Quiz, week 11

Supervised Learning, pt 2

INST414 - Data Science Techniques

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

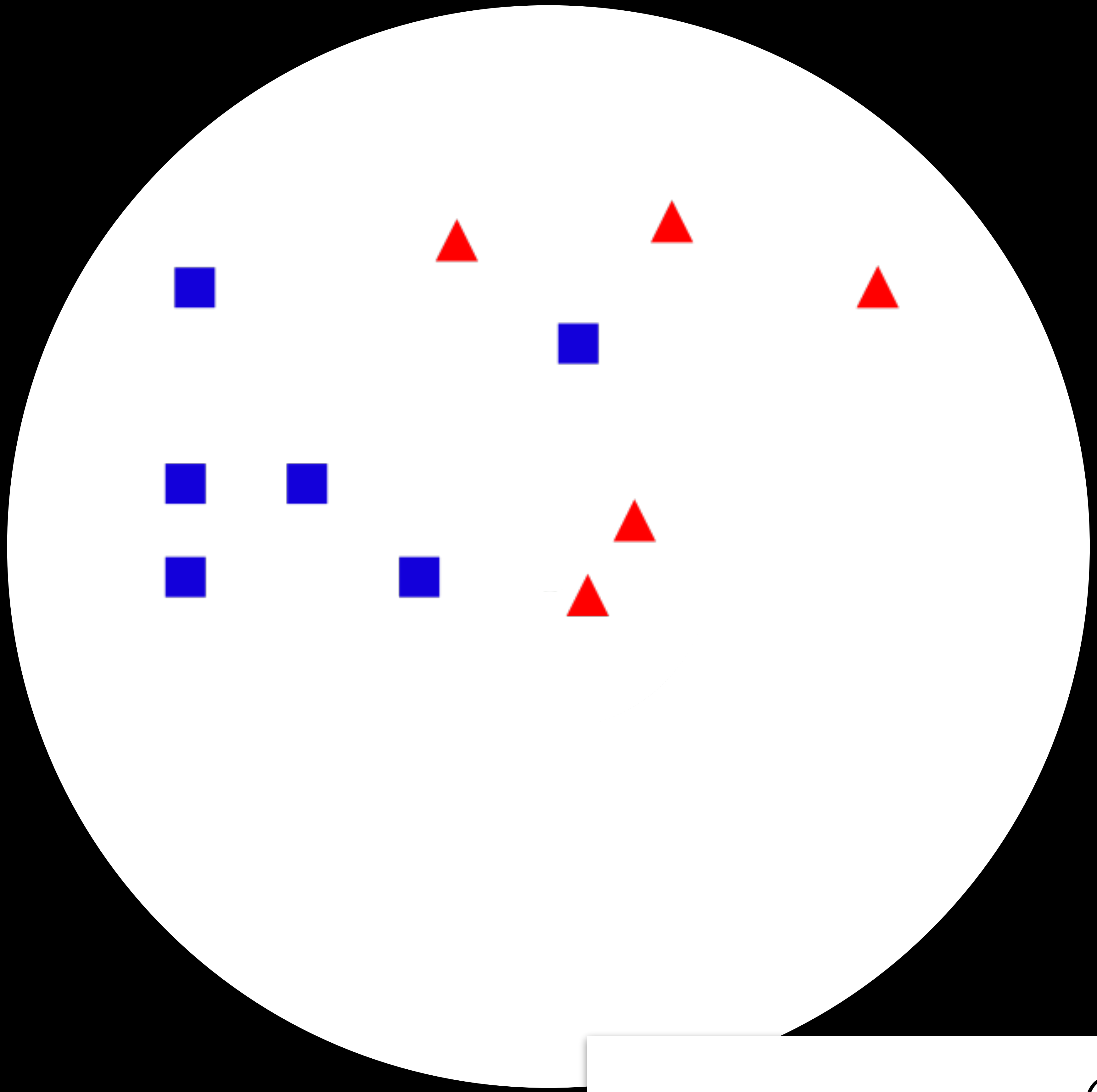
- Differentiate between classification and regression in supervised learning
- Describe how voting is used in for k-nearest neighbors classification
- Differentiate binary, multi-class, and multi-label classification
- Define overfitting and describe its impact on generalizability

Also called the “training data”

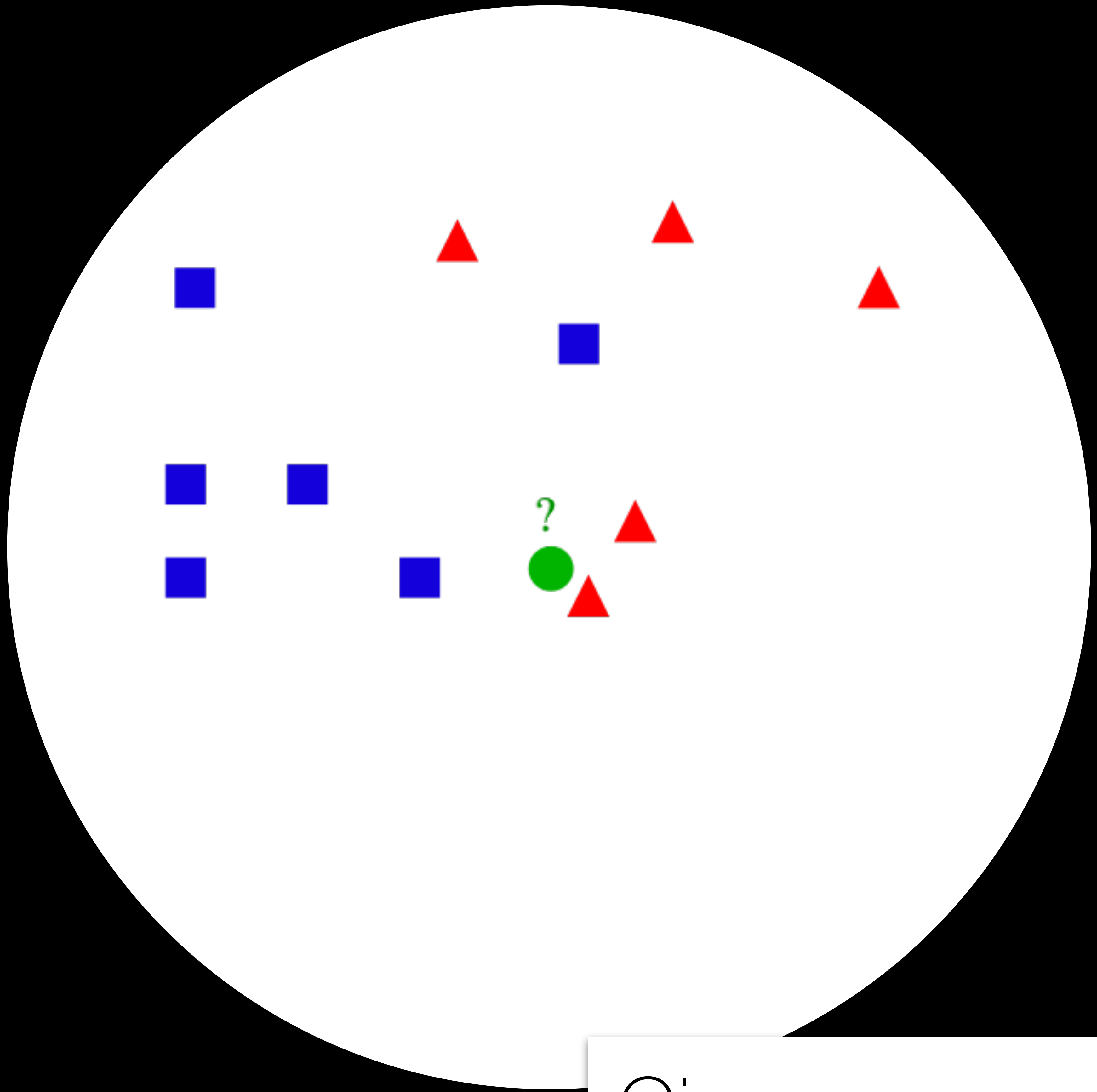
What’s a simple rule for classifying new data based on our sample set?

