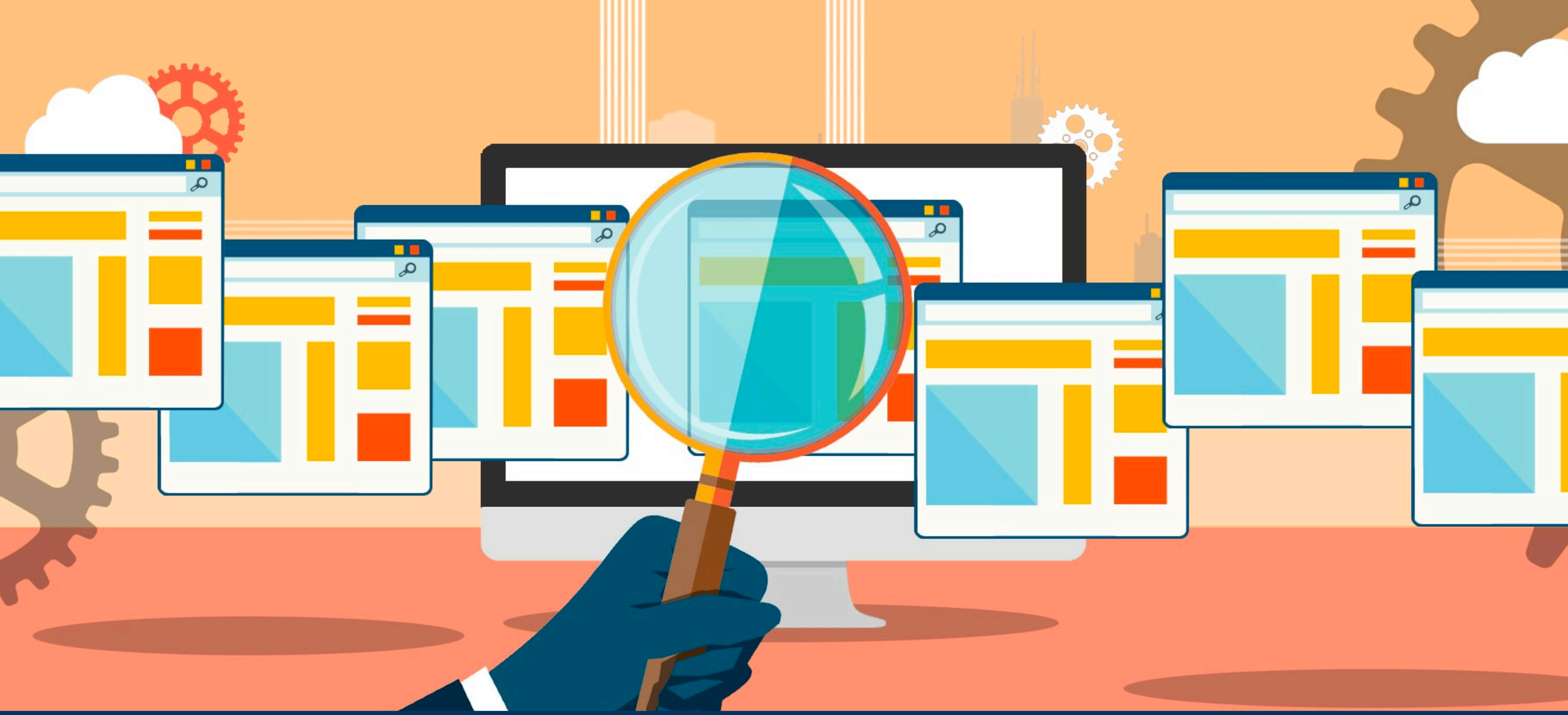


Supervised Learning

INST414 - Data Science Techniques

Six Core Learning Objectives

1. Collect and clean large-scale datasets
2. Articulate the math behind supervised and unsupervised techniques
3. Execute supervised and unsupervised machine learning techniques
4. Select and evaluate various types of machine learning techniques
5. Explain the results coming out of the models
6. Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach

This Module's Learning Objectives

Part 1

- Differentiate between classification and regression in supervised learning
- Describe how voting is used in for k-nearest neighbors classification
- Differentiate binary, multi-class, and multi-label classification
- Define overfitting and describe its impact on generalizability

This Module's Learning Objectives

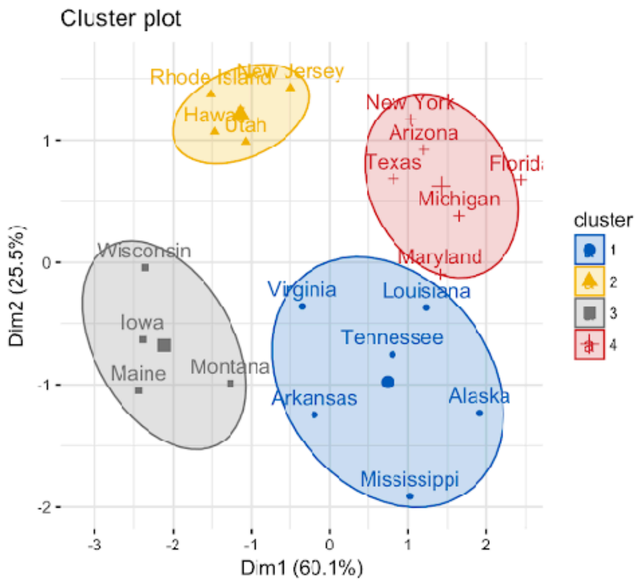
Part 1

- Differentiate between classification and regression in supervised learning
- Describe how voting is used in for k-nearest neighbors classification
- Differentiate binary, multi-class, and multi-label classification
- Define overfitting and describe its impact on generalizability

