Lecture Notes for
Machine Learning in Python

Professor Eric Larson

Week One, Lecture Two

Class Logistics and Agenda

- Canvas Access?
- In-Class Assignments for Distance?
- Participation for Distance?
- Anaconda Installs?
- Agenda:
    - Numpy
    - Data Quality
    - Attributes Representation
        - documents
    - The Pandas eco-system
        - loading and manipulating attributes

synchronous

Demo

Jupyter Notebooks and Numpy

Data Quality

Review of Feature Data

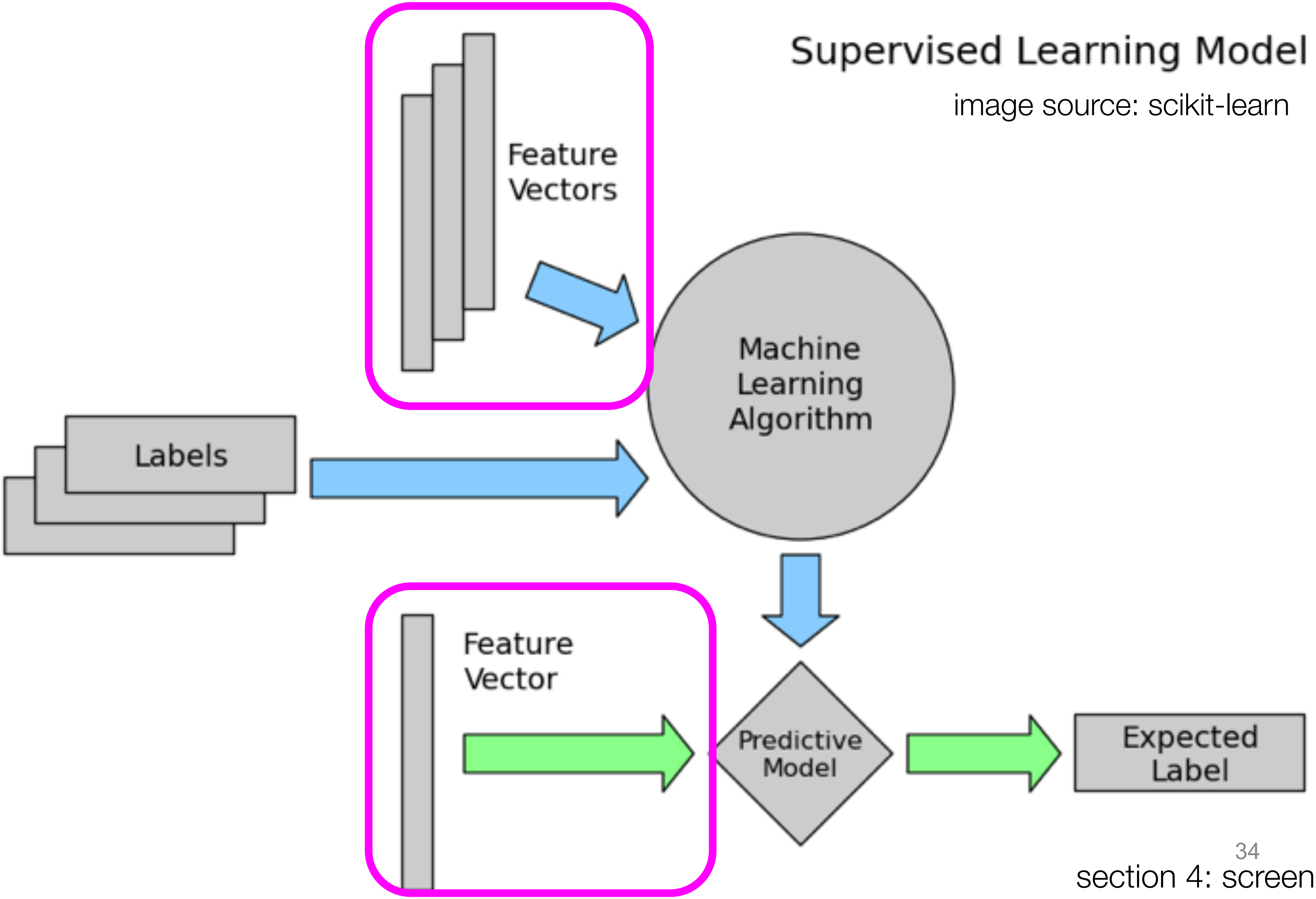

image source: scikit-learn

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)
  - Attributes 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 How?
  - **Ignore** the Missing Value During Analysis
  - Replace with all possible values (talk about later)

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

self test ML1b.1

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

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?
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

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

inverse document frequency: normalize occurrences

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

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

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

TF-IDF

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

self test ML1b.2

Demo

Sklearn and Pandas

TF-IDF

DataFrames

Loading

Indexing

Imputing

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