Use the label of the “closest” sample

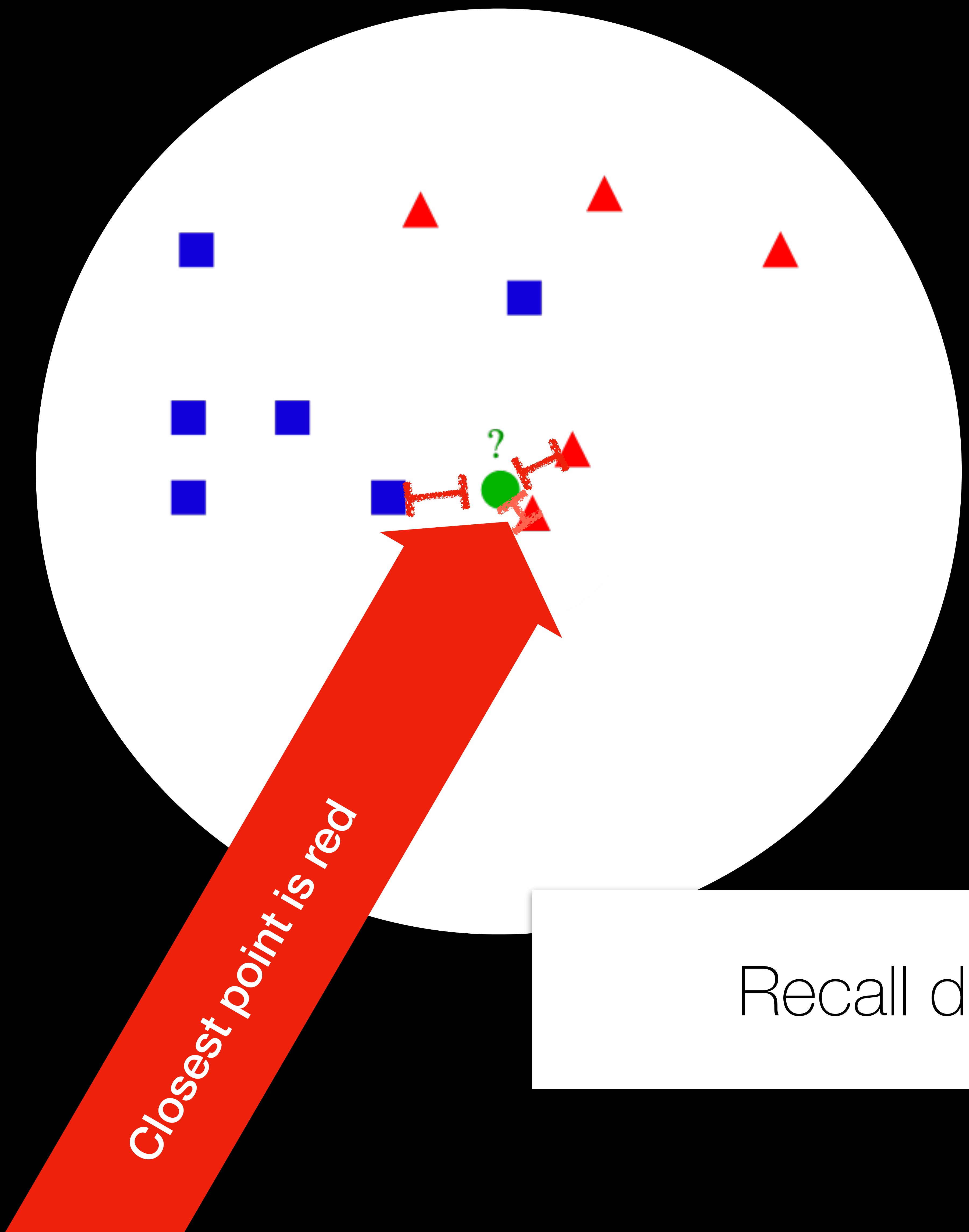
The “nearest neighbor” approach



Given a dataset of two classes
(blue/red labels)