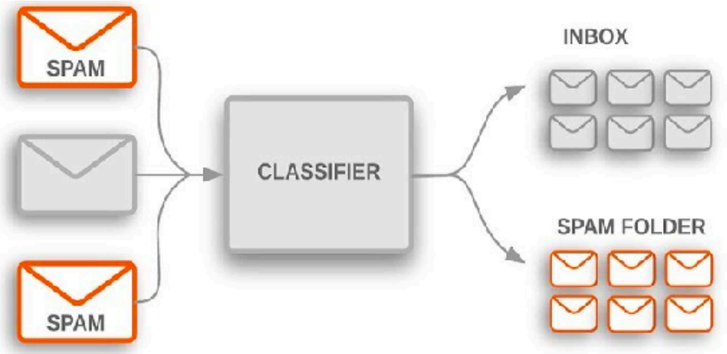
Machine learning is…

An overloaded term

The study of algorithms that “learn” from data



“Learn” some structure in the data



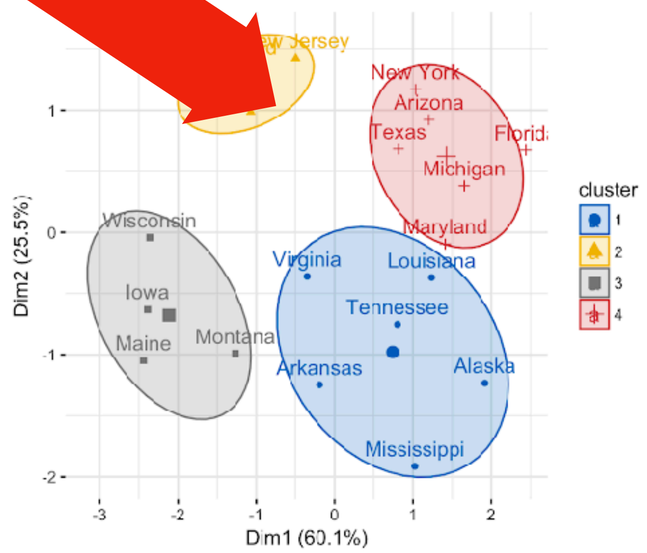
“Learn” to generalize from examples of a task

Two main types of machine learning problems…

Unsupervised learning

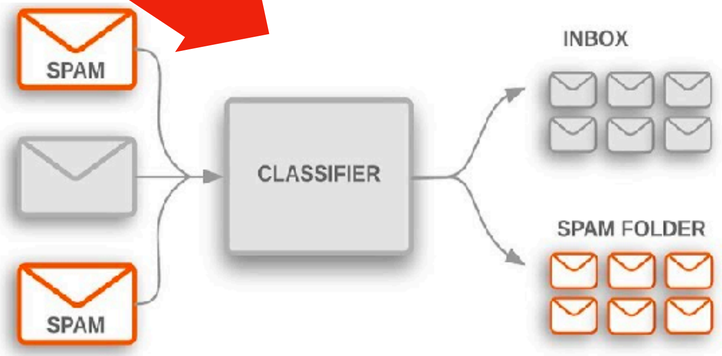
Supervised learning

Unsupervised



"Learn" some structure in the data

Supervised



"Learn" to generalize from examples of a task

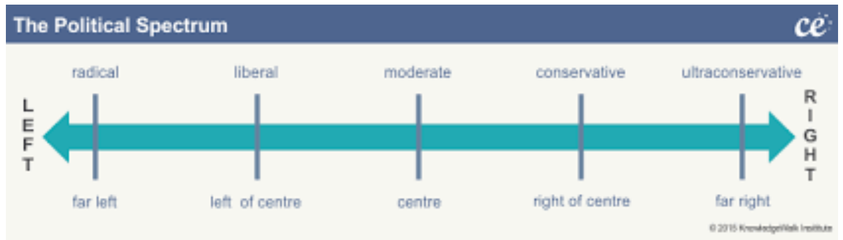
Unsupervised:
You want to find structure in your data, but you don't have examples of this structure

How do I know if the problem is "unsupervised" or "supervised"?

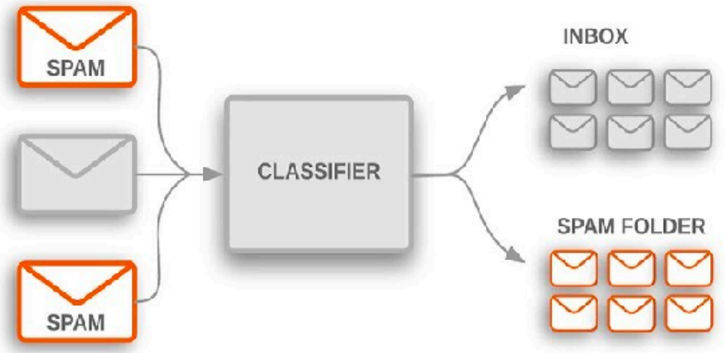
Supervised:
You want to recover (predict) <u>known relationships</u> in your data, and you have examples (i.e., labels) of these relations

Do you have "Labels"?

