
Lecture Notes for Machine Learning in Python

Professor Eric Larson
Introduction, Syllabus, Data Types

Class Logistics and Agenda

- Syllabus
- Overview of Machine Learning
- Types of Data and Representation

Course Syllabus

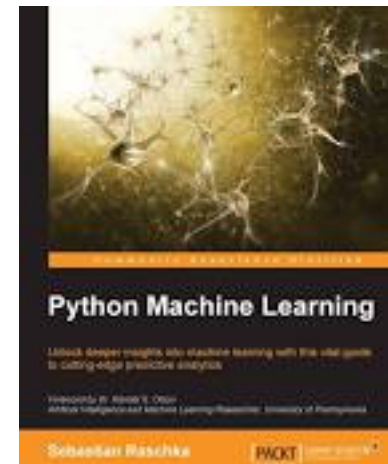


Introductions

- Me
 - Eric Larson
- You
 - Name, department, grad/ugrad
 - Something true or false
- My approach to this course
 - programming
 - math
 - **applications** and **analytics**

FAQ

- Text: None
 - Recommended: Python Machine Learning, Sebastian Raschka
- Use Canvas for posted course material
- Prerequisite: or ability to learn quickly these topics
 - Linear Algebra, Calculus
 - Basic statistics and probability
 - Python programming
- Version of python: 3.X
 - install through anaconda
 - use virtual environments
- Deep Learning Library: Keras over Tensorflow



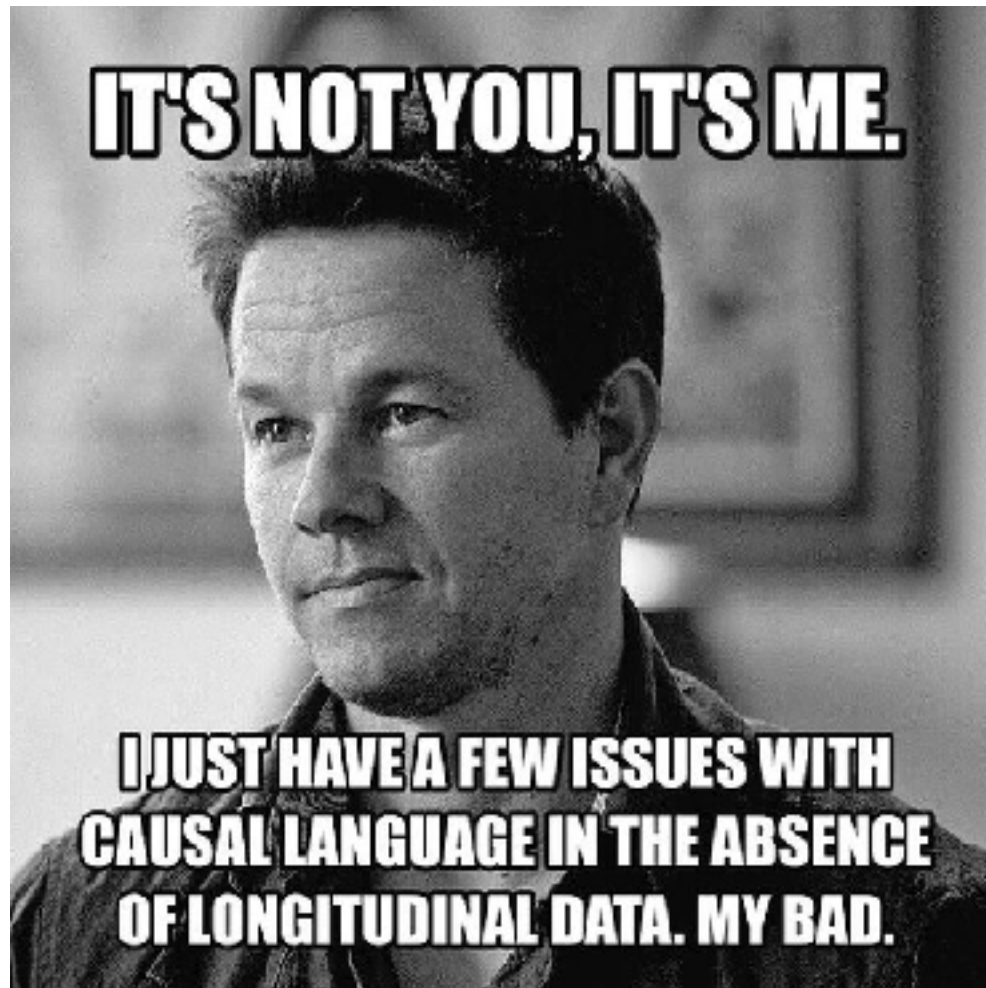
How will you grade participation?

- Participation will be graded in the course:
 - Distance students will answer these questions via canvas upload
 - must upload the questions throughout semester for full credit
- 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: Wait, what...

