Intro to Classification Algorithms

GHW February 2023

First, let’s check in!

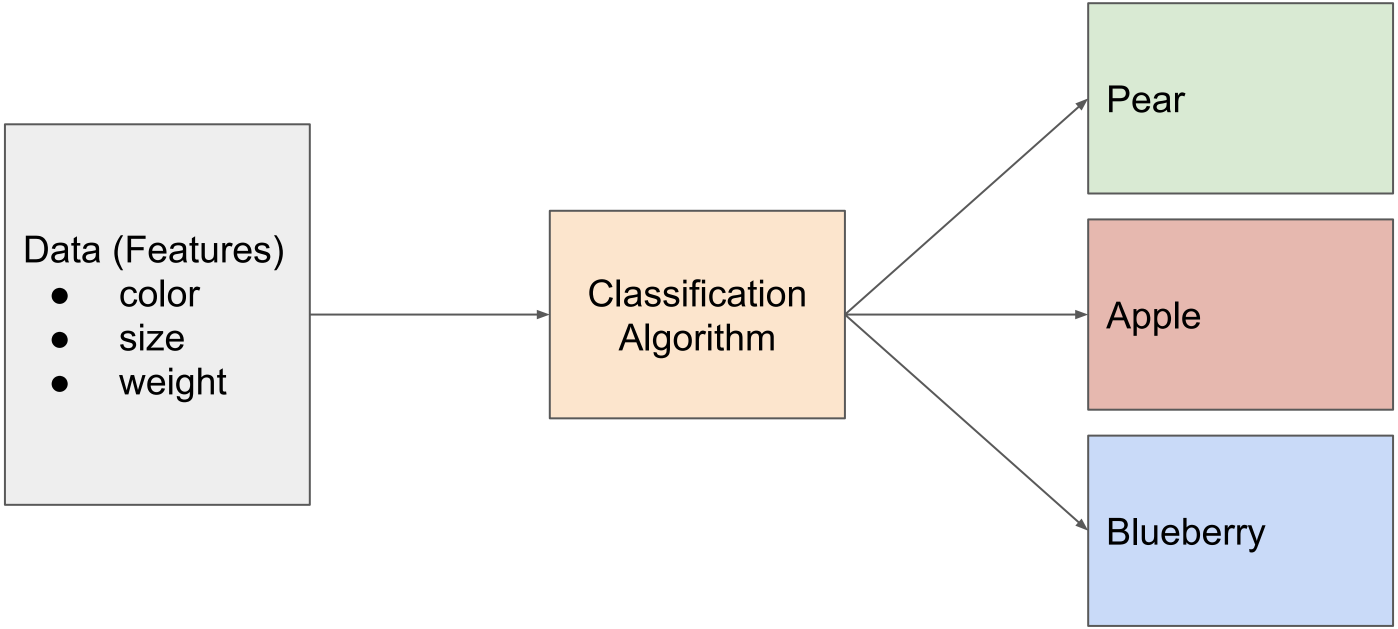
Most Common Classification Algorithms

1. Logistic Regression
2. Naive Bayes
3. K-Nearest Neighbors (KNN)
4. Markov Decision Process (MDP)
5. Decision Trees / Random Forest
6. K-Means Clustering
7. Support Vector Machines

What is a Classification Algorithm?

A classification algorithms takes raw data and predicts its category.

ex: **fruit classification algorithm**…



Data (Features)