Supervised Learning

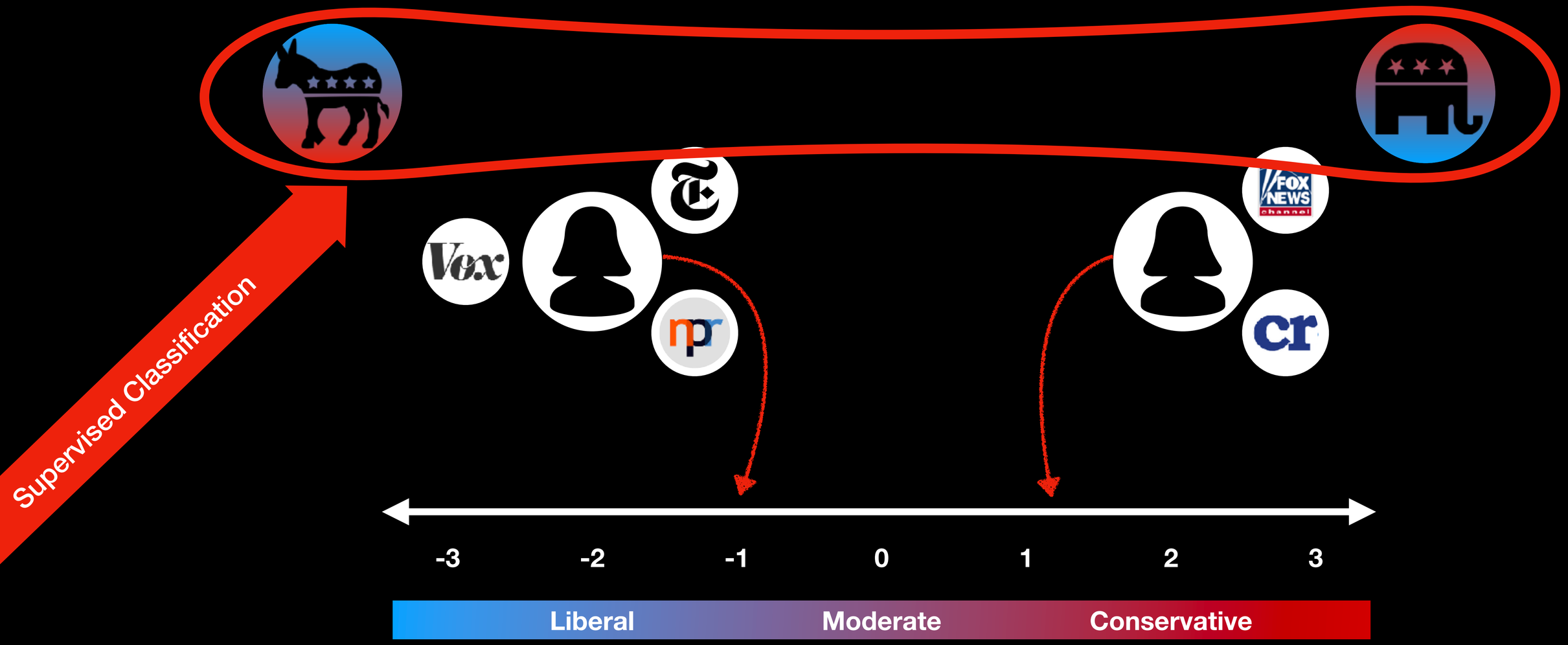


Predicting scores based on known, continuous labels (regression)