Canvas Syllabus

- Assignments
- Grading Rubrics
- Participation
- Course Schedule
- In-Class Assignments

Machine Learning Overview

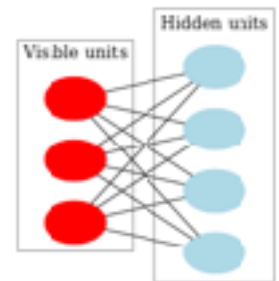
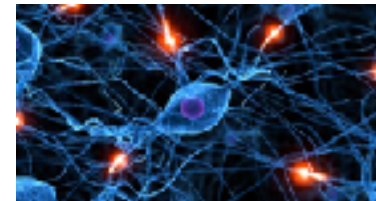


A History of Machine Learning

- Historically builds from disciplines statistics and computer science (algorithms)
- 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
- ~**2010:** Deep Learning Convolutional Networks
- **2015:** Deep Learning becomes buzz word, you hear about it and take this course

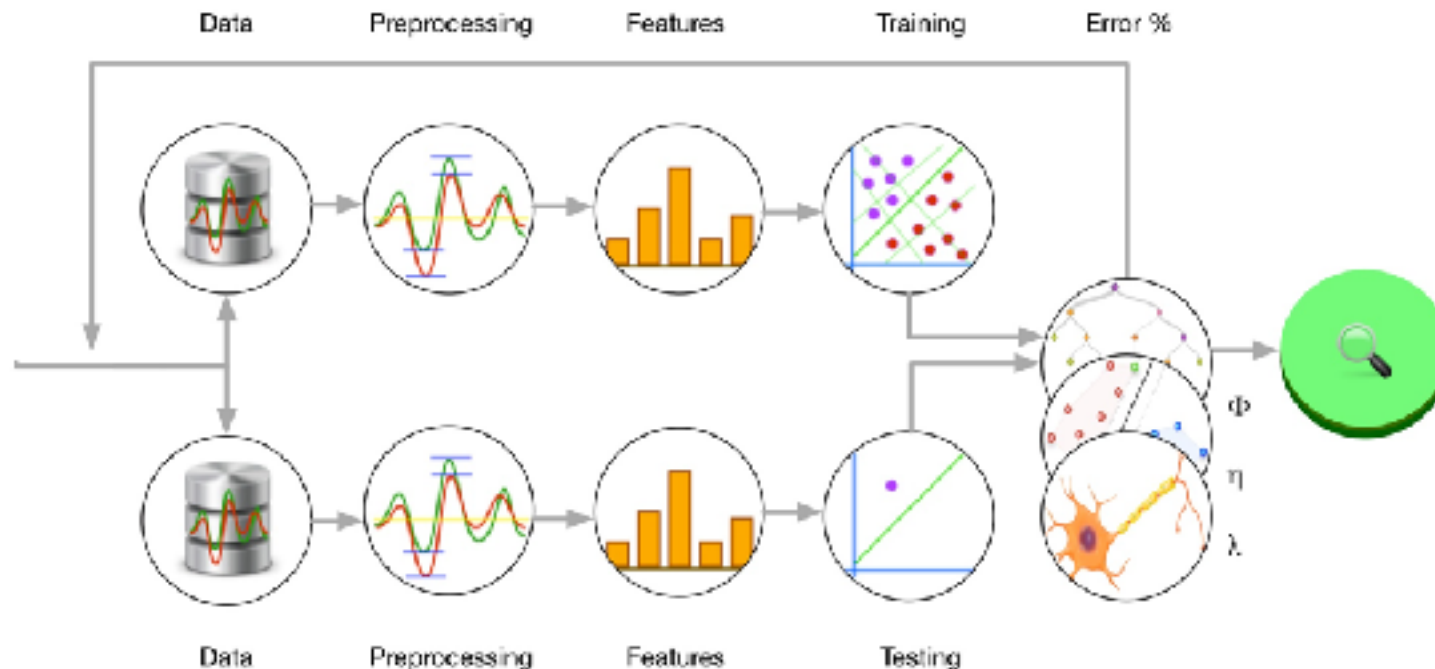


What is Machine Learning?

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](https://www.whatis.techtarget.com/definition/machine-learning)

About this result • Feedback



Machine Learning is part of Data Mining

- Prediction Methods

- Use some variables to predict unknown or future values of other variables

- Description Methods

- Find human-interpretable patterns that describe the data.

