acm.ai

KNNs with Python

Beginner Track Links Sheet:
https://tinyurl.com/f22-ai-beginner
Check out our new Discord server on link tree:
https://linktr.ee/acm_ai_ucla

Announcements!

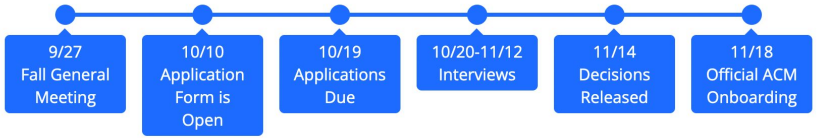
acm.ai

We have a discord now!

Apply to be an ACM intern!

acm.ai

Applications



9/27 Fall General Meeting

Come learn about the different ACM committees!

10/10 Application Form is Open

Fill out the application form for up to 3 committees that you would like to intern for

10/19 Applications Due

Submit your application by 11:59 pm

10/20-11/12 Interviews

A handful of applicants will be chosen for an interview with the committee(s) you applied to

11/14 Decisions Released

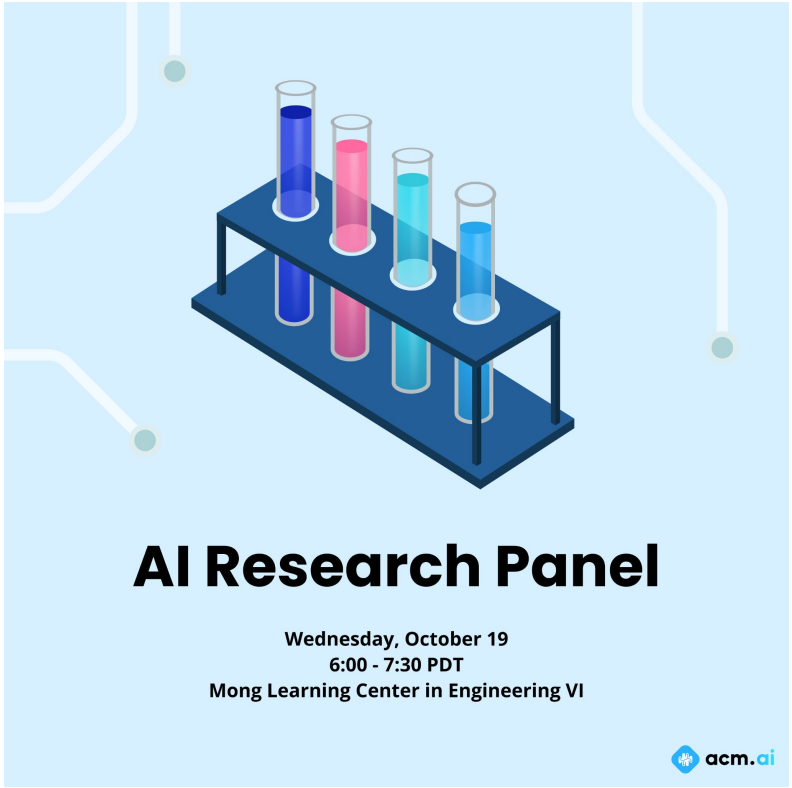
Hear back via Email what committee(s) accepted you (if you are accepted by two, decide which one you would like to intern for)

11/18 Official ACM Onboarding

Learn what you need to know as an ACM Intern

ACM is recruiting interns for Fall 2022!

Apply Now



For more info:
https://www.facebook.com/events/1211994316377417

acm.ai

Today's Content:

- Motivating KNN example
- Kahoot
- KNN Python Demo

You're taking a walk in the sculpture garden

And you see this -> Does it purr or bark?

Some terminology

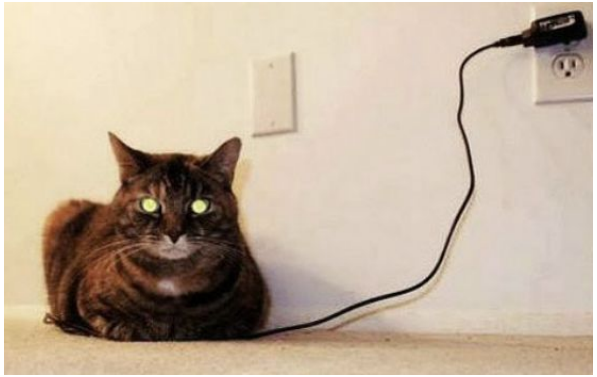
- **Feature** – some property of the object that impacts the target

For an animal, the sharpness of its claws is a feature.