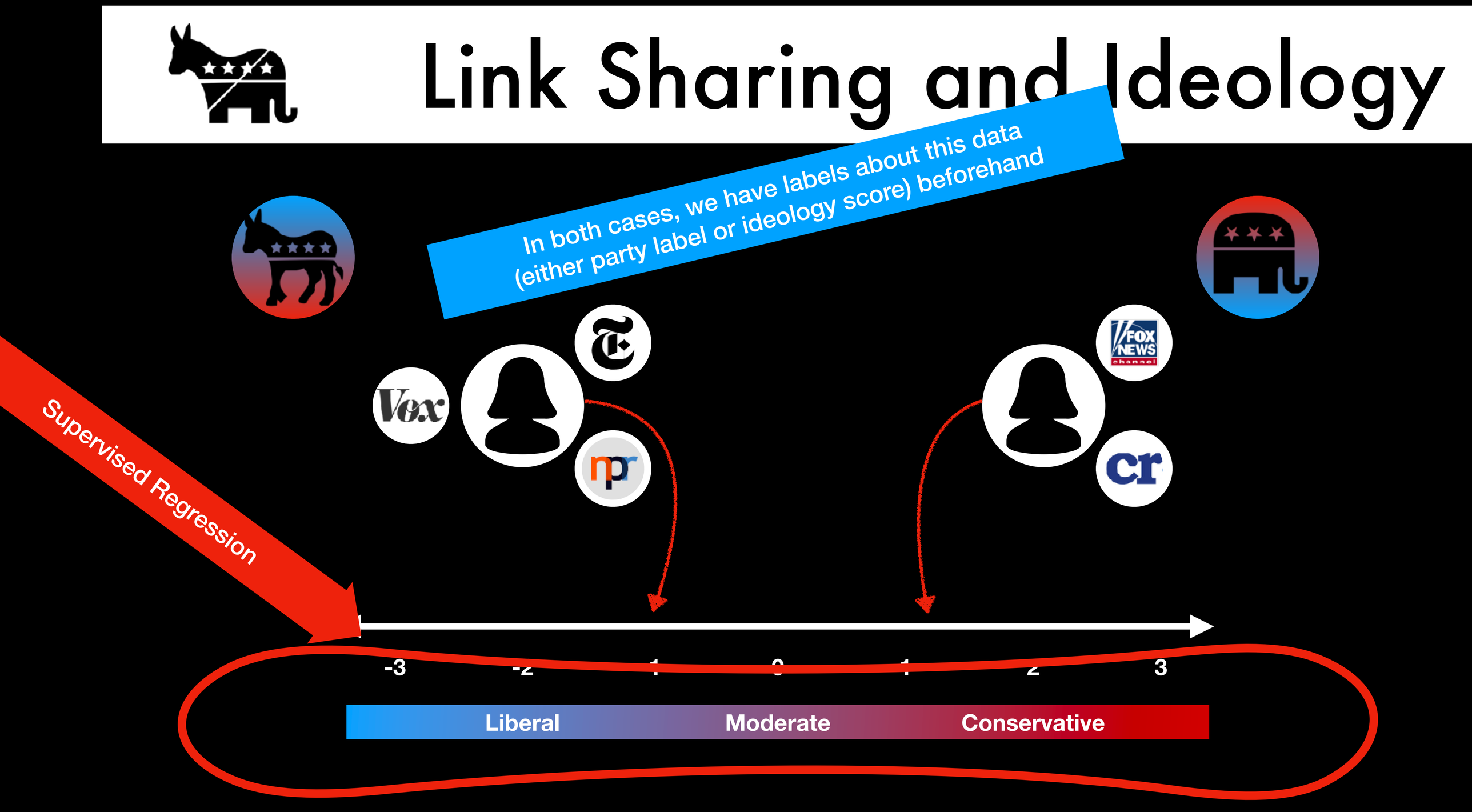


Classifying elements based on known, categorical labels (classification)

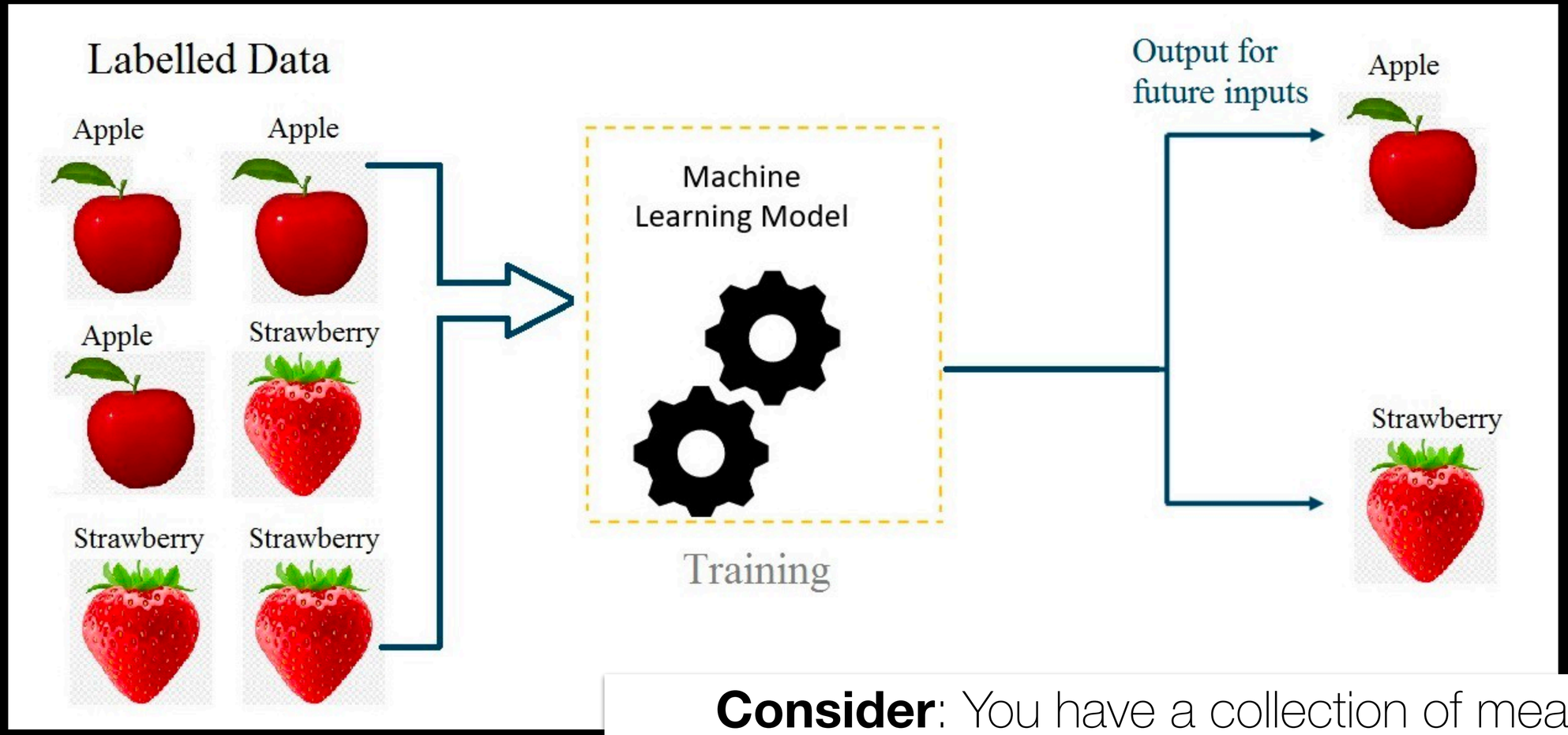
Link Sharing and Ideology



Link Sharing and Ideology



Some Terminology



Consider: You have a collection of measures and want to classify them into different classes of fruit