- Classification [Predictive]

- Regression [Predictive]

- Deviation Detection [Predictive]

- Clustering [Descriptive]

- Association Rule Discovery [Descriptive]

- Sequential Pattern Discovery [Descriptive]

Problem Types in Machine Learning





kaggle

Customer Solutions

Competitions

Community ▾

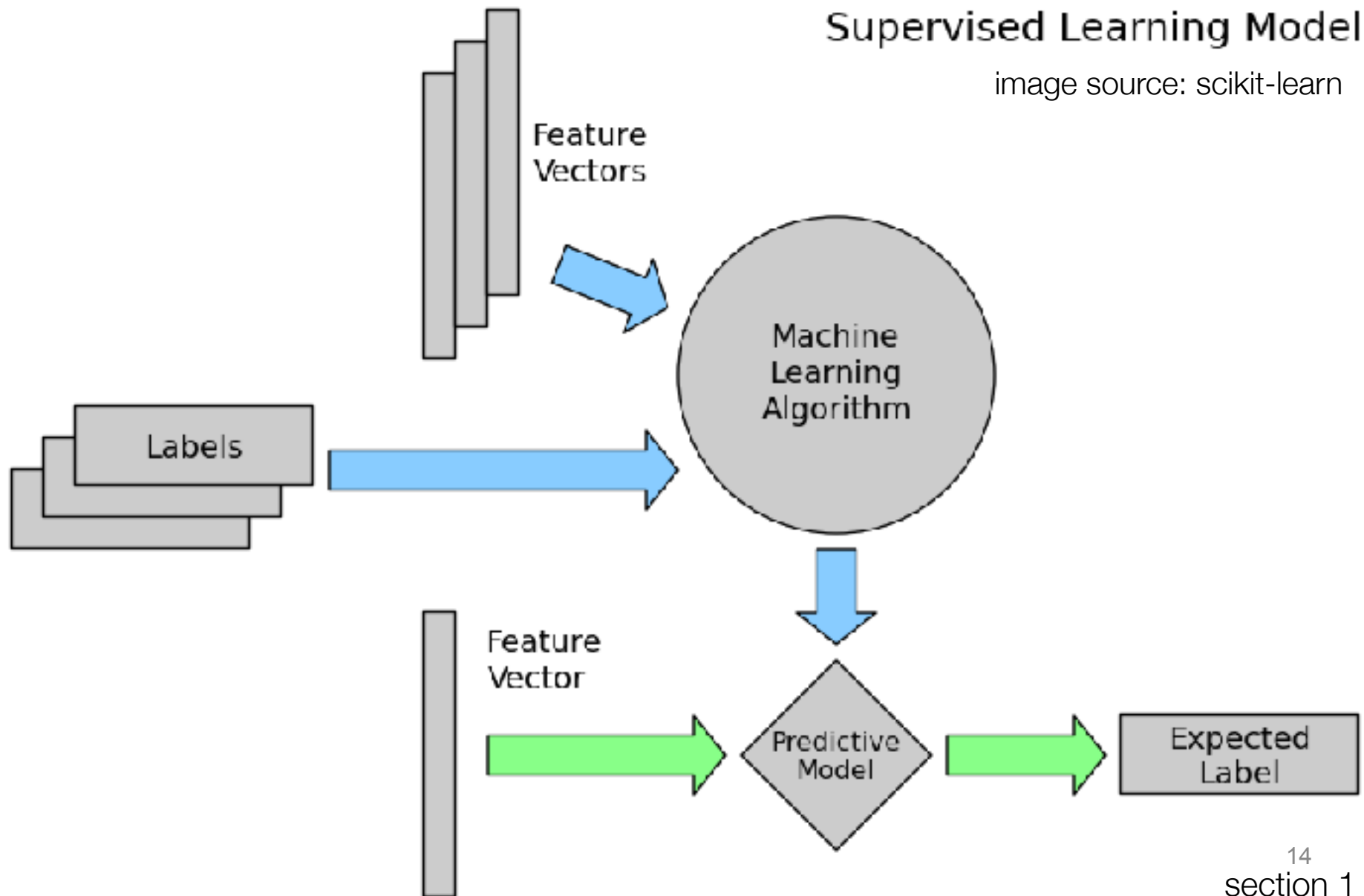
Active Competitions

		Click-Through Rate Prediction Predict whether a mobile ad will be clicked	21 days 1512 teams \$15,000
		National Data Science Bowl Predict ocean health, one plankton at a time	56 days 430 teams \$175,000
		Driver Telematics Analysis Use telematic data to identify a driver signature	56 days 686 teams \$30,000

Classification: Definition

- Given a collection of instances (*training set*)
 - Each instance contains a set of *features*, one of the features is the *class*.
- Find a *model* for class as a function of the values of features.
- Goal: previously unseen instances should be assigned a class as accurately as possible.