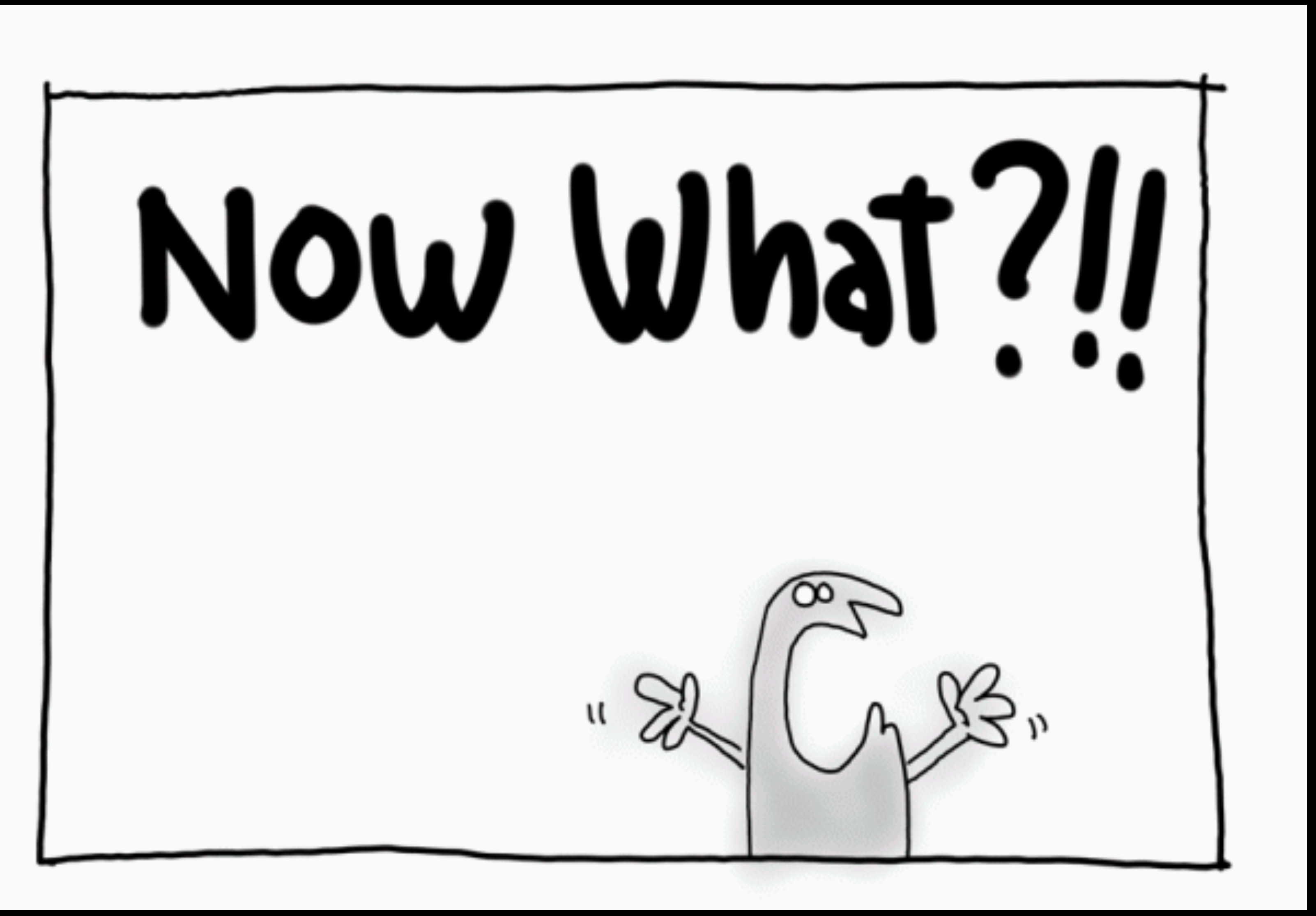


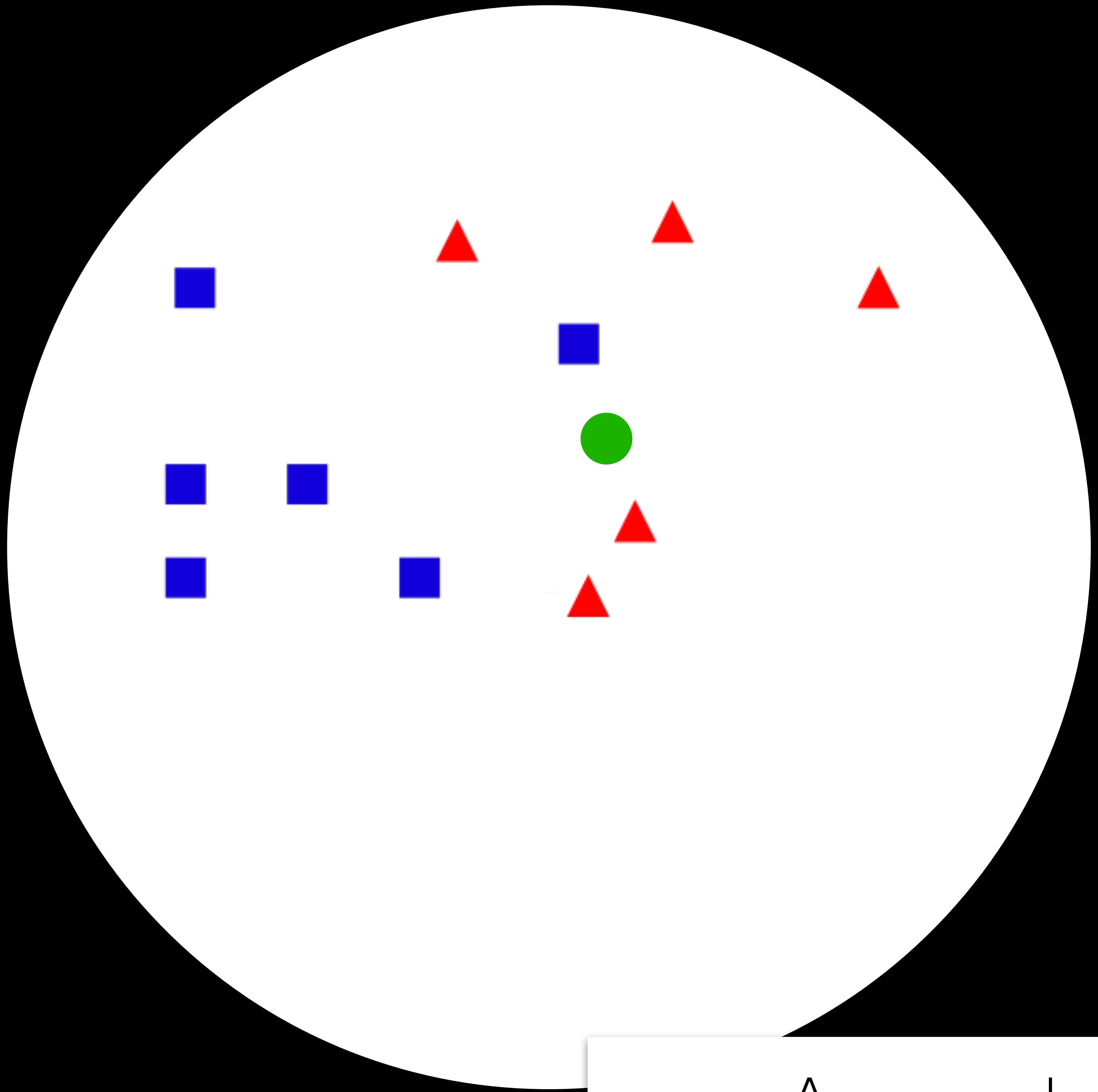
Given a new sample (green data point), what label should it be?



