Lecture Notes for

Machine Learning in Python

[[👨‍🏫, 👨‍💻, 🐍, 👨‍🔬]]

Professor Eric Larson

Data Quality and Imputation

Class Logistics and Agenda

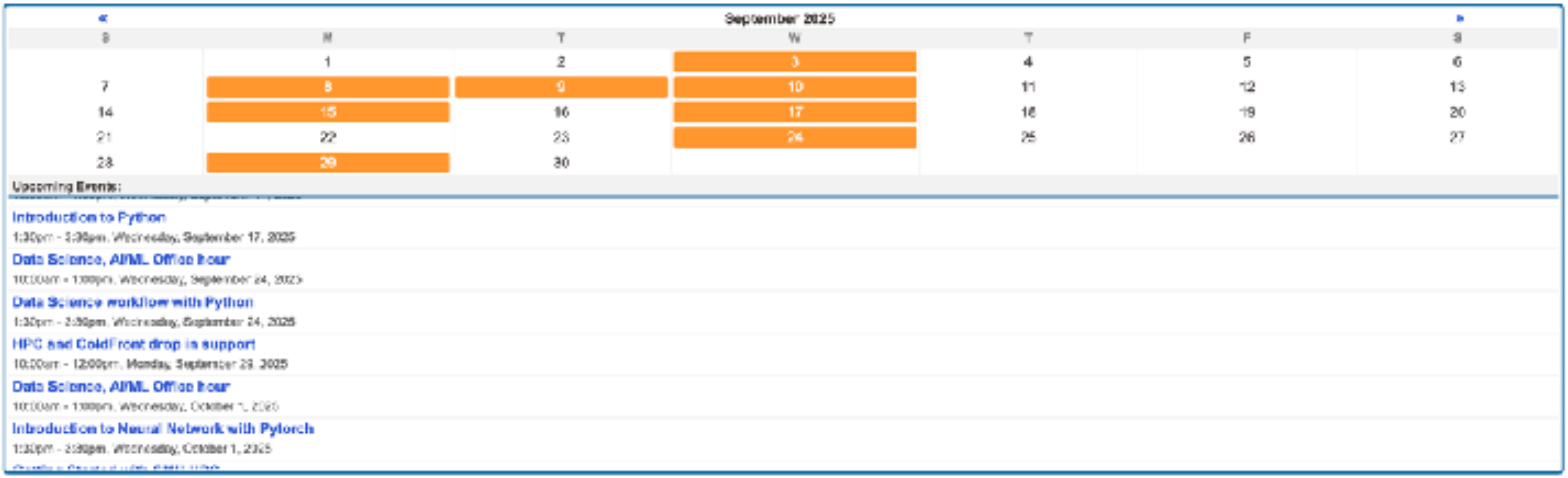
- Logistics:
  - Need help? canvas has links to various resources
    - the class GitHub is also a resource!
  - TA hours!
  - Team Forming Discussion sections
- Agenda:
  - Data Quality
  - Data Representations
  - Imputation methods

Course Github Page:	https://github.com/eclarson/MachineLearningNotebooks
Other Useful Guides:	Helpful Links and Guides for Semester
Participation For Distance Students	Turn in answers to questions here: Participation

Using the SMU HPC

- Tutorials available for various types of analysis:
    - https://www.smu.edu/oit/research

Events This Month



Class Overview, by topic

Table Data Visualization

Numpy, Pandas, Seaborn
Overviews with some in-depth discussion

Dimension Reduction and Image Processing

Scikit-learn, Scikit Image,
Intuition only, Some mathematics

Linear and Logistic Regression

Numpy, Recreate API for Scikit-learn
Detailed mathematics for simple optimization
intuition for advanced optimization

Neural Networks and Back Prop.

Numpy
Detailed mathematics for NN operations

Wide and Deep Networks

Convolutional Networks

Recurrent Networks

Keras, Tensorflow
Intuition, Detailed implement.

Ethics in Language Models

ConceptNet
Case studies

Last Time

Data Quality Problems

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

	Type	Description	Allowed Transforms
Discrete	Nominal or Categorical	Variable could be one value in a set of categories. No ordering of values. *Example: Employee ID*	**Allowed Transforms** permuting values **boolean, one hot encoding, or hash function**
Discrete	Ordinal	Variable could be one value in a set of categories. Ordering matters. *Example: Start Ratings, 1-5*	**Allowed Transforms** $V_{new} = f_{mono}(V_{old}) + b$ **integer (or boolean)**
Continuous	Interval or Numeric	Value is continuous numeric value. Could be in specified range. *Example: BMI, Temperature, etc.*	**Allowed Transforms** $V_{new} = f_{mono}(V_{old}) + b$ **float**
Continuous	Ratio or Numeric	Value is continuous numeric value. Zero is meaningful. Often not treated differently than interval. *Example: Length, Elevation*	**Allowed Transforms** $V_{new} = f_{mono}(V_{old})$ **float**

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

TID	Hair Color	Height	Age	Arrested
1	Brown	5'2"	23	no
2	Hazel	1.5m	12	no
3	Bl	5	929	no
4	Brown	5'2"	23	no

Self Test, Missing data

- Can all missing data be found by searching for NaNs?
  - A. Yes. Missing data should always be a NaN.
  - B. Yes. Pandas defaults all missing data to NaN.
  - C. No. This only works for floats, because that is the data type for NaN.
  - D. No. NaN only represents missing data that is already found.