- color
- size
- weight

Classification Algorithm

Pear

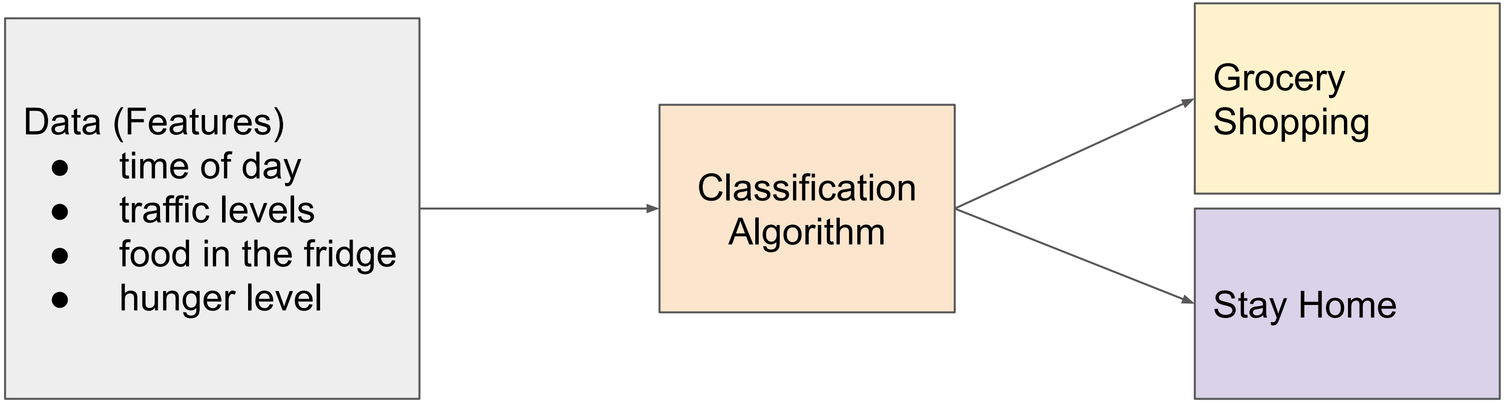
Apple

Blueberry

What is a Classification Algorithm?

A classification algorithms takes raw data and predicts its category.

ex: **email spam classification algorithm**…

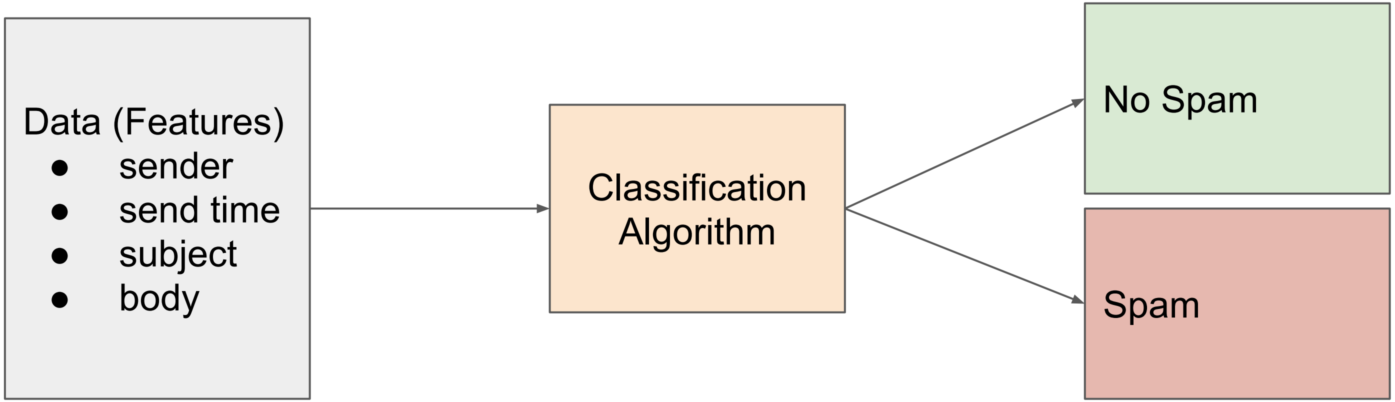


Data (Features)
- time of day
- traffic levels
- food in the fridge
- hunger level

What is a Classification Algorithm?

A classification algorithms takes raw data and predicts its category.

ex: **email spam classification algorithm**…



Data (Features)
- sender
- send time
- subject
- body

Classification Algorithm

No Spam

Spam

How do we construct an algorithm specific to our needs?