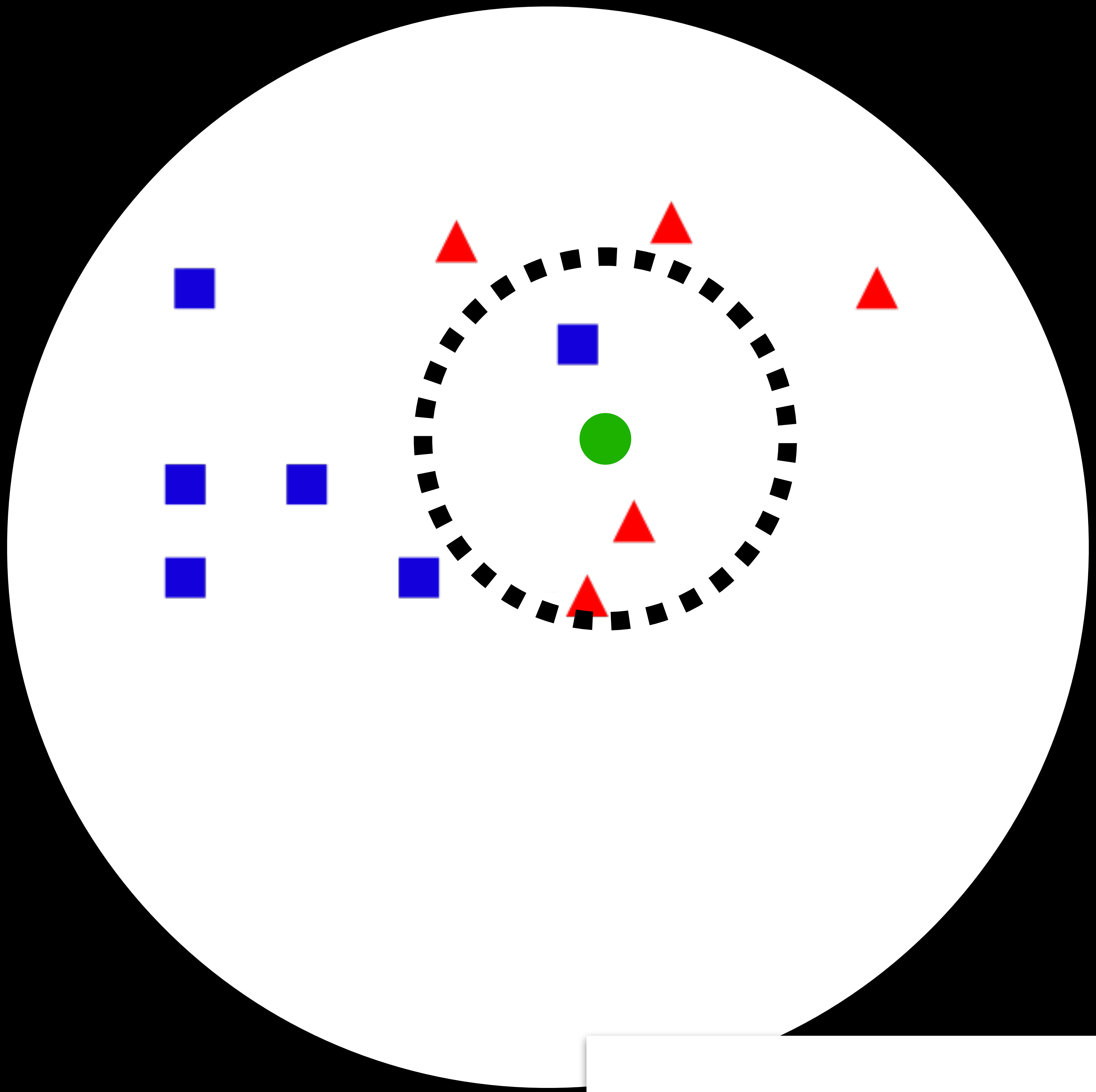
Closest point is red

Recall distance metrics from Module 3

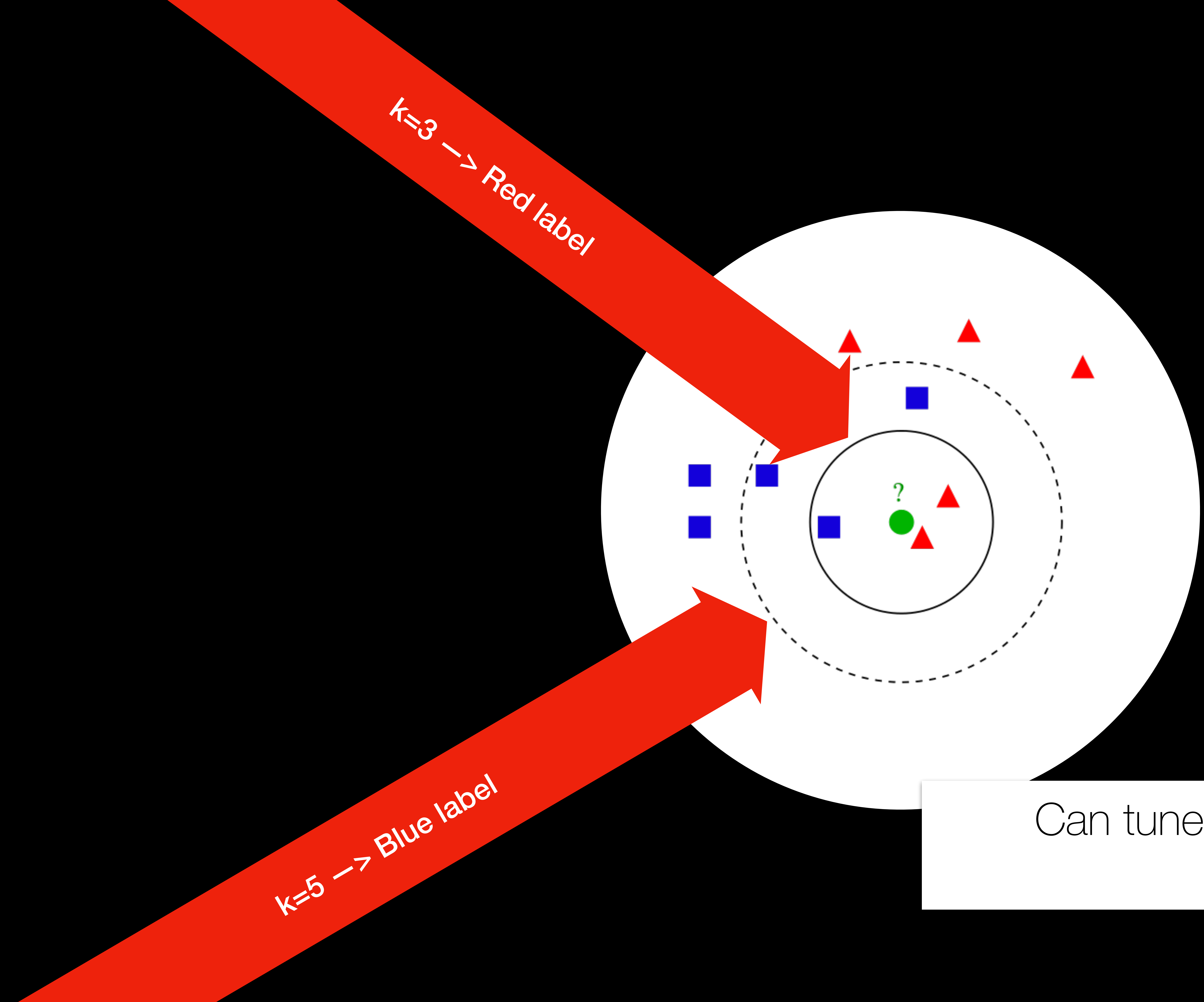




A sample could be equidistant between points



Take a vote among k-nearest neighbors