Classification: Definition



Classification: Malware

- Goal: classify files as malware based on structure, size, and naming.
- Approach:
 - ♦ Use already classified malware files.
 - ♦ *{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

<i>TID</i>	<i>Name</i>	<i>Size</i>
<i>1</i>	asdf.dll	11b

Classifying: Objects in Images

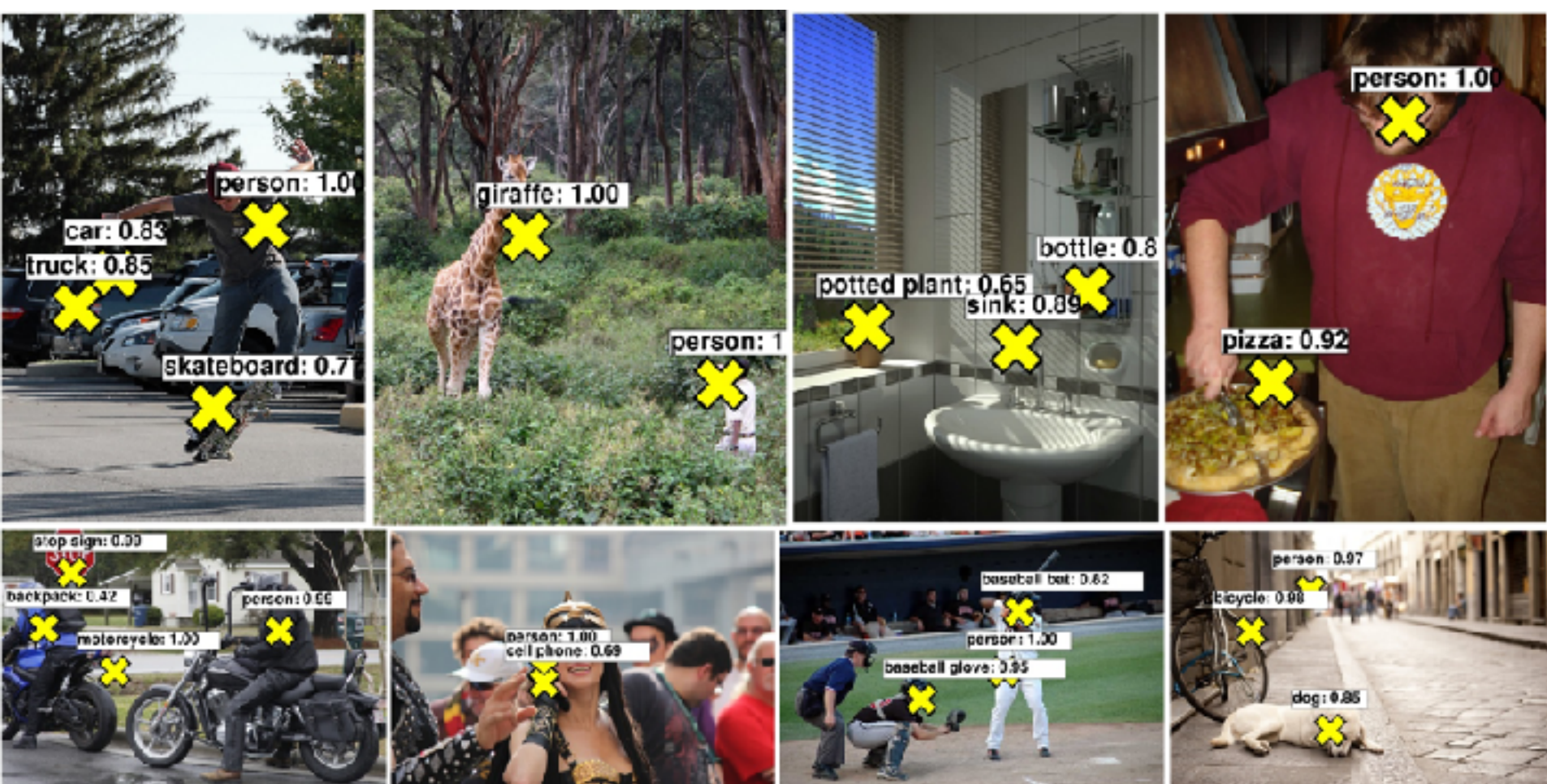


Image Net:

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

Attributes:

- Images

Regression

- 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 lung function as a function of gender, weight, height

Training Set

<i>TI</i>	<i>Gender</i>	<i>Weight</i>	<i>Asthma</i>	<i>LF</i>
1	M	175lbs	N	85%
2	F	150lbs	N	87.3%
3	F	155lbs	Y	90%
4	M	225lbs	Y	65.2%

