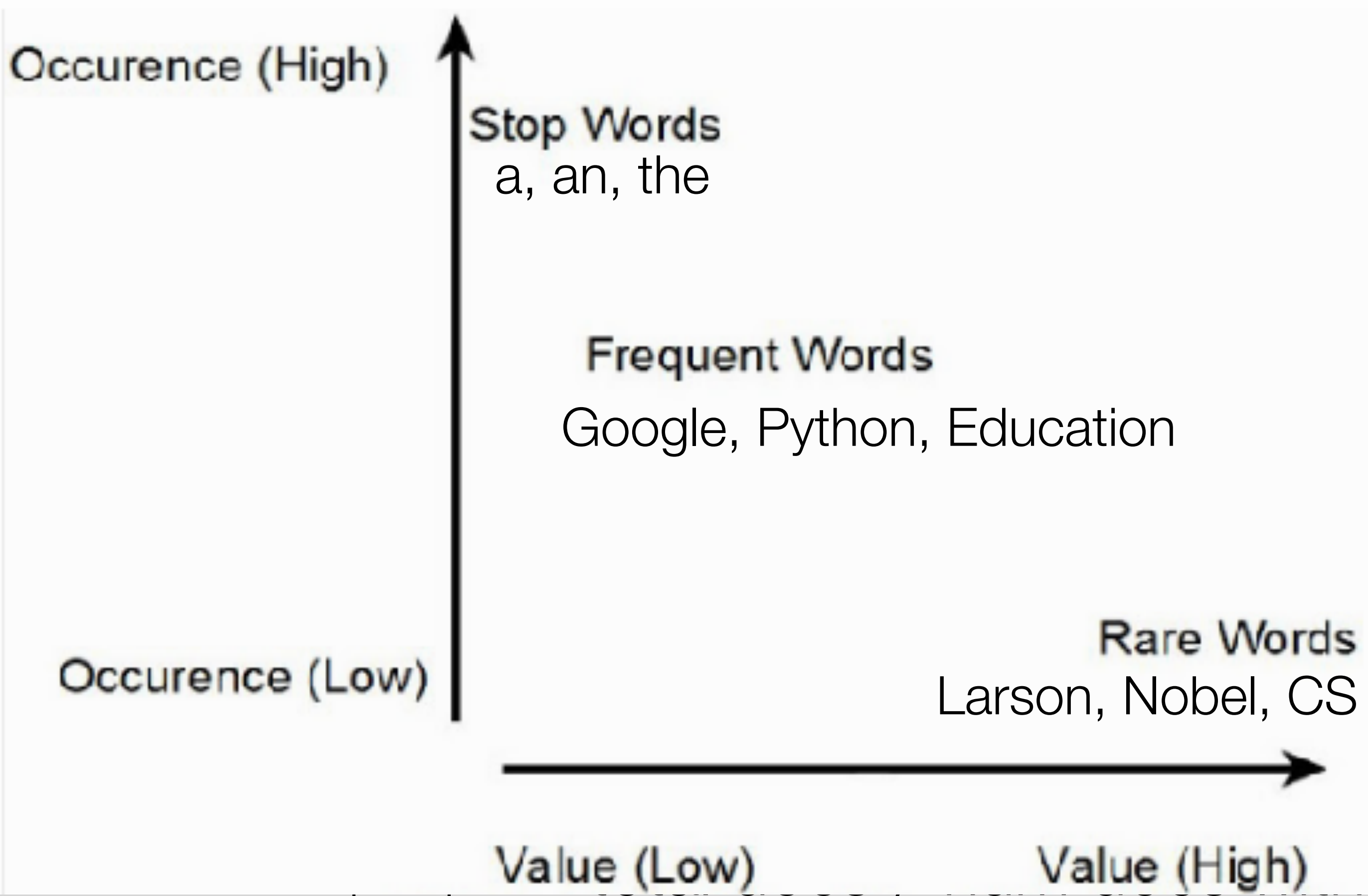
te $\in D$

in d”

inverse

idf($\in d$

t”



https://www.kaggle.com/divsinha/sentiment-analysis-countvectorizer-tf-idf

$$\text{tf-idf}(t,d) = \text{tf}(t,d)\cdot\text{idf}(t,d)$$

$$\text{tf-idf}(t,d) = \text{tf}(t,d)\cdot(1+\text{idf}(t,d))$$ smoothed

TF-IDF

- The tf-idf value can never be greater than one.
  - (A) true
  - (B) false
  - (C) it depends on IDF normalization

term frequency $\mathrm{tf}(t,d) = f_{td},\ t \in T \text{ and } d \in D$

"num occurrences of t in doc d"/"words in d"

inverse document frequency: normalize occurrences

$$\mathrm{idf}(t,d) = \log\frac{|D|}{|n_t|},\ \text{where } n_t = d \in D \text{ with } t \in d$$

"total docs"/"num docs with t"

$$\text{tf-idf}(t,d) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,d)$$

Demo

TF-IDF
DataFrames
Loading
Indexing
Imputing

`02_Document Feature Engineering.ipynb`

Other Tutorials:

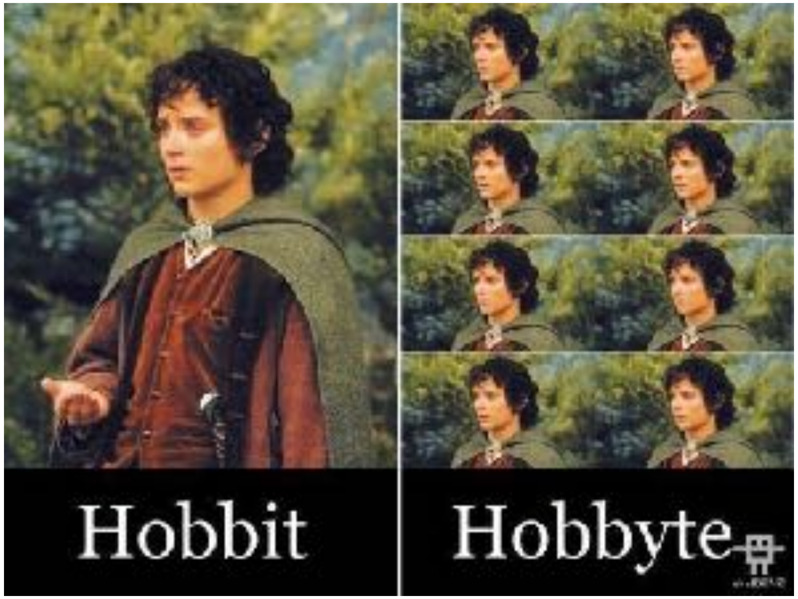
http://vimeo.com/59324550

http://pandas.pydata.org/pandas-docs/version/0.15.2/tutorials.html

For Next Lecture

- Before next class:
  - install seaborn
  - install plotly (or bokeh if you want)
  - look at pandas table data and look at additional tutorials

- Next Week: Data Visualization


Hobbit
Hobbyte


Dear Valued Customers:
Chicken Wings
& Cheesy Crust
Are Currently Out of Stock
Due to a Recent Cyberattack
Which Has Affected Imports
We Apologize
For The Inconvenience