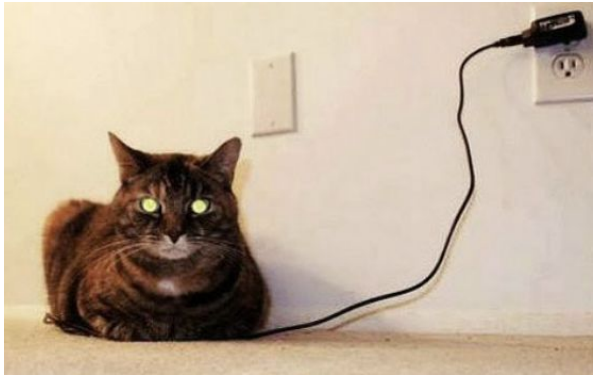
- **Target/Label** – the true value of what we are trying to predict

In our current example, the target would be whether the creature is a cat or a dog.

So: cat or dog?



Feature differences between cats and dogs

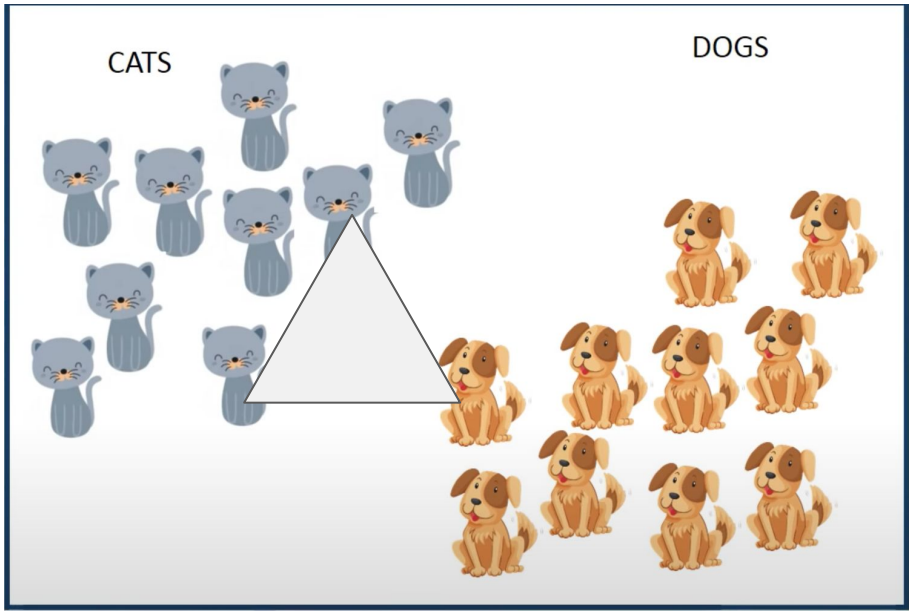


- Sharp claws (used for climbing)
- Shorter ears
- Meows

- Dull claws
- Longer ears
- Barks

Cats v/s Dogs

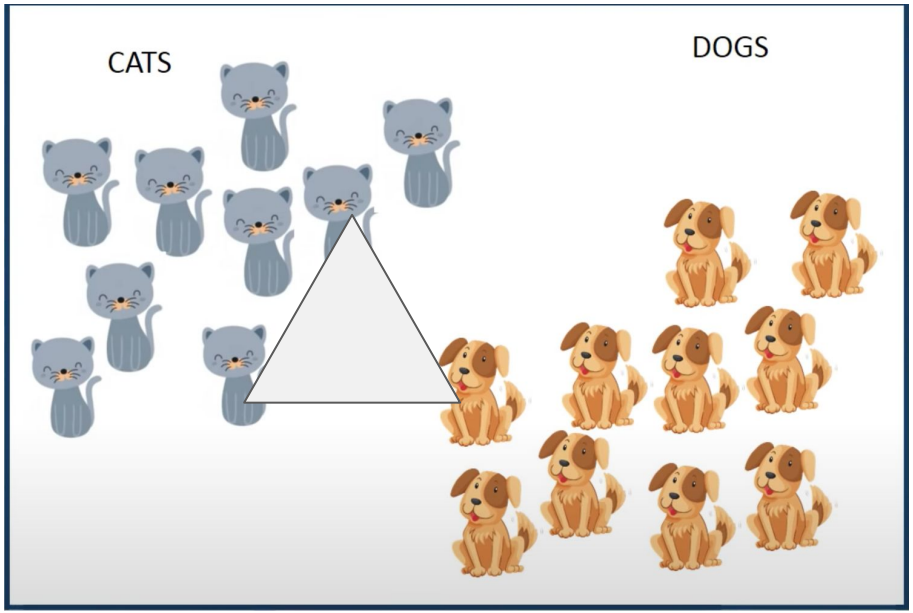
Sharpness of claws



Length of Ears

It's a Cat!

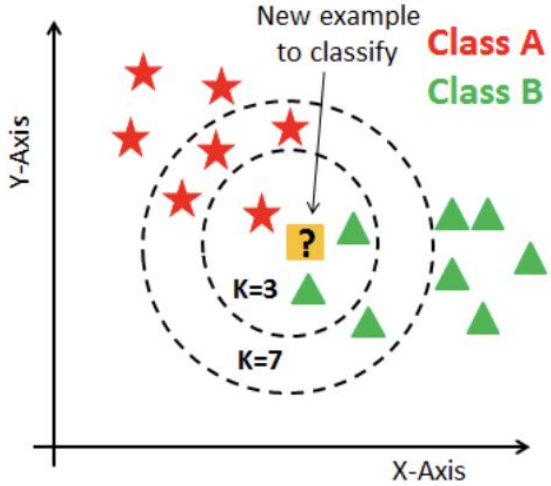