Type of Learning

Supervised

- Model “learns” from training examples, provided by humans.
- Mostly used for “discrimination” (recognition) tasks.
- Train / test data
  - Typically 80% / 20% split
  - **ex: Random Forest (Lecture 3)**

apple

apple

apple

Type of Learning

Unsupervised

- Model “learns” from unlabeled data.
- Mostly used for “generative” (imagination) tasks.
- No split between train / test data.
  - **At the end, we will look at the K-means clustering algorithm in Python for an example! (Note: This is technically a “clustering” algorithm, but it does classify data into categories.)**

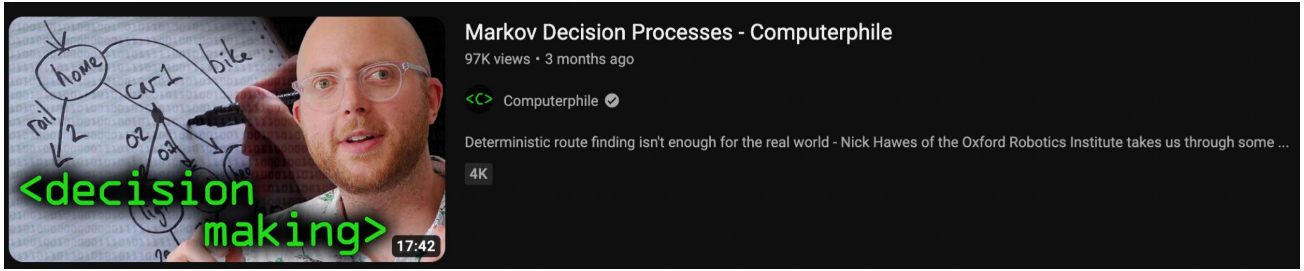
- Age
- Eyes
- Perspective
- Mood

Generative Adversarial Network (GAN)

Type of Learning

Reinforcement Learning

- Model “learns” the best strategy, using a scenario given by humans.
- Scenario: actions + environment
- Mostly used for decision making tasks (ex: robotics).
- Example: Markov Decision Process (MDP) (Lecture 2)