K-Nearest Neighbors Imputation

TID	Pregnant	BMI	Age	Diabetes
1	Y	33.6	31-40	positive
2	N	26.6	31-40	negative
3	Y	23.3	?	positive
4	?	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	21-30	positive

$f_i^{(unk)}$

For $k = 3$, find 3 closest neighbors

TID	Preg.	BMI	Age	Diabetes	Distance d_k
3	Y	23.3	?	positive	0
6	Y	25.6	21-30	negative	(0 + 2.3 + 1)/3
2	N	26.6	31-40	negative	(1 + 3.3 + 1)/3
4	?	28.1	21-30	negative	(4.8 + 1)/2

… repeat for all rows, select 3 closest …

Imputed Age: 21-30

$$d_k = \frac{1}{|F_{valid}|} \sum_{i \in F_{valid}} \| f_i^{(unk)} - f_i^{(k)} \|$$

i^{th} feature, f, in row

Distance can be calculated differently:

- Difference for valid features only
- May need to normalize ranges
- Weight neighbors differently?
- Have min # of valid features?
- Type: Euclidean, city-block, etc.

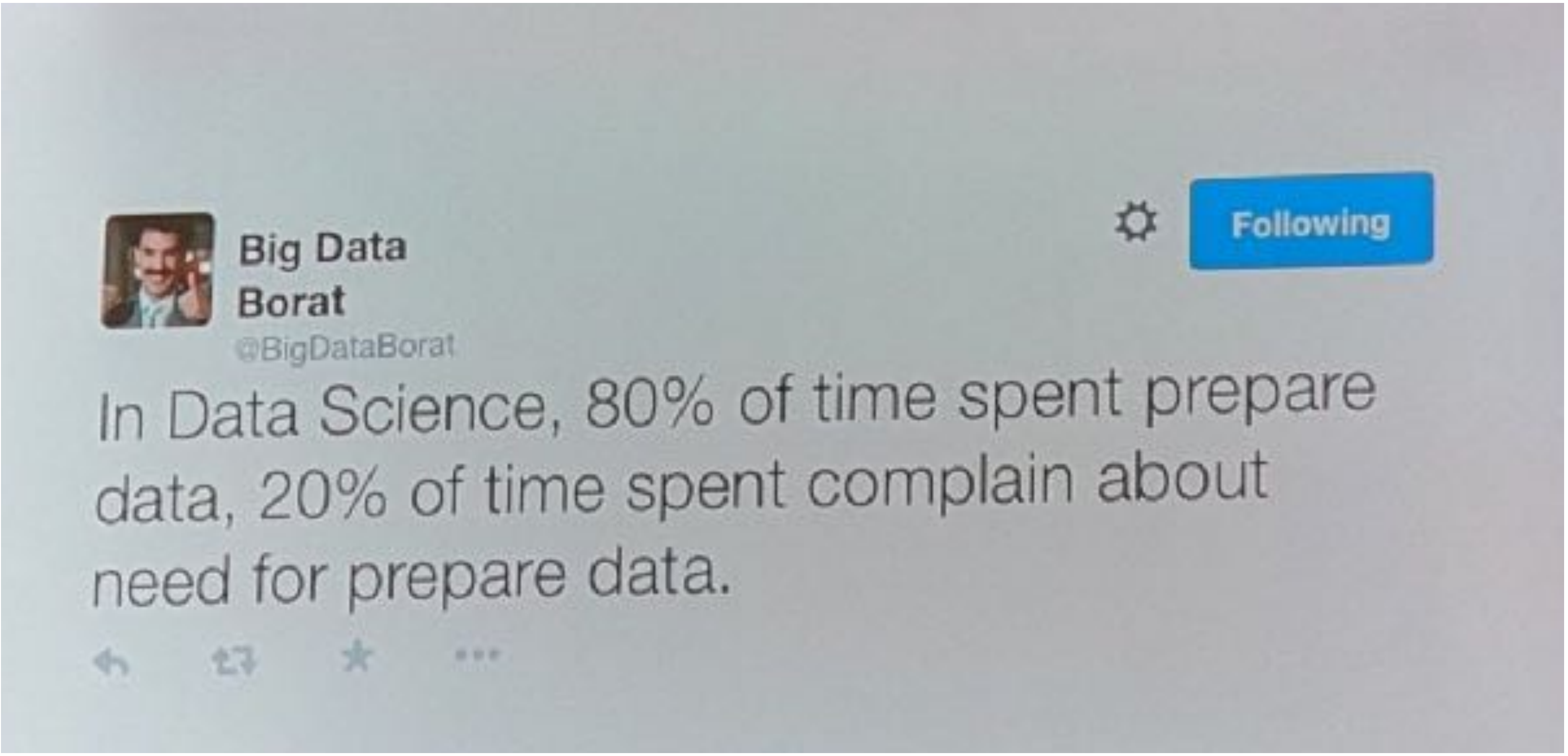
Demo

Pandas and Imputation

Scikit-Learn

`03. Data Visualization.ipynb`

Data Representation and Documents



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

Term-Frequency, Inverse-Document-Frequency

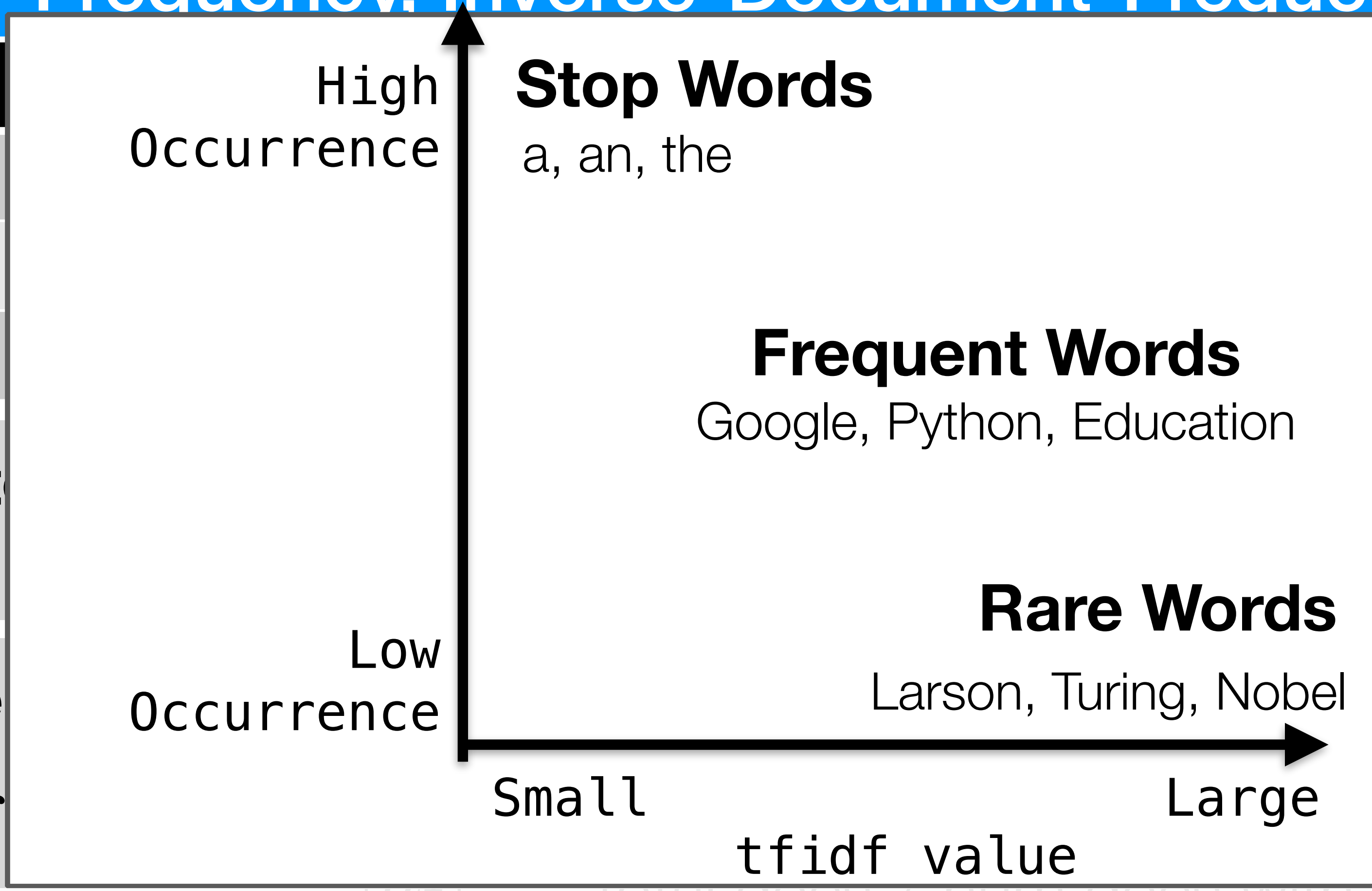
TID	Slee	tc	Stev
1	0	.86	0
2	0.1	.02	0
3	0.1	.1	0

t $\in D$

n d”

inverse

idf $\in d$



This is often used in RAG systems, for Keyword retrieval!!

Want to know more?

Take Natural Language Processing!

For Next Lecture

- Before next class:
  - verify installation of seaborn, plotly, (and/or bokeh if you want)
  - look at pandas table data and additional tutorials

- Next time: Data Visualization

Lecture Notes for

Machine Learning in Python

Professor Eric Larson

Data Quality and Imputation