Sharpness of claws



Length of Ears

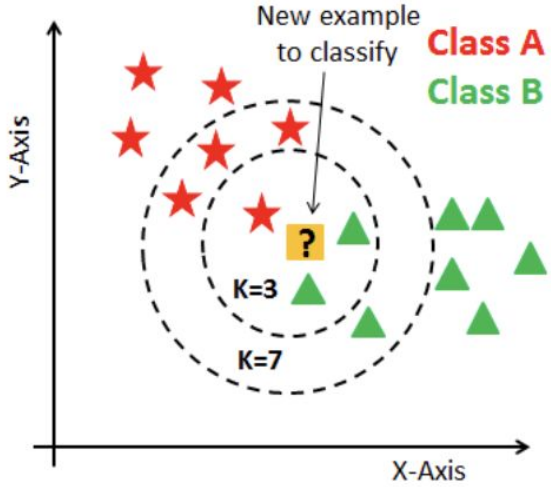
To summarize

- Goal: classify the new data point based on how its neighbors are classified



To summarize

- Observe that the outputs are categorical in our cats vs dogs example



Let’s formalize what we’ve discussed!

Now that we have the K nearest points

- We want to **classify** our new point **p**
- We look at the **classes** of those K nearest points
- Determine the class that is the **most common**
  - i.e. the *mode*
- We can predict that **p** belongs to this class!