Some Terminology

Consider: You have a collection of measures and want to classify them into different classes of fruit

mass	width	height	color_score	fruit_label
192	8.4	7.3	0.55	1
180	8.0	6.8	0.59	1
176	7.4	7.2	0.60	1
86	6.2	4.7	0.80	2
84	6.0	4.6	0.79	2
80	5.8	4.3	0.77	2
80	5.9	4.3	0.81	2
76	5.8	4.0	0.81	2
178	7.1	7.8	0.92	1
172	7.4	7.0	0.89	1
166	6.9	7.3	0.93	1
172	7.1	7.6	0.92	1
154	7.0	7.1	0.88	1
164	7.3	7.7	0.70	1
152	7.6	7.3	0.69	1
156	7.7	7.1	0.69	1
156	7.6	7.5	0.67	1
168	7.5	7.6	0.73	1
162	7.5	7.1	0.83	1
162	7.4	7.2	0.85	1

Some Terminology

Data Instances/Samples/Examples: Individual elements of data (rows in a data matrix)

Features/Independent Variables: Individual properties of a sample (columns in a data matrix)

Target Value/Outcome: The attribute we want to predict/infer for each sample

X and y: *X* is the set of samples in your data, and y is a vector of the target values for rows in *X*

Model/Estimator: The function *f(X) —> y* mapping attributes to outcomes

mass	width	height	color_score	fruit_label
192	8.4	7.3	0.55	1
180	8.0	6.8	0.59	1
176	7.4	7.2	0.60	1
86	6.2	4.7	0.80	2
84	6.0	4.6	0.79	2
80	5.8	4.3	0.77	2
80	5.9	4.3	0.81	2
76	5.8	4.0	0.81	2
178	7.1	7.8	0.92	1
172	7.4	7.0	0.89	1
166	6.9	7.3	0.93	1
172	7.1	7.6	0.92	1
154	7.0	7.1	0.88	1
164	7.3	7.7	0.70	1
152	7.6	7.3	0.69	1
156	7.7	7.1	0.69	1
156	7.6	7.5	0.67	1
168	7.5	7.6	0.73	1
162	7.5	7.1	0.83	1
162	7.4	7.2	0.85	1

No concept of order between labels

$$y \in \mathbb{R}$$

Predicting scores based on known, continuous labels (regression)

$$y \in \text{Set}(L_1, L_2, ..., L_i)$$

Classifying elements based on known, categorical labels (classification)

How do you **evaluate** the **performance/quality** of your model?

Depends on classification or regression

Many, many evaluation metrics exist

Model Evaluation – Classification

Most simplistic evaluation: Accuracy

How often did you predict the right label?

Accuracy = $\frac{1}{|X|} \sum_i f(x_i) == y_i$

|X| = # of samples

Was the prediction the same as the actual label?