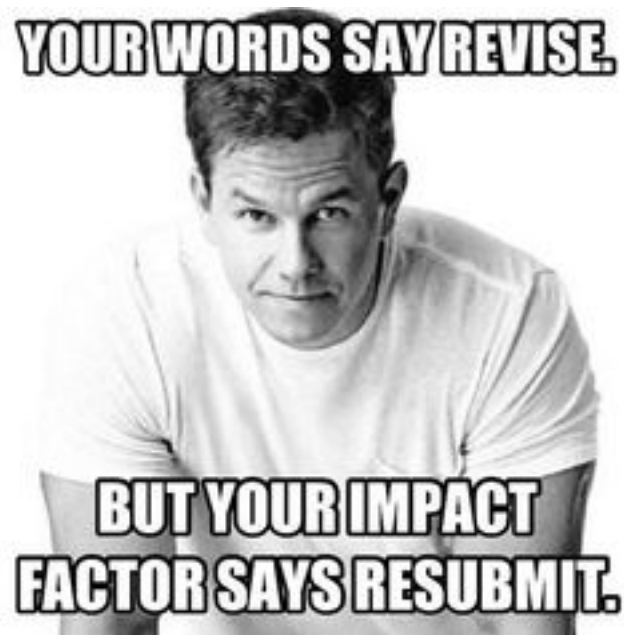
Unknown

<i>TI</i>	<i>Gender</i>	<i>Weight</i>	<i>Asthma</i>
1	M	160lbs	N

Self Test

- **(A. classification)**
- **(B. regression)**
- **(C. not Machine Learning)**
 - Dividing up customers by potential profitability?
 - classification/regression
 - Extracting frequency of sound?
 - NOT ML
 - Finding someone's adipose tissue measure from waist circumference?
 - regression
 - Deciding if a person has diabetes based upon their history and diet?
 - classification
 - Finding the genre of an online article based on the words in it?
 - classification

Types of Data and Categorization



What is Data?

- Collection of data **instances** and their **features**
- A **feature** is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
- A collection of features describe an **instance**

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

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

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>
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

Types of Attributes

- **Nominal**
- ◆ Examples: ID numbers, eye color, zip codes
- **Ordinal**
- ◆ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- **Interval**
- ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- **Ratio**
- ◆ Examples: temperature in Kelvin, length, time, counts

Distinctness: = ≠

Order: < >

Addition: + -

Multiplication: * /

Nominal attribute: distinctness

Ordinal attribute: distinctness & order

Interval attribute: distinctness, order, & addition

Ratio attribute: all properties

Attribute Type	Description	Examples	Operations
Nominal	The values are different names, i.e., only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {male, female}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {good, better, best}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

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., $\text{new_value} = f(\text{old_value})$ 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	$\text{new_value} = a * \text{old_value} + b$ 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	$\text{new_value} = a * \text{old_value}$ float	Length can be measured in meters or feet.

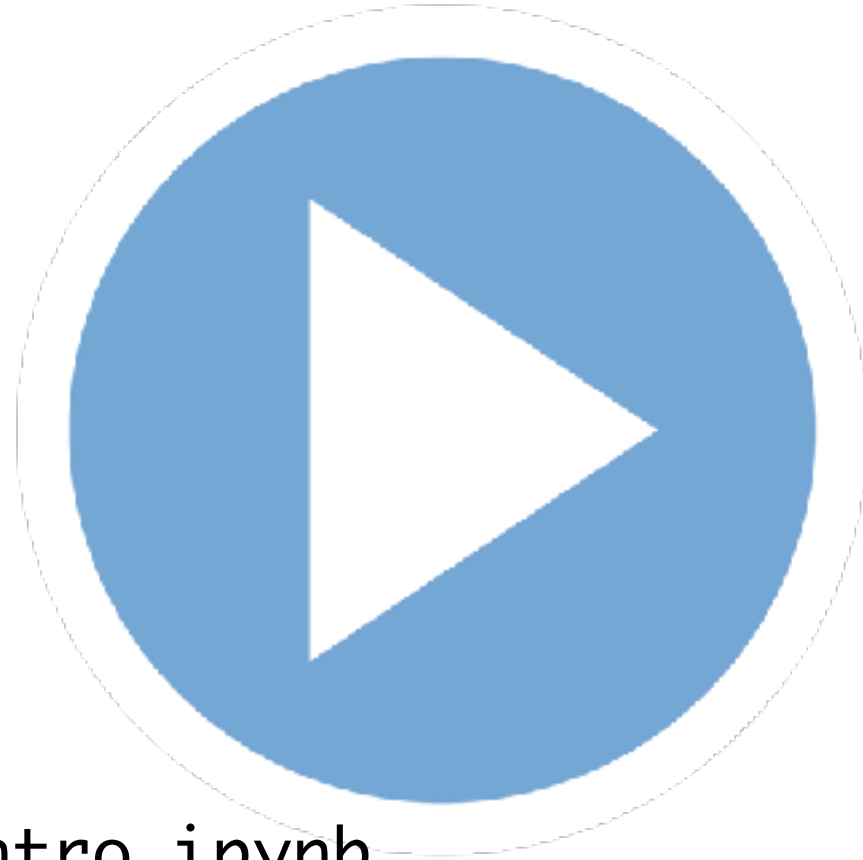
Self Test

- Are these **A. interval** or **B. ratio**:
 - Angle measured 0-360 degrees
 - ratio
 - Height above sea level
 - interval or ratio depending on if sea level is considered arbitrary
- Are these **A. ordinal**, **B. nominal**, or **C. binary**?
 - military rank
 - ordinal
 - coat check number
 - nominal
 - time as AM or PM
 - binary