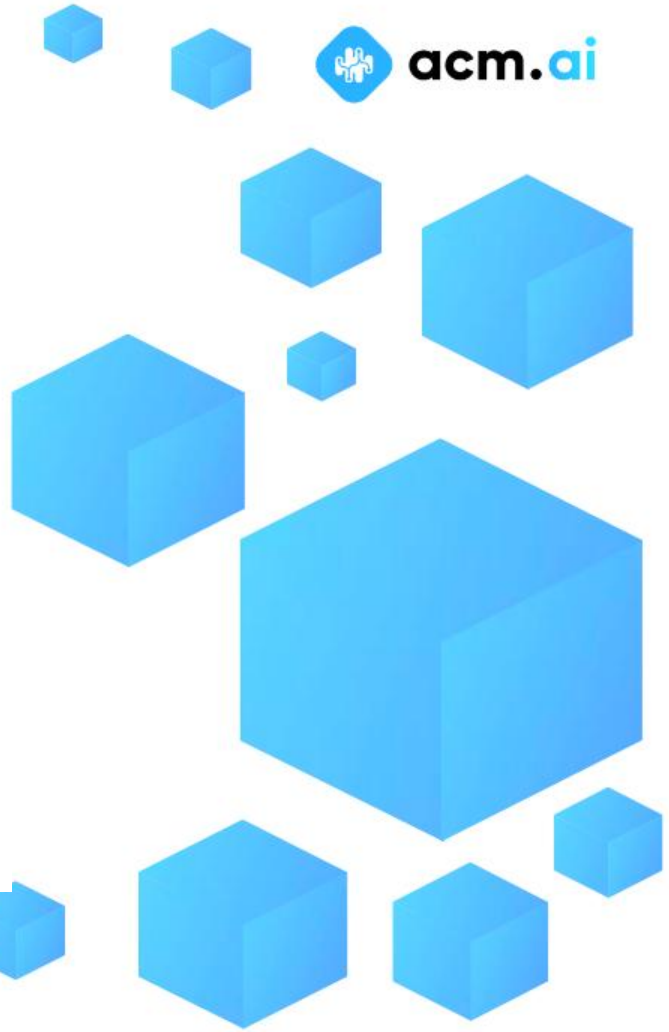
acm.ai

How would you determine what a neighbor is?

How would you determine what a neighbor is?

- Our data-points live in 2 dimensional space
- We can compute distances using the Euclidean distance measure
- So for our points (x1, y1) and (x2, y2)

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



We know what ***neighbors*** are, what's the ***nearest*** part?

The **nearest** neighbors to a given point ***p*** are the ones which have the **minimal Euclidean distances** to the that point

What does the **K** in KNNs represent?

K is just a **hyperparameter** (a value that *we* choose) to determine **how many** nearest neighbors to look at

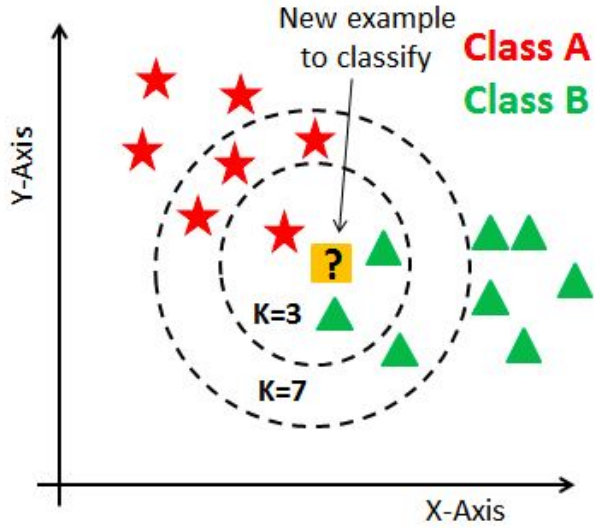
The algorithm to find the K nearest neighbors

- Let's say we choose one point **p** = (X, Y)
- Compute the distance between **p** and every other point in our data set using the Euclidean formula
- Choose the *K* data-points that are the closest to **p**
- We can then classify **p** using those K nearest points

Putting it all together

Given a dataset of points and a new point **p** to classify:

1) Choose a value for **K**
2) Compute the distances between **p** and every point in the dataset.