Model Evaluation – Regression

Common metric: average square of delta between true and predicted value

"Error" = predicted value - true value

Mean Squared Error =

$$\frac{1}{|X|}\sum_{i}(f(x_i)-y_i)^2$$

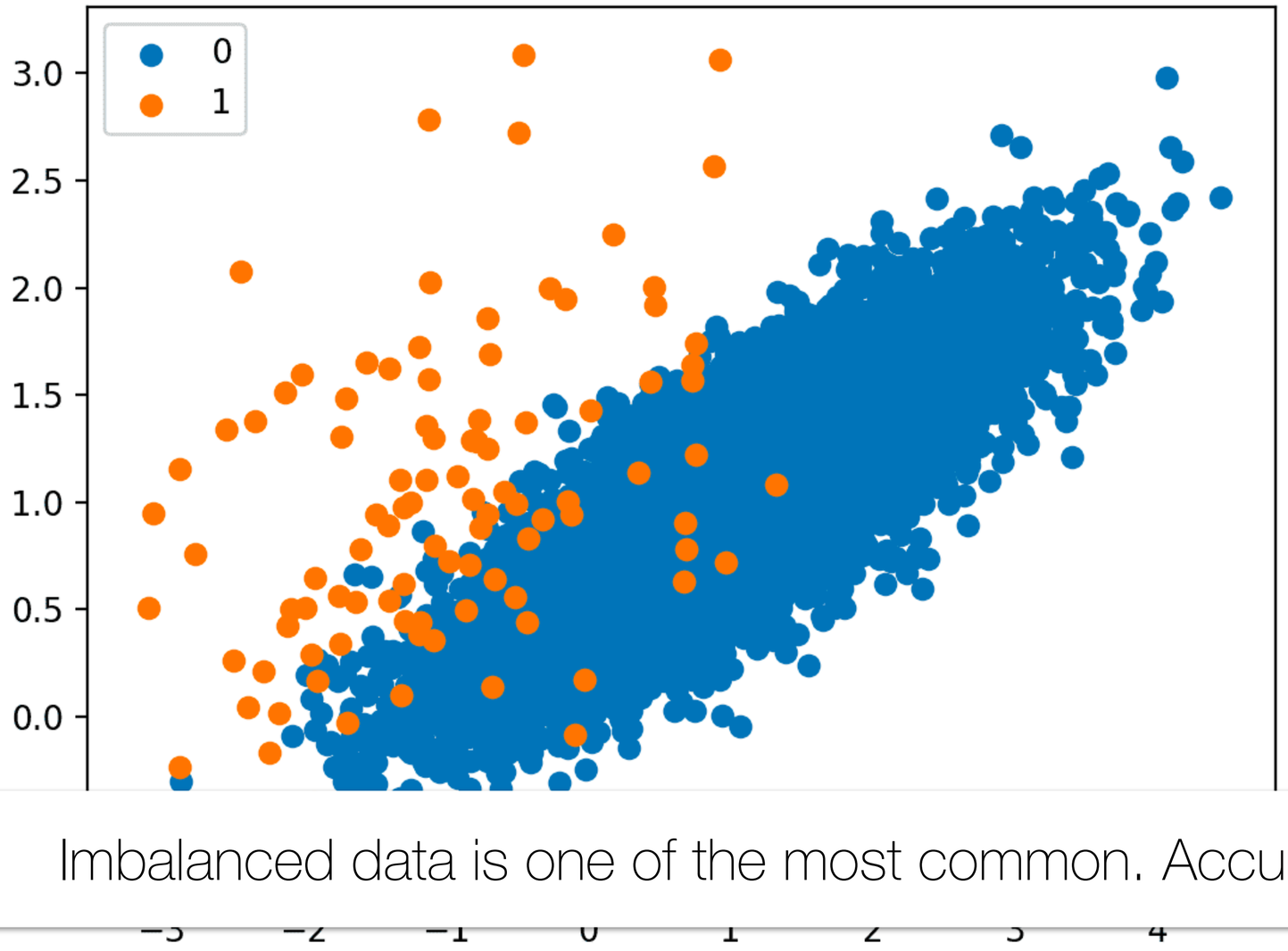
|X| = # of samples

Delta between true and predicted value

do you **evaluate** the **performance/quality** of your model?

