A Quick Look at Machine Learning

Chapter 24

How do we learn?

- Memorization – Declarative Knowledge
  - Lists of facts
  - Limited by:
    - Time available
    - Ability to retain information
- Generalization – Interpretive Knowledge
  - Deduce new facts from old
  - Reasoning
  - Problem Solving
  - Limited by:
    - Accuracy of deduction process
    - Future following the past

How do computers solve problems?

- Humans program them to perform a sequence of logical statements and calculations
- Machine Learning (ML)

What is Machine Learning?

- “Field of study that gives computers the ability to learn without being programmed” (Arthur Samuel)

- Automatically learn to make useful inferences from implicit patterns in data.
  - Observe a set of **training data**
  - Use inference techniques to create a model
    - Types of model
    - Create the model
    - Parameters
  - Apply model to previously unseen data

Types of ML

- Supervised
  - Given a set of feature/label pairs, find a rule that predicts the label associated with previously unseen data
- Unsupervised
  - Give a set of feature vectors (no labels) group them into “natural” clusters

Feature vectors

2021-22 PLAYER STATISTICS

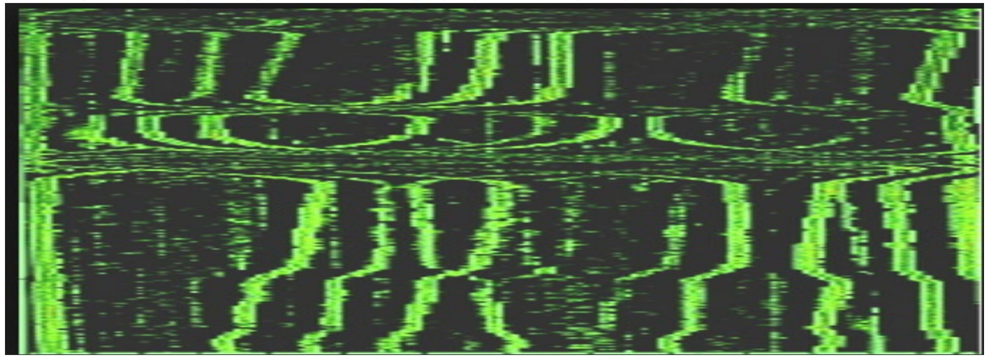
Player		G	PTS	FG	FG%	3P%	FT%	OREB	DREB	REB	AST	STL	TO	PF
	Justin Anderson 10 • F-G	9	5.1	1.7	36.6%	25%	90%	0.3	2	2.3	1.9	0.4	0.6	1.4
	Goga Bitadze 88 • C-F	45	6.3	2.5	52.9%	28.4%	62.9%	1.4	2	3.4	1	0.4	0.8	2
	Oshae Brissett 12 • F-G	60	8.2	2.8	40.4%	34%	69.4%	1.6	3.5	5.2	0.9	0.6	0.8	1.6
	Malcolm Brogdon 7 • G	36	19.1	6.8	44.8%	31.2%	85.6%	0.9	4.2	5.1	5.9	0.8	2.1	2
	Chris Duarte 3 • G	55	13.1	4.9	43.2%	36.9%	80.4%	0.7	3.4	4.1	2.1	1	1.6	1.7
	Tyrese Haliburton 0 • G	68	15	5.5	46.4%	41.2%	83.3%	0.8	3.3	4	7.9	1.8	2.6	1.5
	Buddy Hield 24 • G	72	15.4	5.4	40%	36.3%	87.5%	0.8	3.5	4.3	2.7	0.9	1.8	2.2
	Isaiah Jackson 23 • F	31	7.5	2.9	55.2%	30.8%	65.3%	1.6	2.5	4.1	0.3	0.5	1	2.5

https://www.nba.com/pacers/stats

Feature representation

- All features are not complete or perfect

- Feature engineering
  - Represent examples by feature vectors that will facilitate generalization
    - Weight, height, strength, speed, agility
  - Dimensionality
    - Number of features v. number of samples
  - Try to maximize the amount of useful information versus irrelevant information
    - Raise the Signal to Noise Ratio



ML common elements

- A representation of the **model**
- An **objective function** for assessing the goodness of the model
  - Briefly discussed in chapter 18
- An **optimization method** for learning a model that minimizes or maximizes the objective function
  - Chapters 12 & 13

Feature vectors

Name	Aspects		
Abraham Lincoln	American	President	193 cm tall
Benjamin Harrison	American	President	168 cm tall
George Washington	American	President	189 cm tall
James Madison	American	President	163 cm tall
Charles de Gaulle	French	President	196 cm tall
Louis Napoleon	French	President	169 cm tall

A	Abraham Lincoln, George Washington, Charles de Gaulle
B	Benjamin Harrison, James Madison, Louis Napoleon