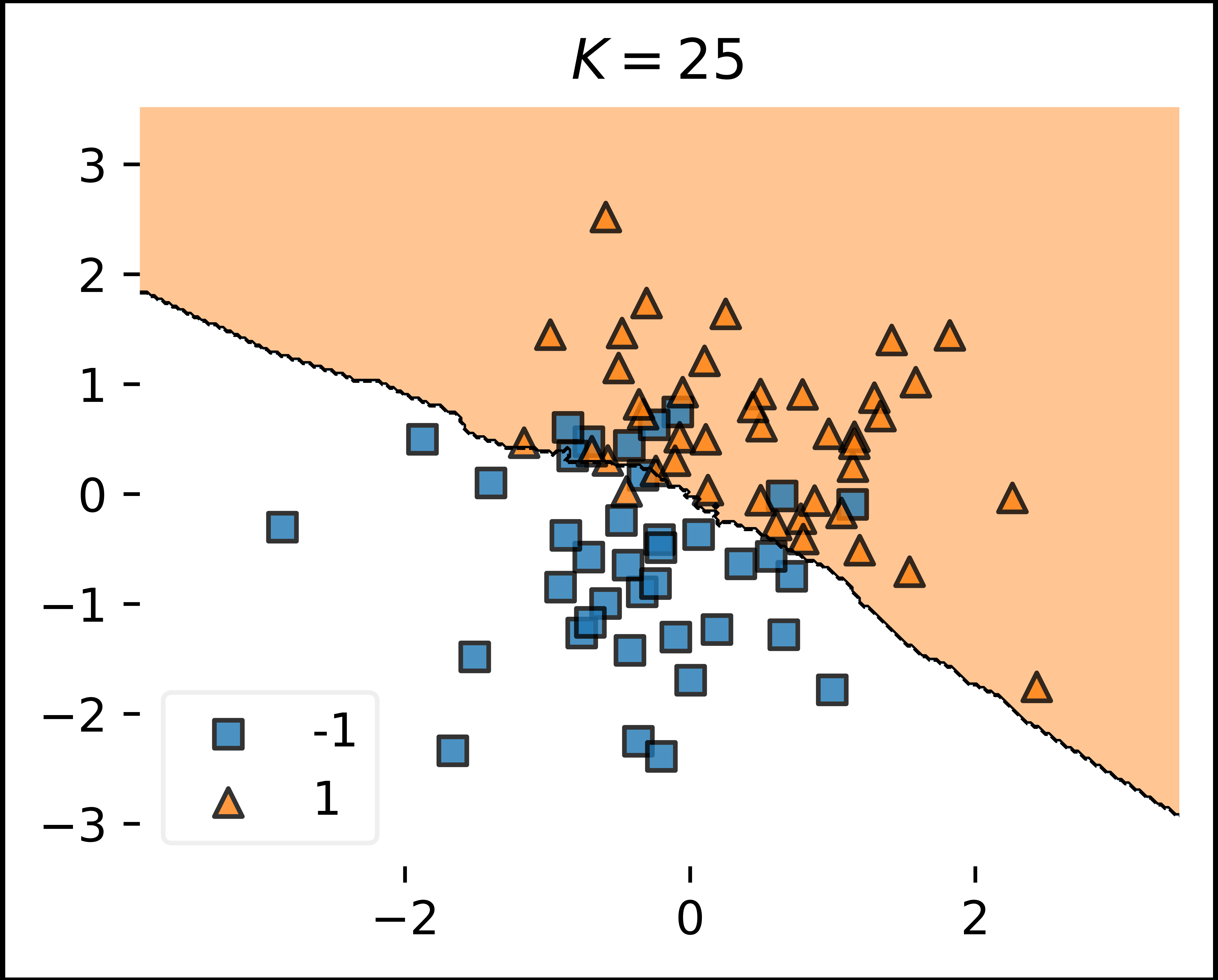


k=3 —> Red label

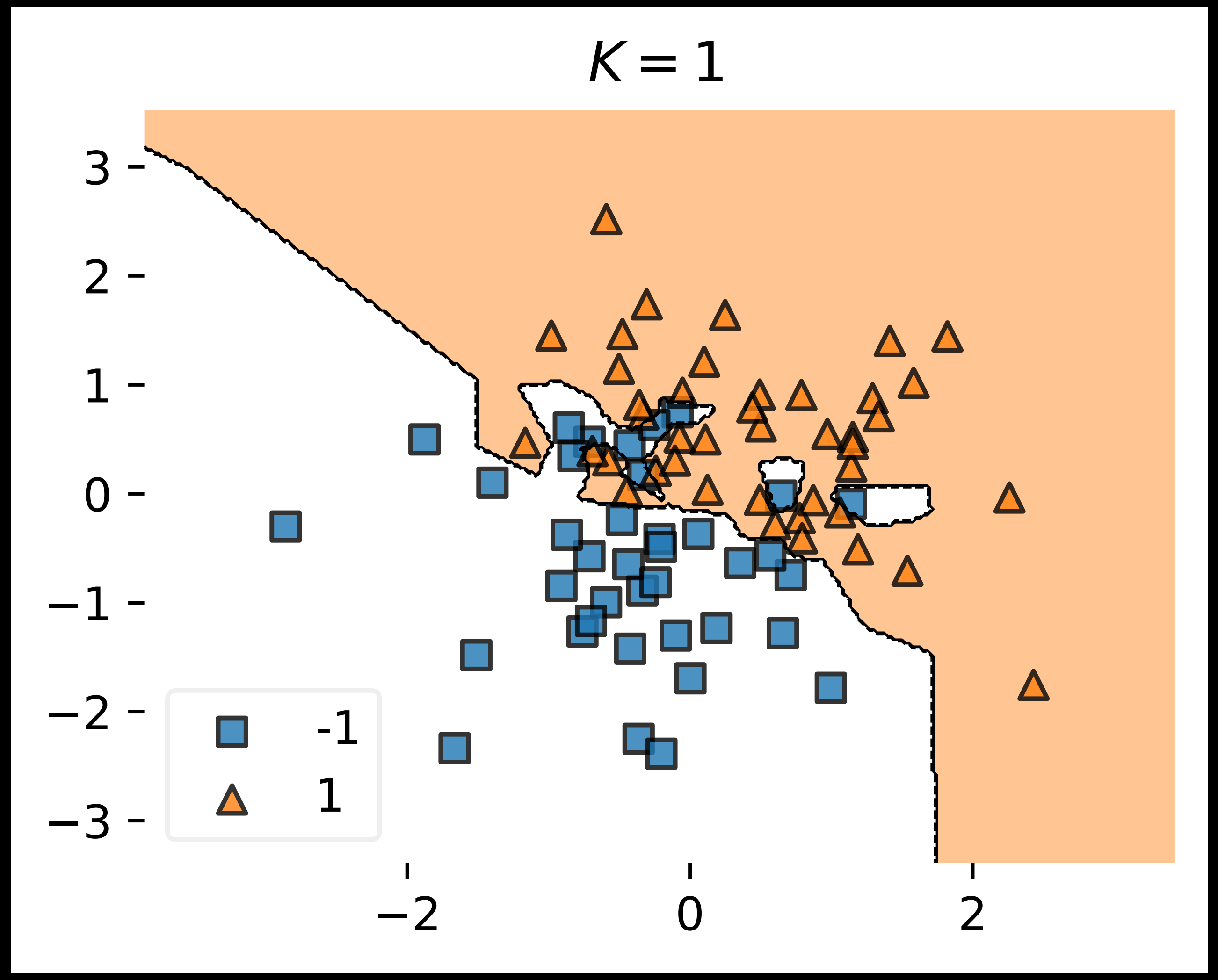
k=5 —> Blue label

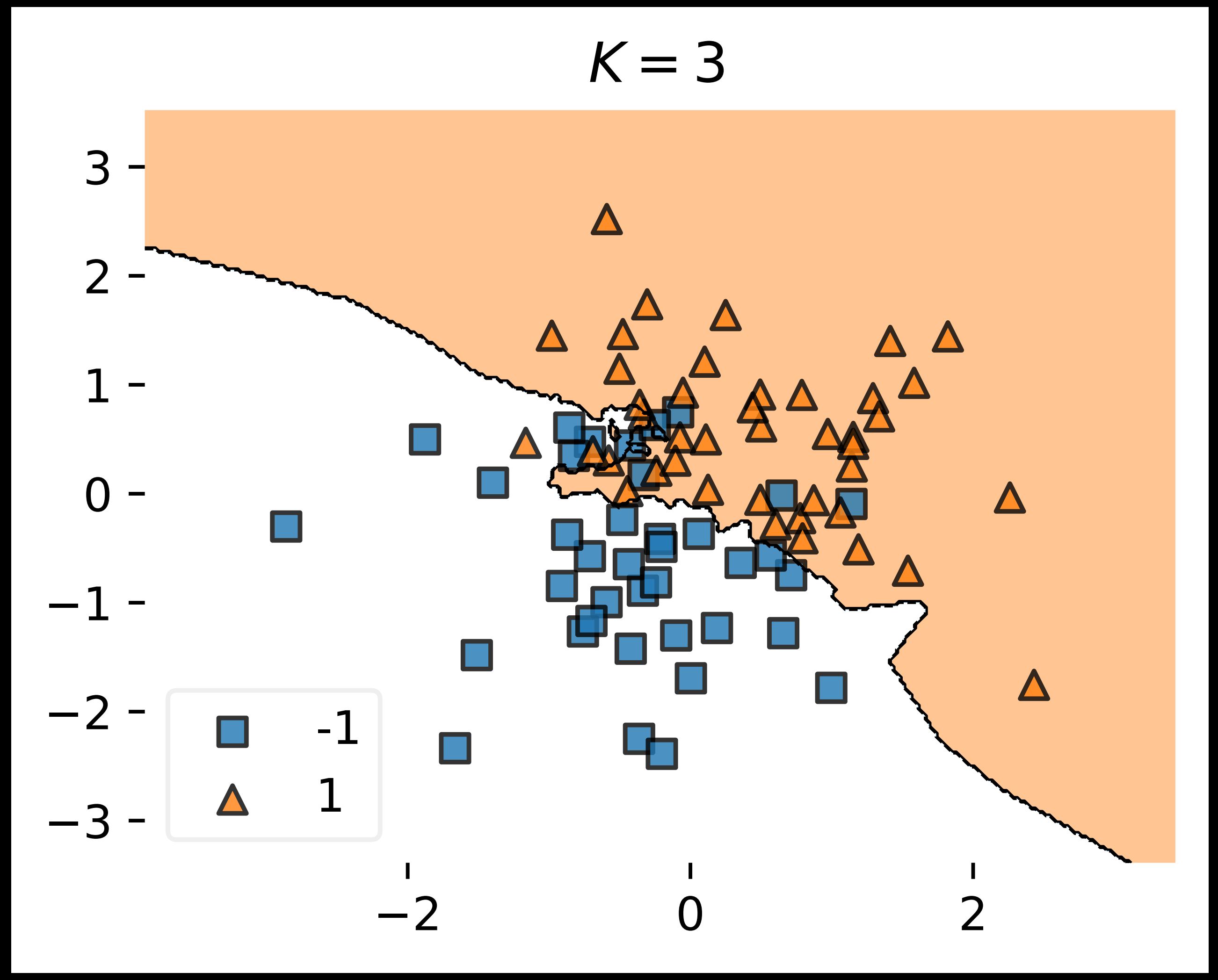
Can tune k-value as needed to improve classification accuracy