Markov Decision Processes - Computerphile

97K views • 3 months ago

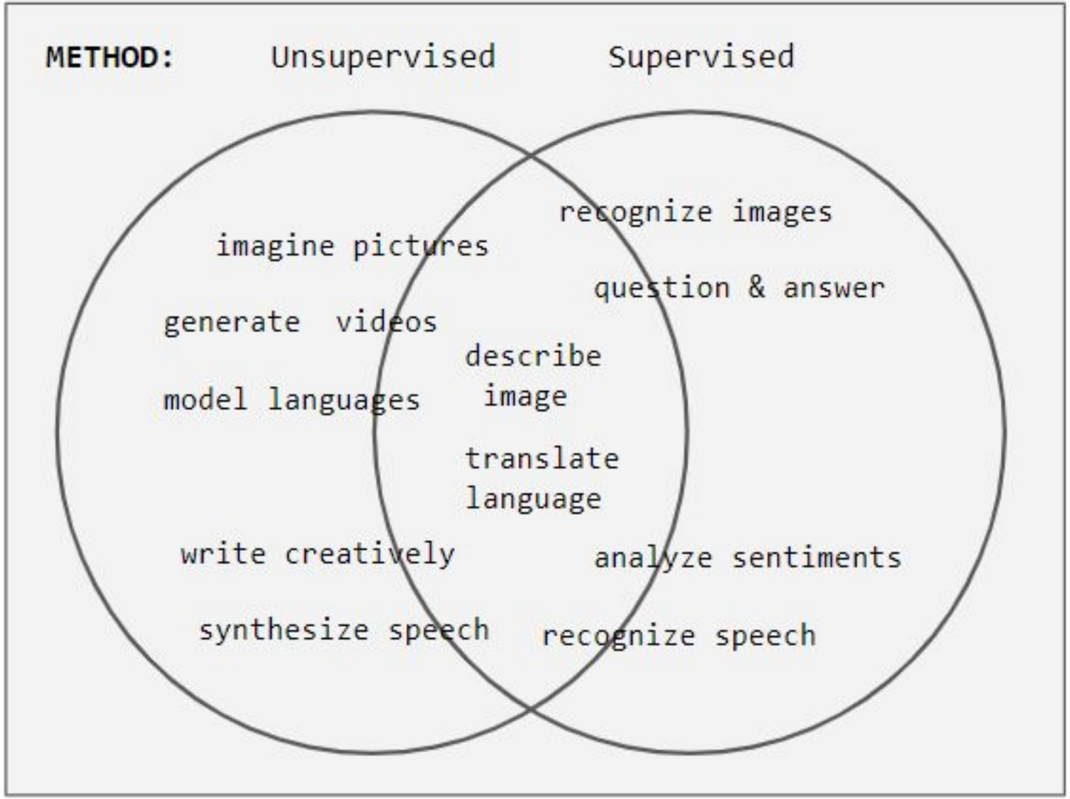
<C> Computerphile

Deterministic route finding isn't enough for the real world - Nick Hawes of the Oxford Robotics Institute takes us through some ...

4K

https://youtu.be/2iF9PRriA7w

Type of Learning: Select Learning Mode based on Application!



METHOD: Unsupervised Supervised

imagine pictures

generate videos

model languages

write creatively

synthesize speech

describe image

translate language

recognize images

question & answer

analyze sentiments

recognize speech

Other Factors to Consider

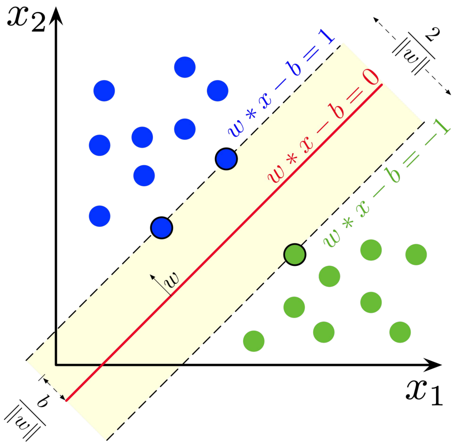
Can you reduce the number of features to simplify your model? (Feature Selection)

Number of Features

- How many variables do you have? (age, height, ….)

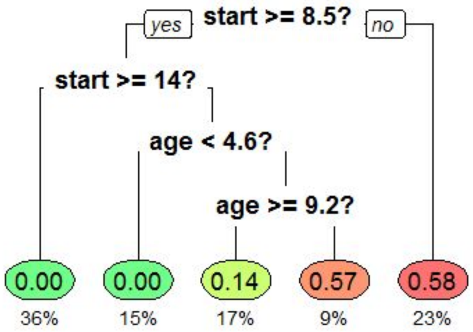
More features:

consider a **support vector machine**



Less features:

consider a **decision tree**



Other Factors to Consider

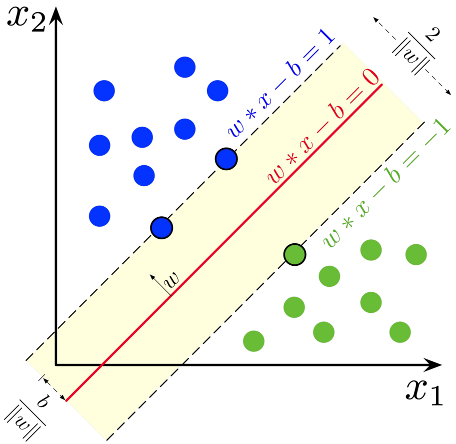
Can you transform your data to make it linear? (We will discuss this later!)

Linearity

- Is your data linear?