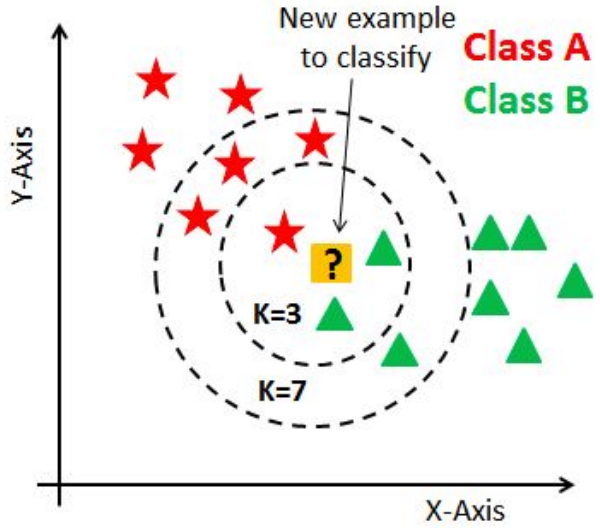


Putting it all together

3) Using these distances, determine the K points that are closest to **p**

4) Look at the classes of these K points and choose the class that is the most common

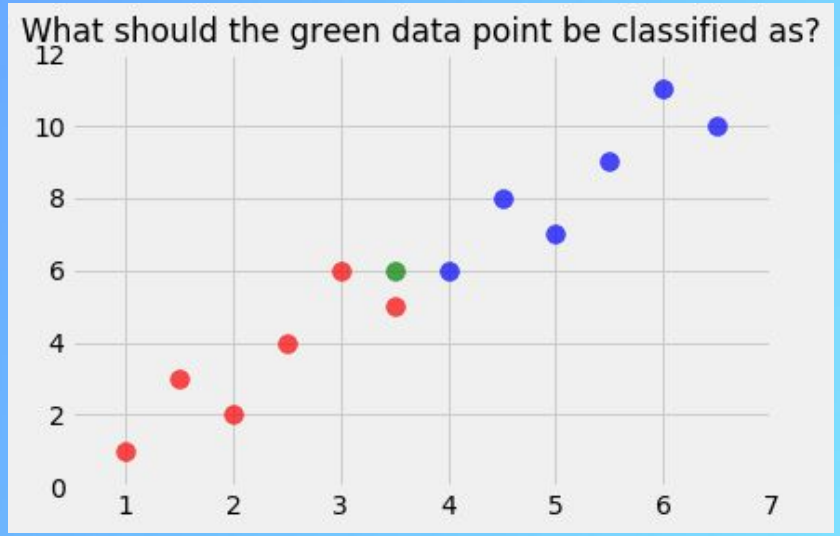
5) This is the class that **p** belongs to!



Quick quiz

What should the green data point be classified as? Choose k = 3.

a. Red
b. Blue
c. Green
d. None of the above

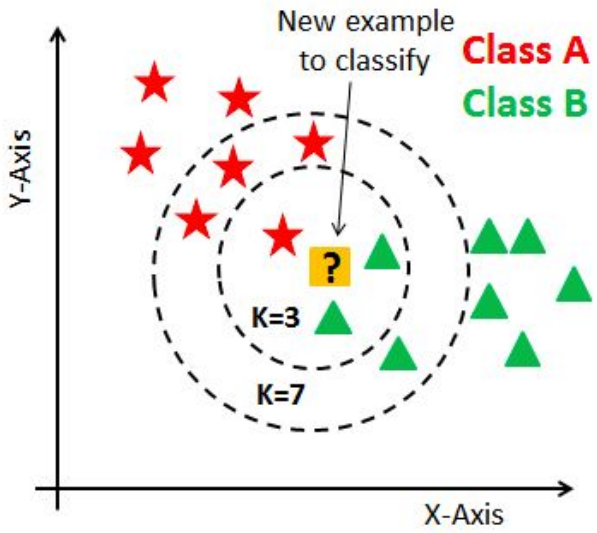


acm.ai

The Hyperparameter “K”

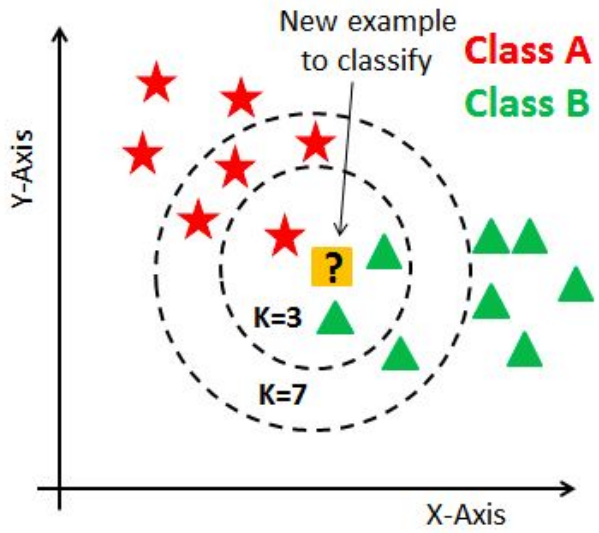
How do we choose the factor ‘K’?