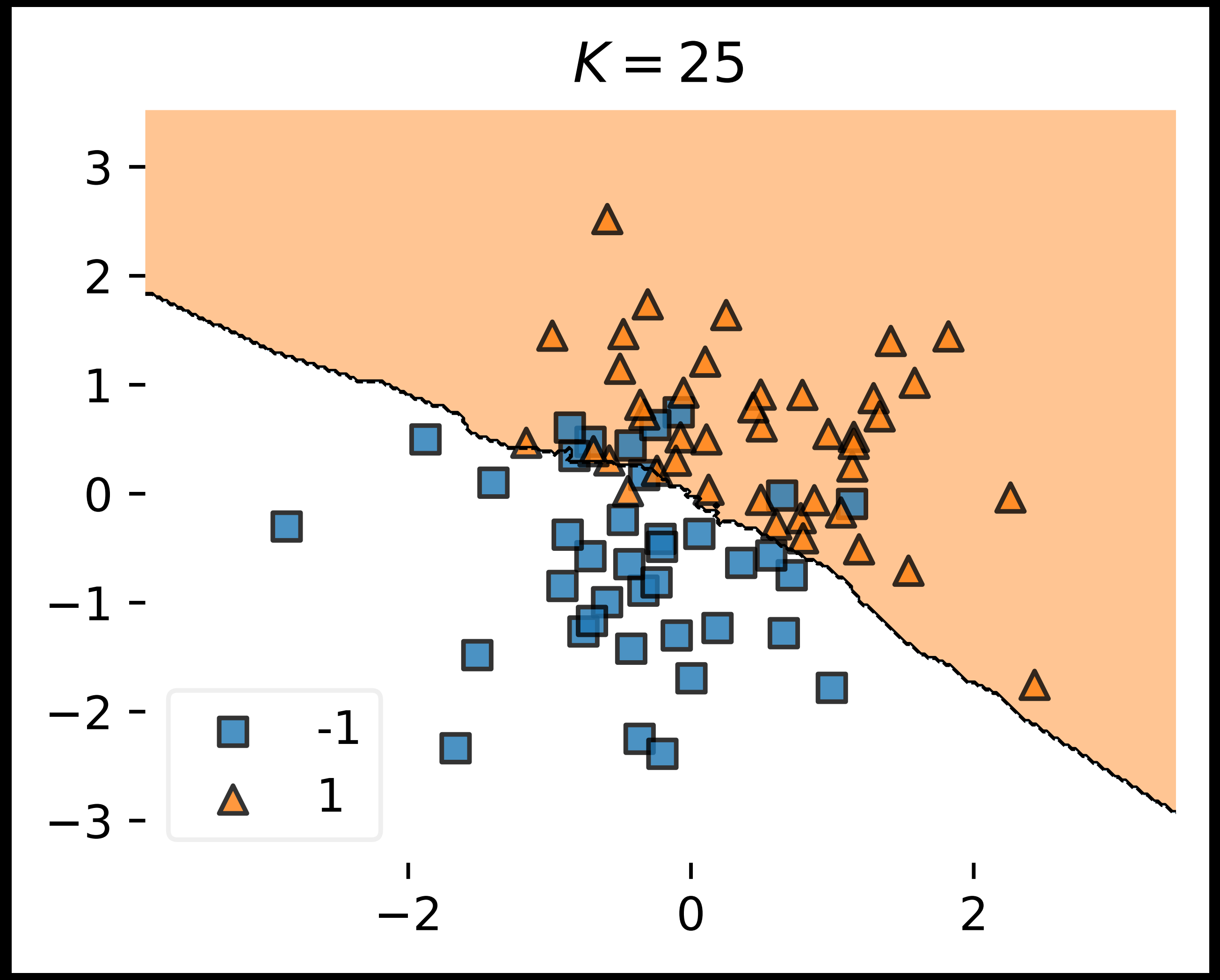
KNN for Classification (example: 2 features)



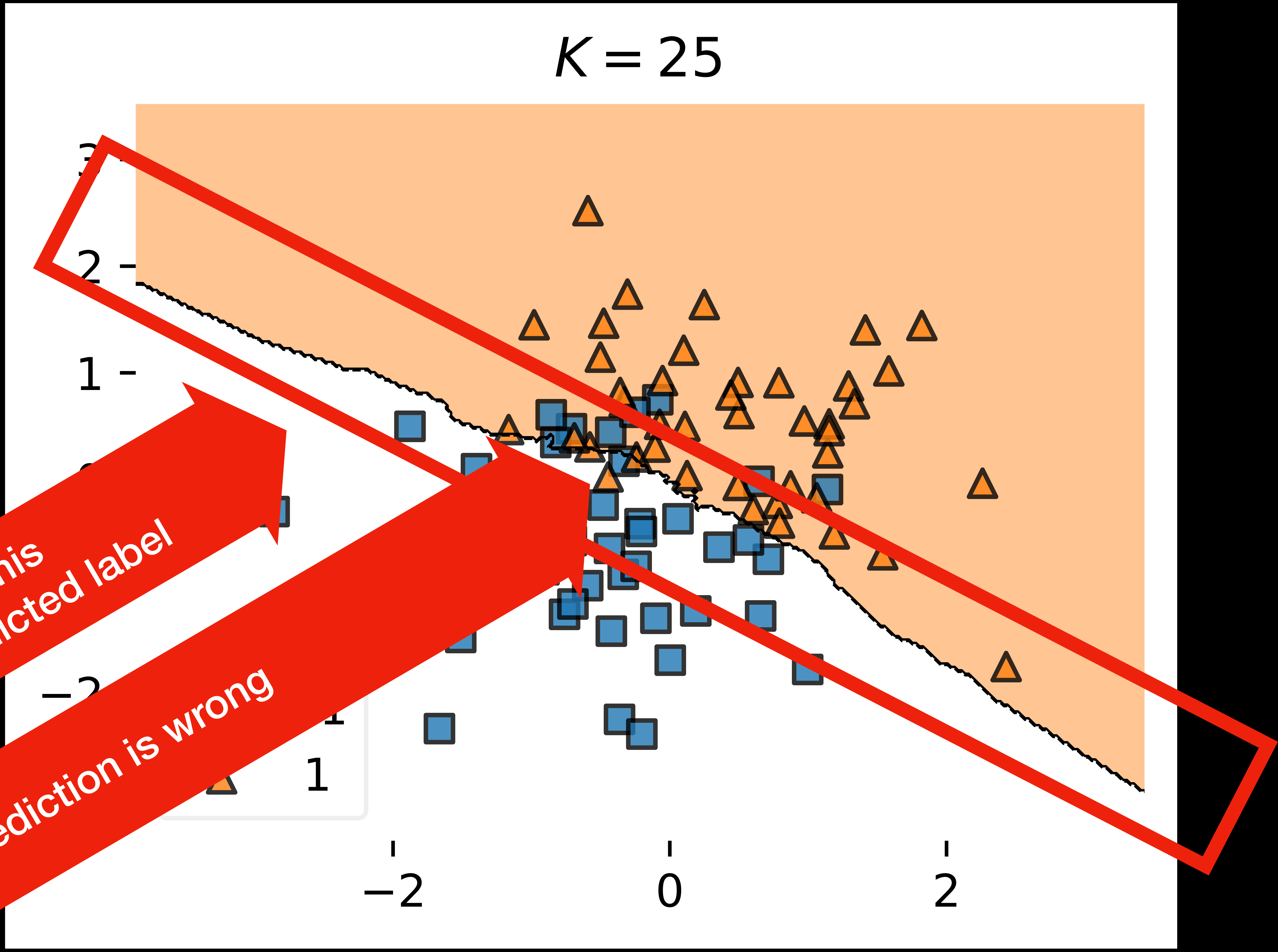
KNN for Classification (example: 2 features)