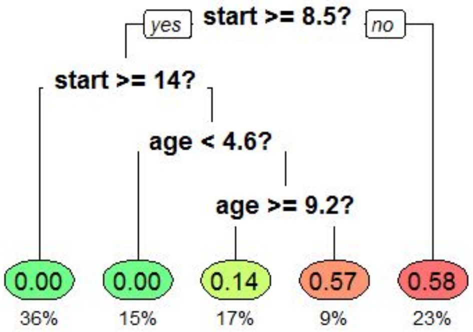
Linear:

consider a **support vector machine**



Nonlinear:

consider a **decision tree**



Other Factors to Consider

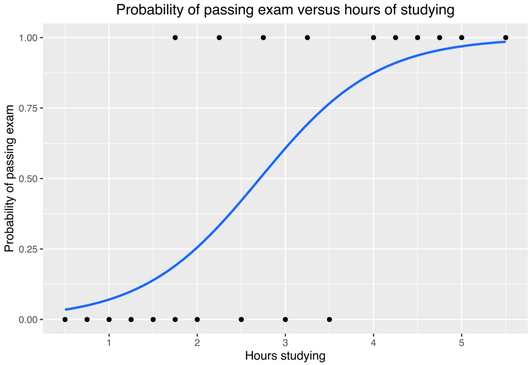
Generally speaking, more training = more accuracy (but not always!)

Training Time

- How much data do you have, and how fast is your computer?

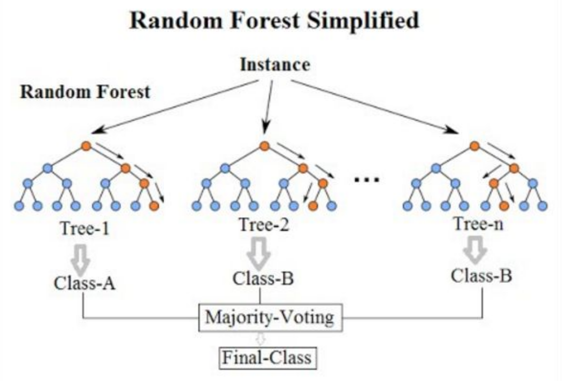
Faster Training:

consider a **logistic regression**



Slower Training:

consider a **random forest**



Other Factors to Consider

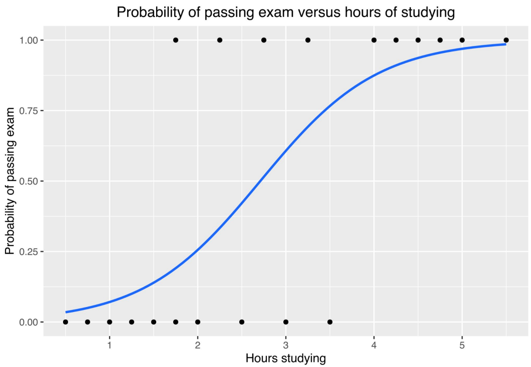
Split into train / **validation** / test to examine the parameter space.

# of Parameters (parameters = model specs. # iterations, error, etc.)

- How much flexibility do you want in your training?

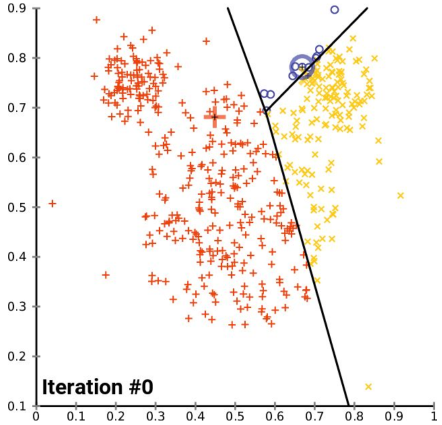
Less Parameters:

logistic regression (4)



More Parameters:

K-means clustering (8)