Supervised learning techniques

- Regression models (linear, logistic)
- Classification (chapter 26)
- Similarity learning
- Support vector machines
- Linear discriminate analysis
- Decision tree
- Neural networks

Unsupervised learning techniques

- Clustering (chapter 25)
- Anomaly detection
- Neural networks
- Latent variable learning techniques

Reptile(?) example

	Features					Label
Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes

Initial model:

- Not enough information to generalize

Reptile(?) example (continued)

	Features					Label
Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes

Initial model:

- Egg laying
- Has scales
- Is poisonous
- Cold blooded
- No legs

Reptile(?) example - refinement

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes

Current model:

- Has scales
- Cold blooded
- No legs

Boa doesn't fit model, but is labeled as reptile.
Need to refine model

Reptile(?) example – non-reptile

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No

Current model:

- Has scales
- Cold blooded
- No legs

Reptile(?) example – later gator

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes

Current model:

- Has scales
- Cold blooded
- Has 0 or 4 legs

Alligator doesn't fit model, but is labeled as reptile.
Need to refine model

Reptile(?) example – getting complicated

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes
Dart frog	True	False	True	False	4	No

Current model:

- Has scales
- Cold blooded
- Has 0 or 4 legs

Reptile(?) example – 2 more

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes
Dart frog	True	False	True	False	4	No
Salmon	True	True	False	True	0	No
Python	True	True	False	True	0	Yes

Current model:
- Has scales
- Cold blooded
- Has 0 or 4 legs

No (easy) way to add to rule that will correctly classify salmon and python (since identical feature values)

Nobody is perfect

	Features					Label
Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes
Dart frog	True	False	True	False	4	No
Salmon	True	True	False	True	0	No
Python	True	True	False	True	0	Yes

Good model:
- Has scales
- Cold blooded

Not perfect, but no false negatives (anything classified as not reptile is correctly labeled); some false positives (may incorrectly label some animals as reptile)

Distance metrics

- We can think of our reptile feature vector as containing four binary values (eggs, scales, poisonous, cold-blooded) and one integer (legs)
  - Rattlesnake = [1,1,1,1,0]
  - Boa constrictor = [1,1,0,1,0]
  - Dart frog = [1,0,1,0,4]

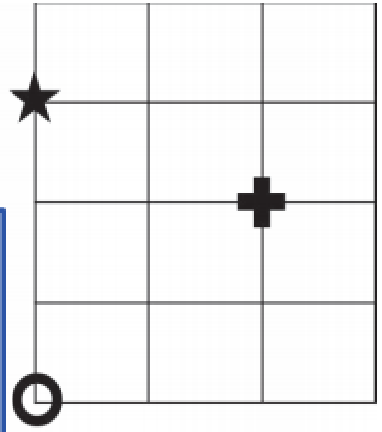
- We can compute the distance between pairs of examples
  - Cluster data into common class (nearest neighbor) *unsupervised*
  - Find classifier surface that optimally groups different labelled collections separating them from others collections *supervised*

Minkowski Metric

$$dist(X1, X2, p) = \left(\sum_{k=1}^{len} abs(X1_k - X2_k)^p\right)^{1/p}$$

p = 1: Manhattan Distance
p = 2: Euclidean Distance

Need to measure distances between feature vectors



Typically use Euclidean metric; Manhattan may be appropriate if different dimensions are not comparable

Is circle closer to star or cross?

- Euclidean distance
  - Cross – 2.8
  - Star – 3
- Manhattan Distance
  - Cross – 4
  - Star - 3

Minkowski distance (in Python)

```
def minkowski_dist(v1, v2, p):
    """Assumes v1 and v2 are equal-length arrays of numbers
       Returns Minkowski distance of order p between v1 and v2"""
    dist = 0.0
    for i in range(len(v1)):
        dist += abs(v1[i] - v2[i])**p
    return dist**(1.0/p)

# scipy.spatial.distance.minkowski(u, v, p=2, w=None)
# u and v are 1-D feature vectors
# p gives Manhattan v. Euclidian distance
# w gives optional weights to the feature vectors
```

Euclidean distance

	rattlesnake	boa constrictor	dart frog
rattlesnake	–	1.414	4.243
boa constrictor	1.414	–	4.472
dart frog	4.243	4.472	–

Add in Ally Gator

	rattlesnake	boa constrictor	dart frog	alligator
rattlesnake	--	1.414	4.243	4.123
boa constrictor	1.414	--	4.472	4.123
dart frog	4.243	4.472	--	1.732
alligator	4.123	4.123	1.732	--

- Why does the alligator seem to have more in common with the dart frog (amphibian) than the snakes (reptiles)?

Make ‘has 0 or 4 legs’ Boolean

	rattlesnake	boa constrictor	dart frog	alligator
rattlesnake	--	1.414	1.732	1.414
boa constrictor	1.414	--	2.236	1.414
dart frog	1.732	2.236	--	1.732
alligator	1.414	1.414	1.732	--