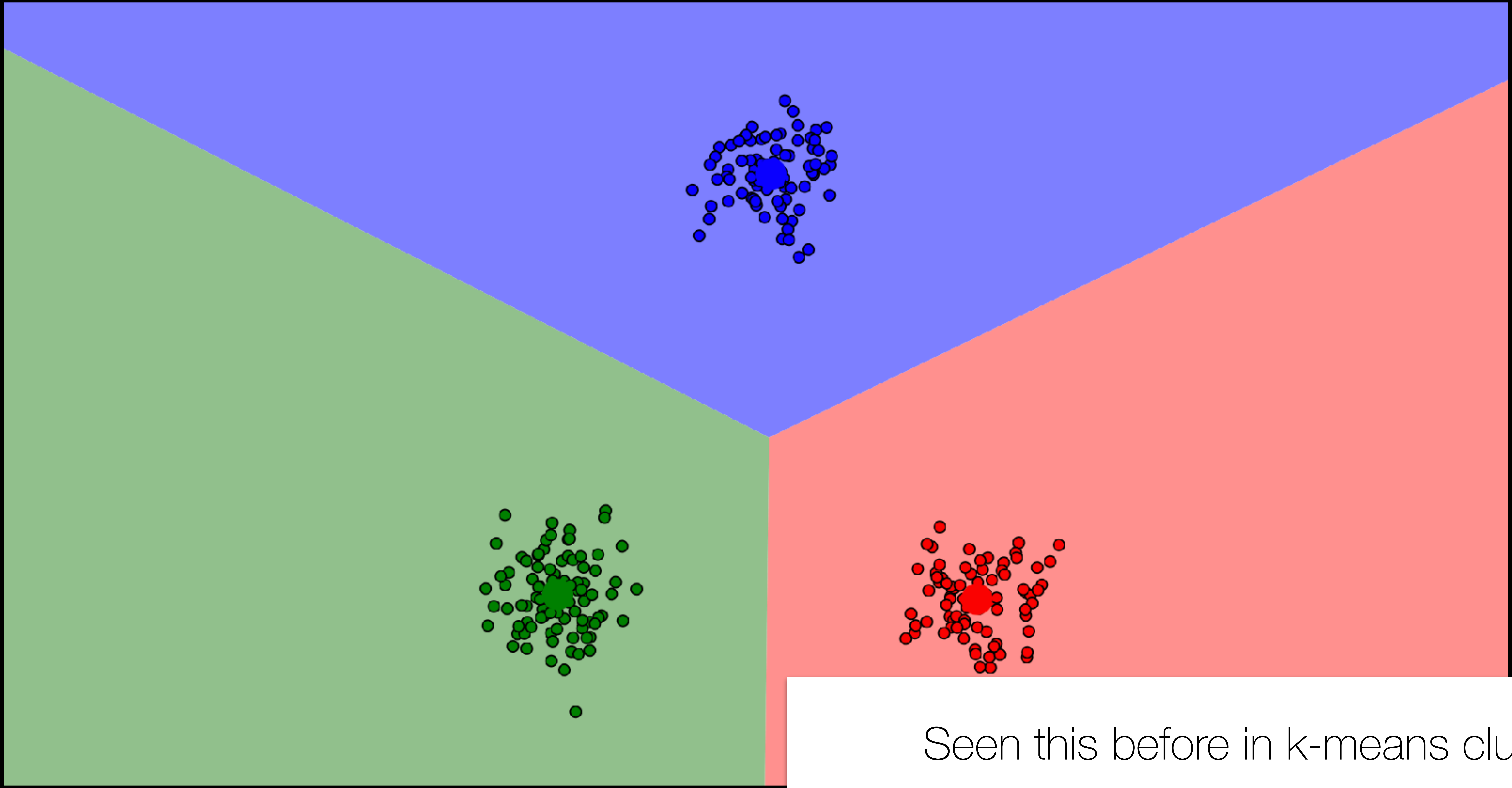


Decision Boundary



All elements on the same side of this decision boundary have the same predicted label