Before Next Lecture

- Before next class:
 - install python on your laptop
 - install anaconda distribution of python
- Look at Python primer if you need an intro to Python
 - I made ~4 hours of YouTube content...
 - <https://www.youtube.com/playlist?list=PL7IPdRN5E0YKCnVI-fvx8jOOCWVeGTsrV>

**If time:
Jupyter Notebooks
and Numpy**



`01_Numpy and Pandas Intro.ipynb`

Lecture Notes for Machine Learning in Python

Professor Eric Larson
Numpy, Pandas, Document Features

Class Logistics and Agenda

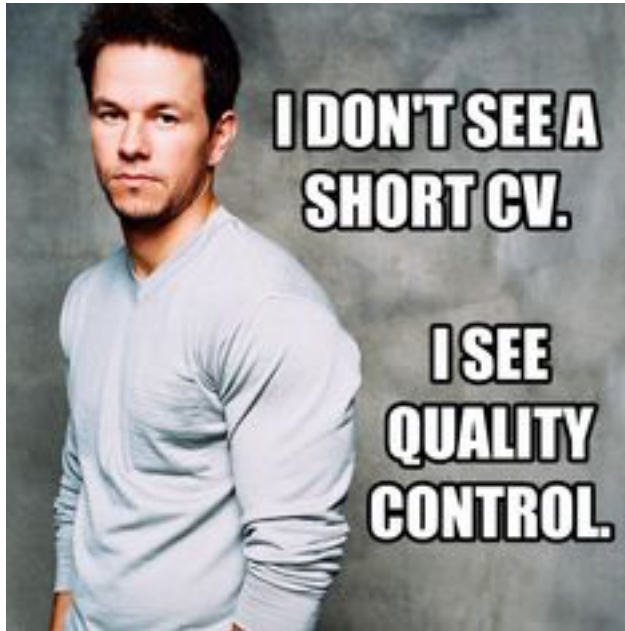
- Canvas Access?
- Anaconda Installs?
- Agenda:
 - Numpy
 - Data Quality
 - Attributes Representation
 - documents
 - The Pandas eco-system
 - loading and manipulating attributes