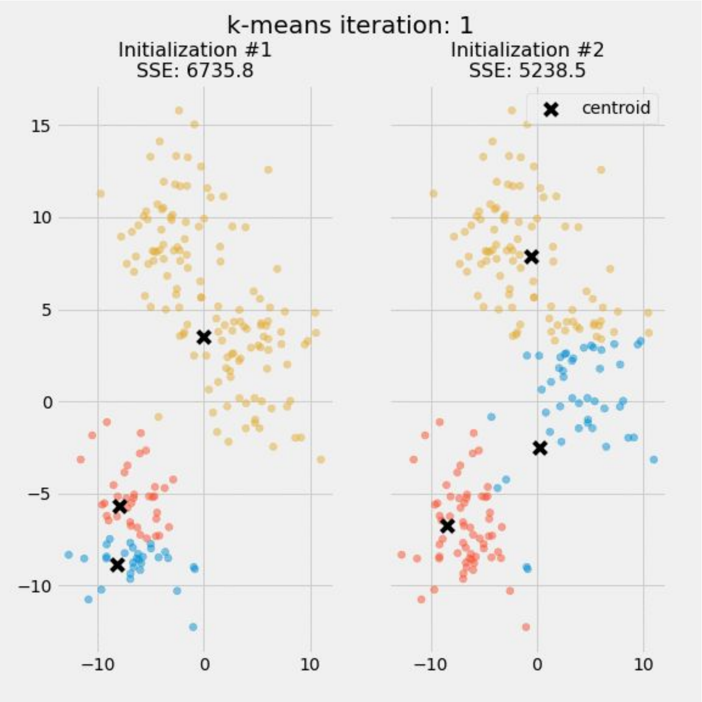


10-minute break

then: K-means Clustering!

What is K-Means Clustering

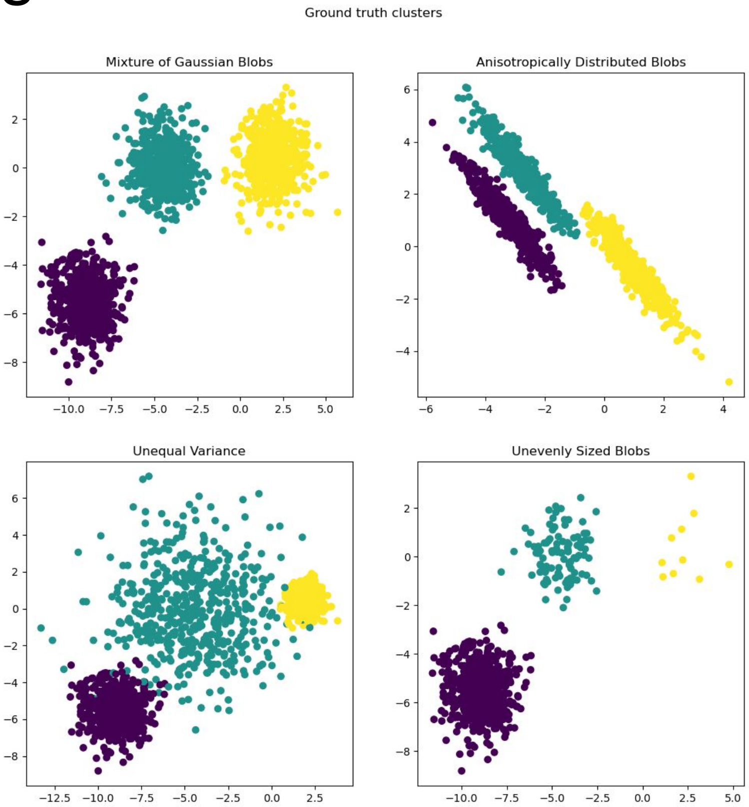
- Clustering algorithm that categories data into “k” groups.
- Random initialization means that the answer will be slightly different each time (**nondeterministic**).
- Each iteration works to minimize the squared errors from a centroid.



https://realpython.com/k-means-clustering-python/

K-Means Clustering Assumptions

- There a K clusters, each of roughly the same size
- Features are isotropically distributed.
- Each feature has a spherical variance.
- However, there is flexibility in many of these assumptions (see right panel from sklearn documentation).



https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html

Let’s do an example!