- Choosing the right value of k is important for better accuracy
- Here’s an example of a dilemma
  - At k = 3, we classify ? as a triangle
  - But at k = 7, we classify ? as a star

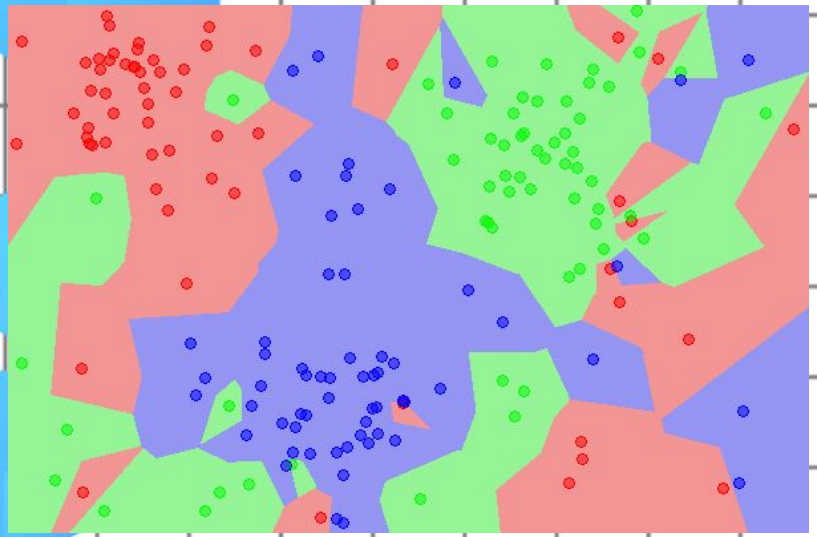


How do we choose the factor ‘K’?

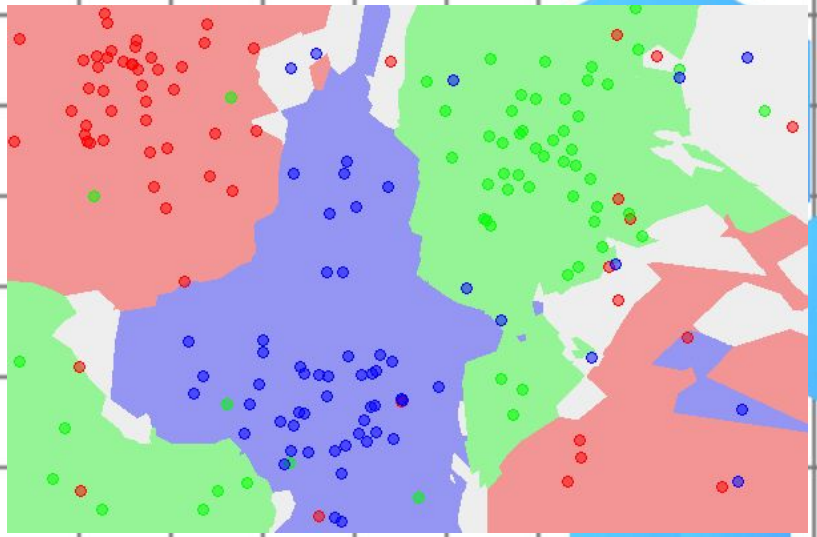
- So which one is correct?
  - We should try both and see which gives us a better accuracy



K = 1



K = 5



How do we choose the factor ‘K’?

How do we choose the factor ‘K’?

- What would happen if we pick a k that is too low?
  - Think about noise/outliers in the dataset
- What would happen if we pick a k that is too high?
  - As an extreme, think about if k = size of dataset

How do we choose the factor ‘K’?