“Finish” Jupyter Notebooks and Numpy

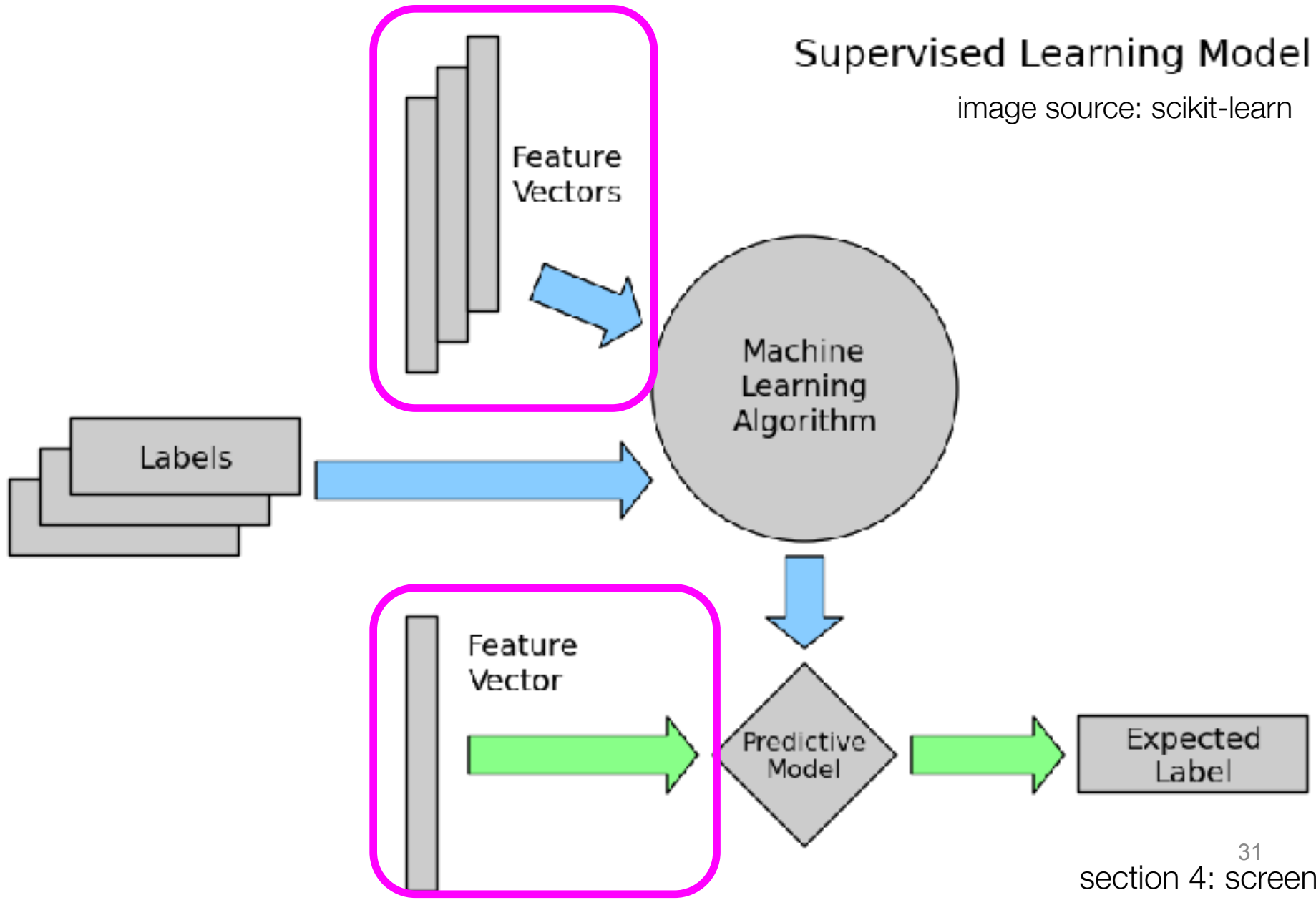


`01_numpy and Pandas Intro.ipynb`

Data Quality



Review of Feature Data



Data Quality Problems

- Noise and outliers
 - remove if you know its noise/outlier
- Missing values
 - replace or ignore
- Duplicate data
 - clean entries or merge

Missing Values

- Reasons for missing values
 - Information is **not collected**
(e.g., people decline to give their age and weight)
 - Features may **not be applicable** to all cases
(e.g., annual income for children)
 - **UCI ML Repository**: 90% of repositories have missing data
- Handling missing values
 - **Eliminate** Data Objects
 - **Impute** Missing Values
 - **Ignore** the Missing Value During Analysis
 - Replace with all possible values (talk about later)

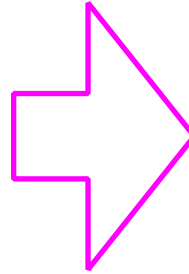
Stats:
mean
median
mode

How?

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



<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>
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

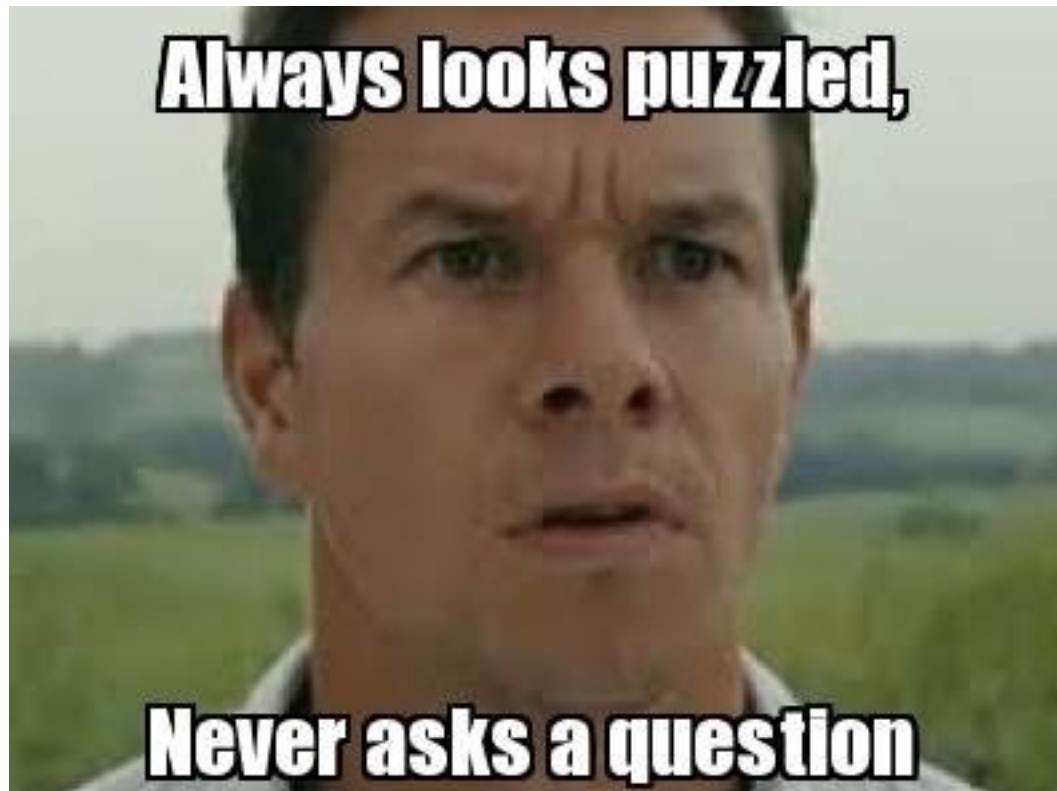
<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>
1	Y	>32	41-50	positive
8	Y	>32	?	negative
10	Y	>32	51-60	positive