Depends on classification or regression

Many, many evaluation metrics exist

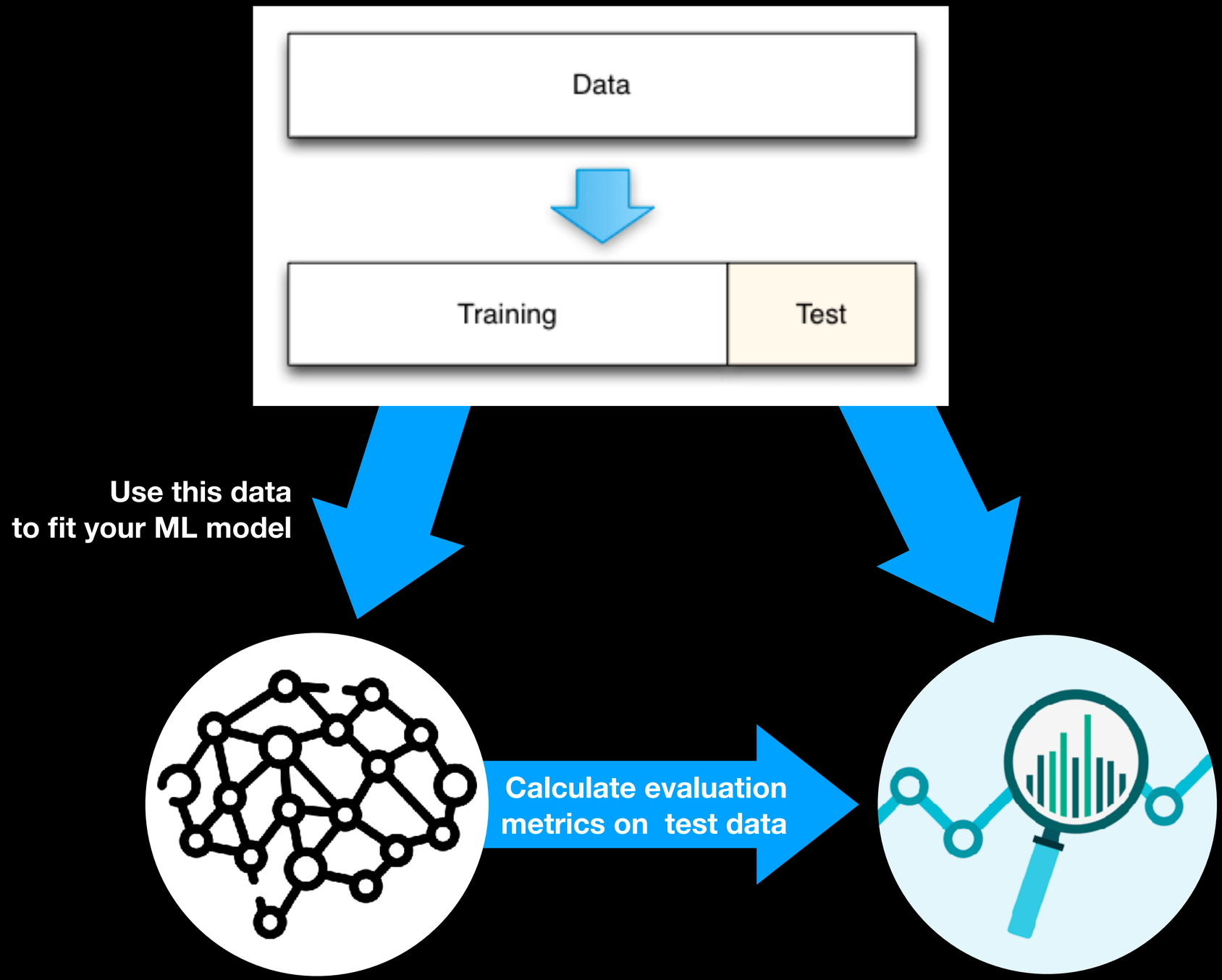


Imbalanced data is one of the most common. Accuracy means less here

Should **not** evaluate your model based on the data used to train it

"Train" = fitting the model/determining parameters

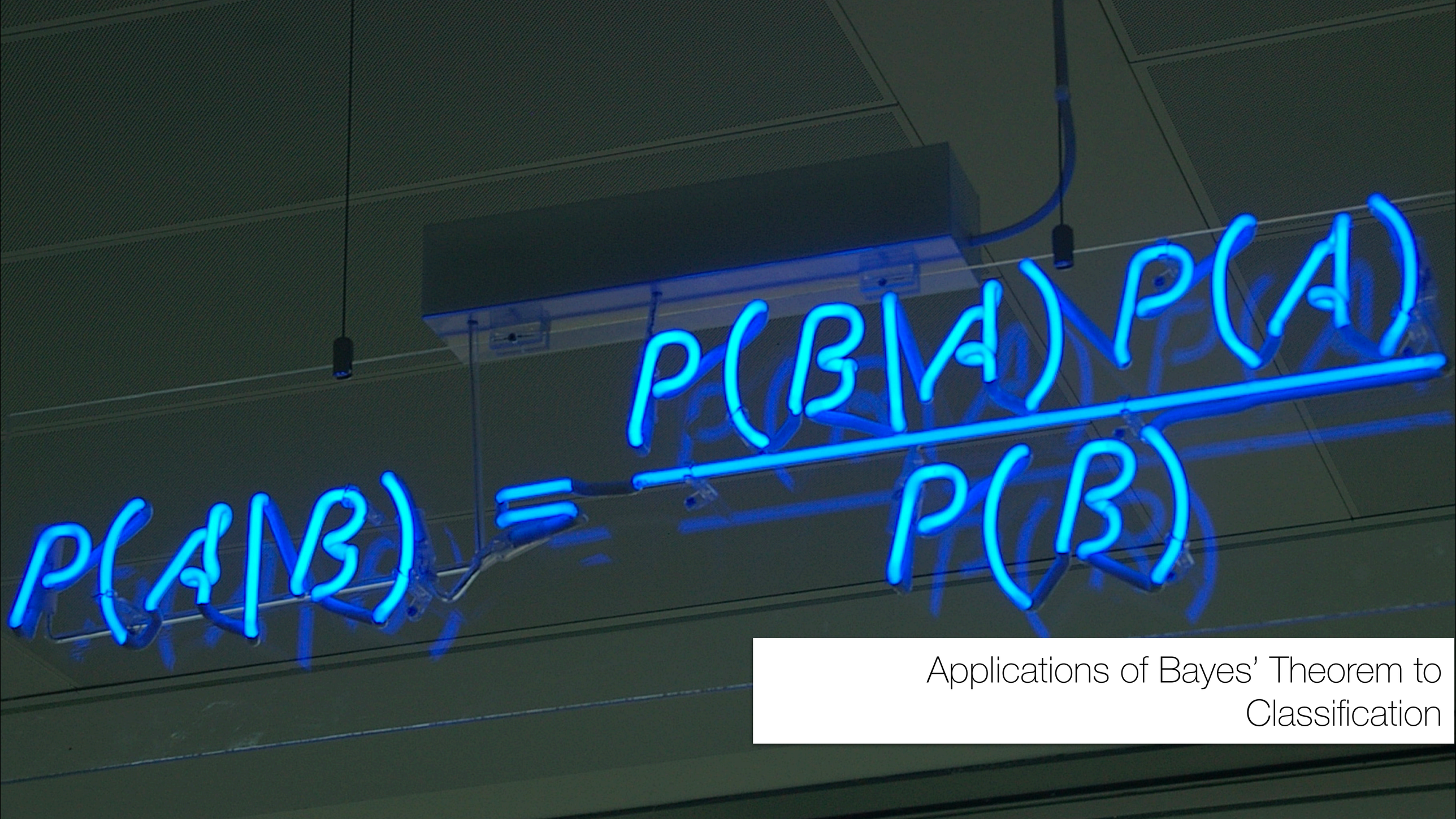
Should evaluate on held-out "test" data


Data
Training
Test
Use this data
to fit your ML model
Calculate evaluation
metrics on test data

Split your data into training and testing sets

More training data => less testing data

Often 90/10 or 80/20 split



Applications of Bayes' Theorem to Classification

Into which cluster will a Sports movie go?

$$Pr(Cluster = X | Genre = Sport)$$

$$= \frac{Pr(Genre = Sport | Cluster = X) Pr(Cluster = X)}{Pr(Genre = Sport)}$$

Probability of a specific cluster given Sport genre...