- A general rule of thumb for choosing k
  - k = sqrt(n), where n is the total number of data points
  - **Odd** value of k to avoid a potential tie in the voting process
  - This is a good *starting point* for k. You should try different values to see what gives the best result.

Quick Quiz

If there are 170 data points in your KNN algorithm, what k value should you pick?

a) 11
b) 12
c) 13
d) 14

acm.ai

Where does the *training* come in?

- The point of training is to give the model information from our dataset
- In the case of of a KNN we can just use the dataset directly

Where does the *training* come in?

- So there's no explicit training process
- That's why KNN is called a **lazy** learning algorithm as opposed to an **eager** learning algorithm

acm.ai

```python
class KNearestNeighbor(object):

  def __init__(self):
    pass

  def train(self, X, y):
    self.X_train = X
    self.y_train = y

  def test(self, x_test, k=1):

    dists = np.linalg.norm(self.X_train.T - x_test, axis=1).T
    sortedIdxs = np.argsort(dists)
    closest_y = self.y_train[sortedIdxs[:k]]
    y_pred = np.argmax(np.bincount(closest_y));

    return y_pred
```

Can you identify any potential pitfalls of KNNs?

Pros and cons

- Pros:
  - No training required!
  - Simple to implement
  - Works well for small datasets
- Cons:
  - Does not scale well to large datasets
  - Curse of dimensionality – computationally and in representations

Case study

How does KNN work in practice?

- Consider a dataset having 2 features: **X** and **Y**
- Each sample is labeled as **0** or **1**

X	Y	Label
51	167	0
62	182	1
69	176	1
64	173	1
65	172	1
56	174	0
58	169	1
57	173	1
55	170	0

How does KNN work in practice?

57	170	?

- Given this train dataset, we have to classify this new point **p** as either in class 0 or 1 using KNN
- To find nearest neighbors, we will use Euclidean distance

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

How does KNN work in practice?

- Here, the last column shows the distance between each training sample and **p**
- Total number of data points = 9
- **k** = sqrt(9) = 3

X	Y	Label	Distance
58	169	1	1.4
55	170	0	2
57	173	1	3
56	174	0	4.1
51	167	0	6.7
64	173	1	7.6
65	172	1	8.2
62	182	1	13
69	176	1	13.4

How does KNN work in practice?

- To find the K-Nearest neighbors to **p**, we need to sort our table by each sample's distance to **p**

X	Y	Label	Distance
58	169	1	1.4
55	170	0	2
57	173	1	3
56	174	0	4.1
51	167	0	6.7
64	173	1	7.6
65	172	1	8.2
62	182	1	13
69	176	1	13.4

Recap of KNN

- Given a dataset and a new point to classify
- Compute the distance between **p** and each sample in the dataset
- Sort the dataset by each sample's distance to **p**

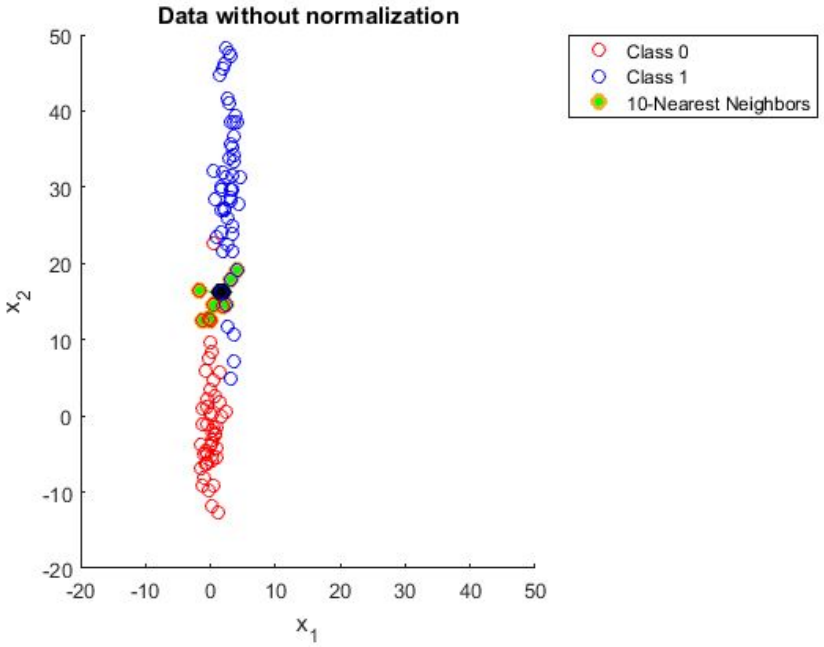
Recap of KNN

- Select **K** entries from our dataset closest **p**
- Find the most common classification (mode) of these **K** entries
- This is the classification we give to the test data point **p**