Mode: none, can't impute

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

Mode: 21-30

Data Representation



Feature Type Representation

		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., $\text{new_value} = f(\text{old_value})$ 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	$\text{new_value} = a * \text{old_value} + b$ 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	$\text{new_value} = a * \text{old_value}$ float	Length can be measured in meters or feet.

Data Tables as Variable Representations

Table

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Eye Color</i>	<i>Diabetes</i>
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.

<i>TID</i>
1
2
3
4
5
6

Data Tables as Variable Representations

Table

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Eye Color</i>	<i>Diabetes</i>
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.

<i>TID</i>	<i>Binary</i>	<i>Float</i>	<i>Ordinal</i>	<i>Object</i>	<i>Diabetes</i>
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

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Chart Notes</i>	<i>Diabetes</i>
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

Bag of Words

Vocabulary						
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)

Term-Frequency, Inverse-Document-Frequency

Given a vocabulary of words:

TID	Slee	Fati	Wei	Ras	First	Sigh	Why	Fox	Bro	Lazy	Dog	Etc	Stev
1	0	0.05	0	0	0.34	0	0	0	0	1	0	0.86	0
2	0.1	0.05	0	0	0.12	0.25	0	0	1.21	0	1	1.02	0
3	0.1	0.05	0	0.27	0.12	0.25	0.02	0	0.45	0	0	0.1	0

term frequency $tf(t, d) = f_{td}$, $t \in T$ and $d \in D$
“num occurrences in doc”/“word in doc”

inverse document frequency: normalize occurrences

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

“num docs”/“num docs with t”

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

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

TF-IDF

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

term frequency $\text{tf}(t, d) = f_{td}$, $t \in T$ and $d \in D$
“num occurrences in doc”/“word in doc”

inverse document frequency: normalize occurrences

$\text{idf}(t, d) = \log \frac{|D|}{|n_t|}$, where $n_t = \{d \in D \mid t \in d\}$
“num docs”/“num docs with t”

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

Sklearn and Pandas

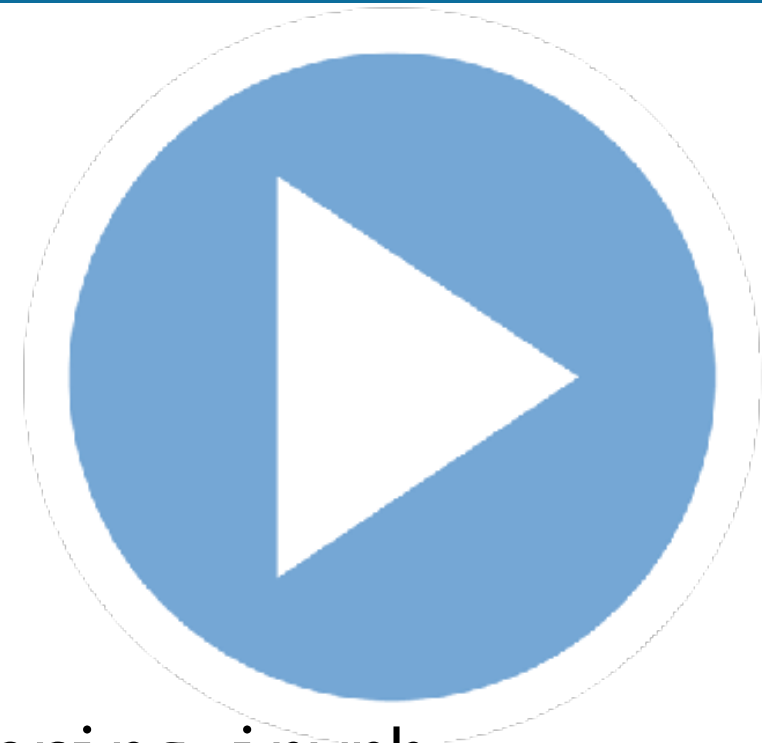
TF-IDF

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
 - mess with pandas and look at additional tutorials
- Next Week: Data Visualization and Dimensionality Reduction
- End of Next Week: **Lab One Due, Table Data**