$$\text{argmax}_{X \in Clusters} Pr(Cluster = X | Genre = Sport)$$

Really want this:

$$\text{argmax}_{X \in Clusters} Pr(Cluster = X | Genre = Sport)$$

Really want this:

Maximize Over:

$$= \frac{Pr(Genre = Sport | Cluster = 1) Pr(Cluster = 1)}{Pr(Genre = Sport)}$$

$$= \frac{Pr(Genre = Sport | Cluster = 2) Pr(Cluster = 2)}{Pr(Genre = Sport)}$$

...

$$= \frac{Pr(Genre = Sport | Cluster = C_n) Pr(Cluster = C_n)}{Pr(Genre = Sport)}$$

Each value has the same denominator

$$\text{argmax}_{X \in Clusters} Pr(Cluster = X | Genre = Sport)$$

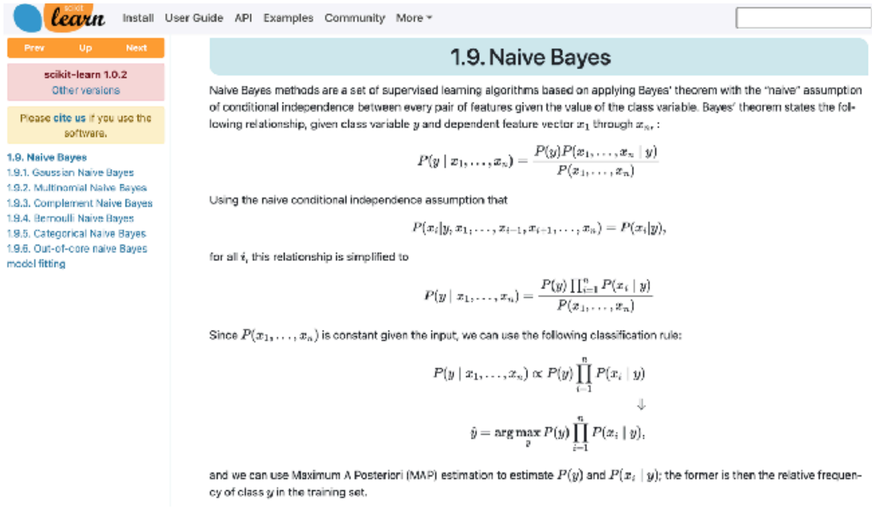
$$= \frac{Pr(Genre = Sport | Cluster = 1) Pr(Cluster = 1)}{Pr(Genre = Sport)}$$

$$= \frac{Pr(Genre = Sport | Cluster = 2) Pr(Cluster = 2)}{Pr(Genre = Sport)}$$

...

$$= \frac{Pr(Genre = Sport | Cluster = C_n) Pr(Cluster = C_n)}{Pr(Genre = Sport)}$$

Standard classification algorithm



scikit-learn Install User Guide API Examples Community More

Prev Up Next

scikit-learn 1.0.2
Other versions

Please cite us if you use the software.

1.9. Naive Bayes
1.9.1. Gaussian Naive Bayes
1.9.2. Multinomial Naive Bayes
1.9.3. Complement Naive Bayes
1.9.4. Bernoulli Naive Bayes
1.9.5. Categorical Naive Bayes
1.9.6. Out-of-core naive Bayes model fitting

1.9. Naive Bayes

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive conditional independence assumption that

$$P(x_i | y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i | y),$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

-SkLearn

Could also predict gross box office returns (regression problem)

Box Office Mojo by IMDbPro

Search for Titles

IMDbPro

Domestic | International | Worldwide | Calendar | All Time | Showdowns | Indices

Daily | Weekend | Weekly | Monthly | Quarterly | Yearly | Seasons | Holidays

mestic 2022 Weekend 14

022

« Week « Year

Week »

Key: New This Week

Rank	LW	Release	Gross	%± LW	Theaters	Change	Average	Total Gross	Weeks	Distributor
1	**-**	**Sonic the Hedgehog 2**	**$72,105,176**	**-**	**4,234**	**-**	**$17,030**	**$72,105,176**	**1**	**Paramount Pictures**
2	1	Morbius	$10,201,332	-73.8%	4,268	-	$2,390	$57,078,553	2	Columbia Pictures
3	2	The Lost City	$9,027,813	-38.6%	3,797	-486	$2,377	$68,716,972	3	Paramount Pictures
4	**-**	**Ambulance**	**$8,699,630**	**-**	**3,412**	**-**	**$2,549**	**$8,699,630**	**1**	**Universal Pictures**
5	3	The Batman	$6,458,465	-41.3%	3,254	-478	$1,984	$358,960,613	6	Warner Bros.
6	9	Everything Everywhere All at Once	*$6,059,263*	*+461.8%*	*1,250*	*+1,212*	*$4,847*	*$8,428,962*	3	A24
7	4	Uncharted	$2,634,821	-28.1%	2,318	-746	$1,136	$142,937,039	8	Sony Pictures Entertainment (SPE)

This Module's Learning Objectives

Part 1

- Differentiate between classification and regression in supervised learning
- Describe how voting is used in for k-nearest neighbors classification
- Differentiate binary, multi-class, and multi-label classification
- Define overfitting and describe its impact on generalizability

What are your questions?

Prof. Cody Buntain | @codybuntain | cbuntain@umd.edu
Director, Information Ecosystems Lab