The Importance of Data Normalization for KNNs

What is the issue with this dataset?

Specifically, what is it about this dataset that might make our KNN model perform suboptimally?

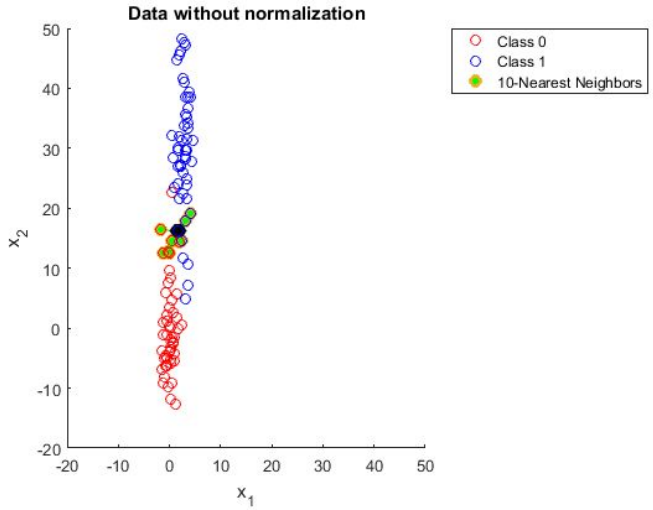


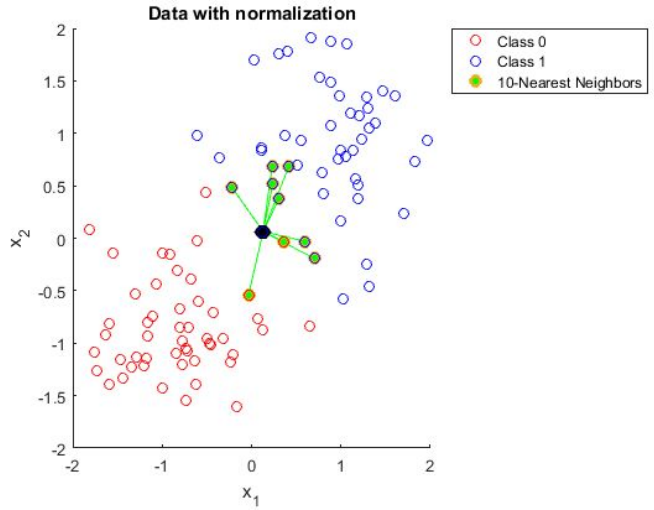
What do we mean by ‘normalization’?

- ***f*** *is one of our features (a column)*
- μ is the mean value in that column
- σ is the standard deviation of that column

$$f_{normalized} = \frac{f - \mu}{\sigma}$$

Normalize the data beforehand?





Ethics: Bias in AI

Quick Review of Ethics!

Remember the ML Pipeline?

The next two weeks of ethics content, will be focused on the first stage: Data!

Can anybody tell me about the three parties: Developer, Deployer and User?

Ethics

AI Models can be biased when collected data does not accurately represent the problem at hand:

- Data collection was not representative
- Data reflects bias that is already present in society

Ethics

- COMPAS: Used an algorithm to predict whether a defendant would be rearrested for a similar crime. It predicted twice as many false positives for black offenders than it did for white offenders.

- Amazon: Used a hiring algorithm that looked at the resumes submitted over a 10 year period, and since most of the resumes were submitted by men, the algorithm was shown to be biased against women.

Game Time

KNN Demo

tinyurl.com/f22-ws2

acm.ai

Thank you!

Feedback Form: tinyurl.com/f22-feedback