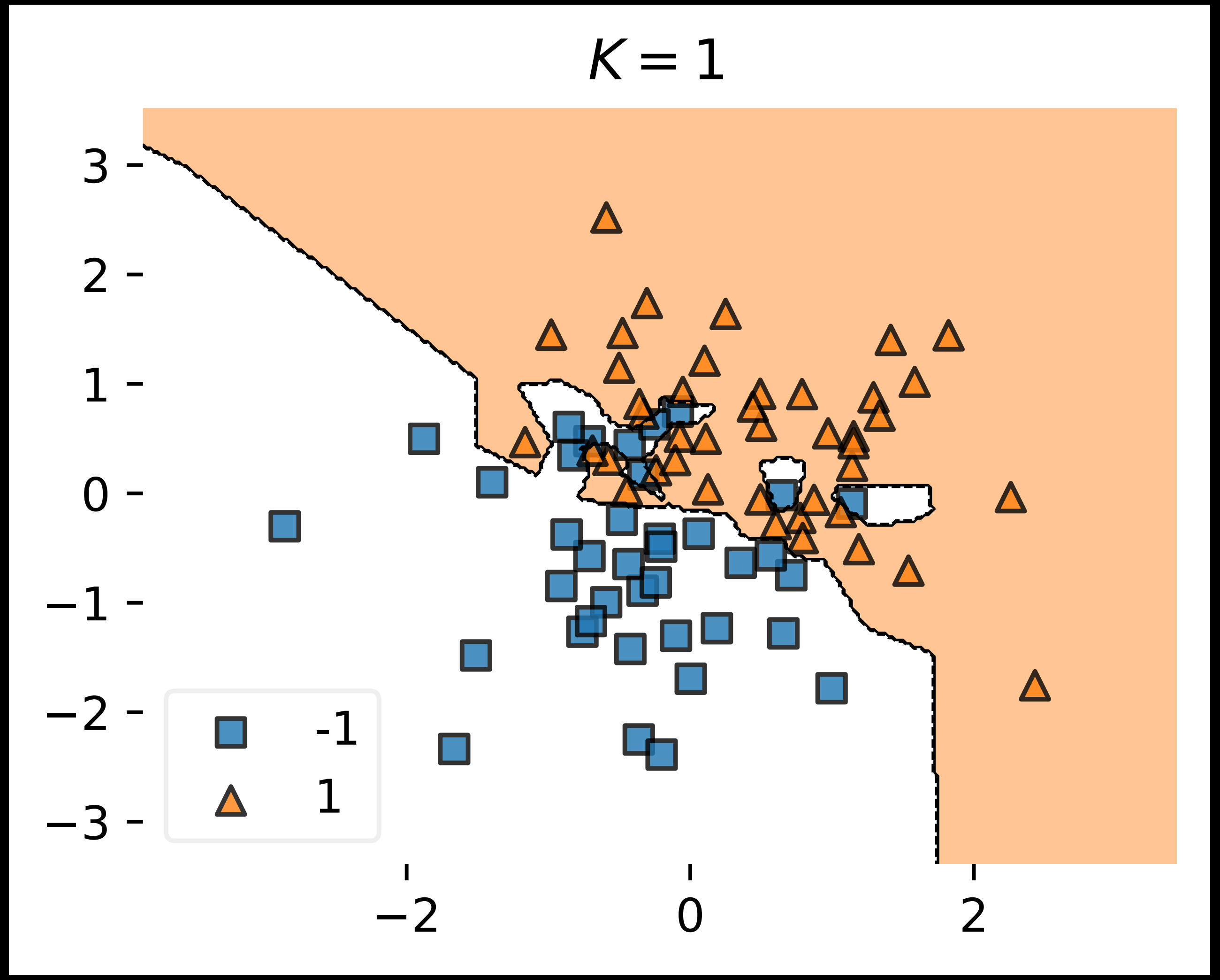
Event if prediction is wrong



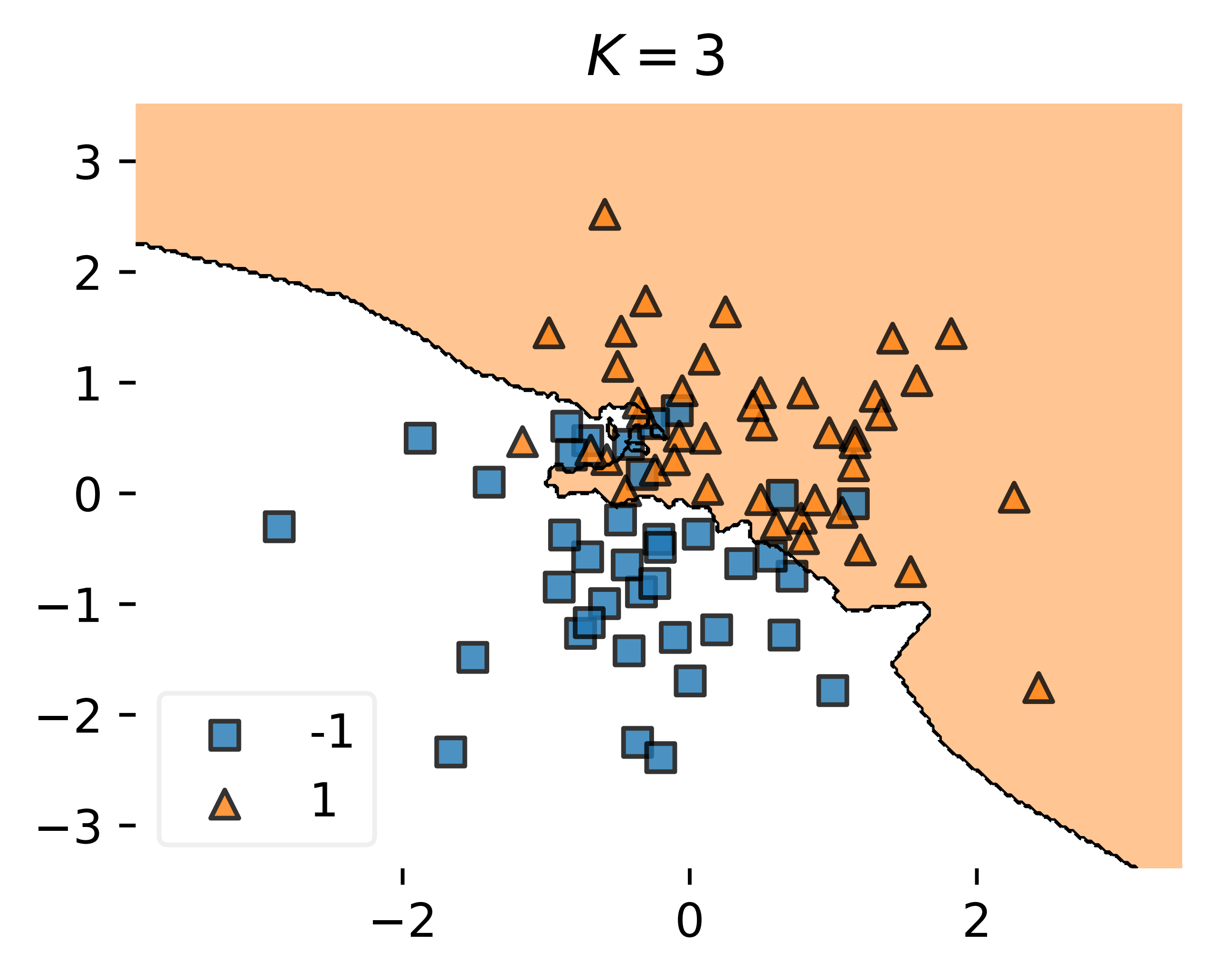
Seen this before in k-means clustering

KNN Classifier

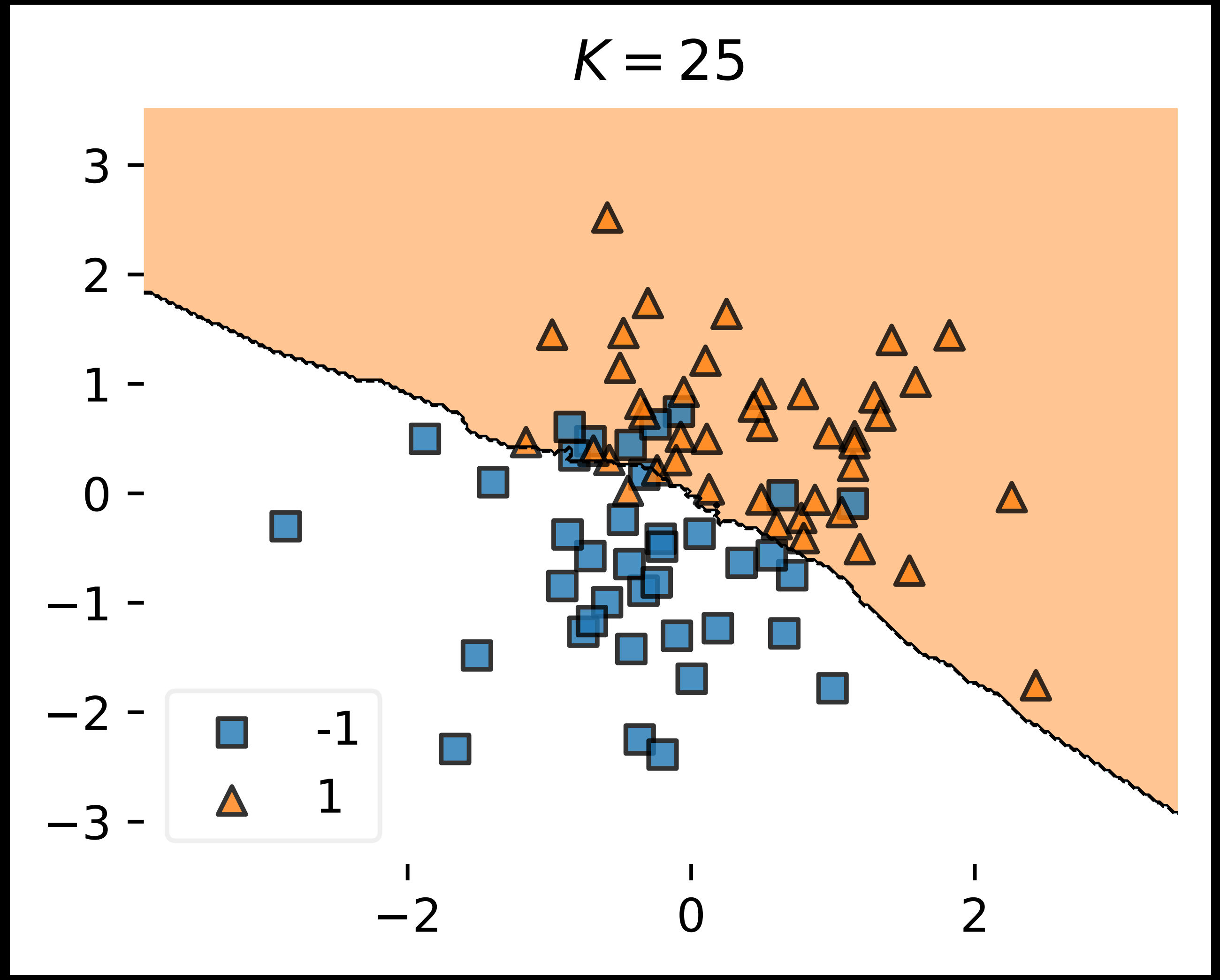
- Model complexity controlled by k (# of nearest neighbors)
- Larger k indicates less complex decision boundary



$K = 1$

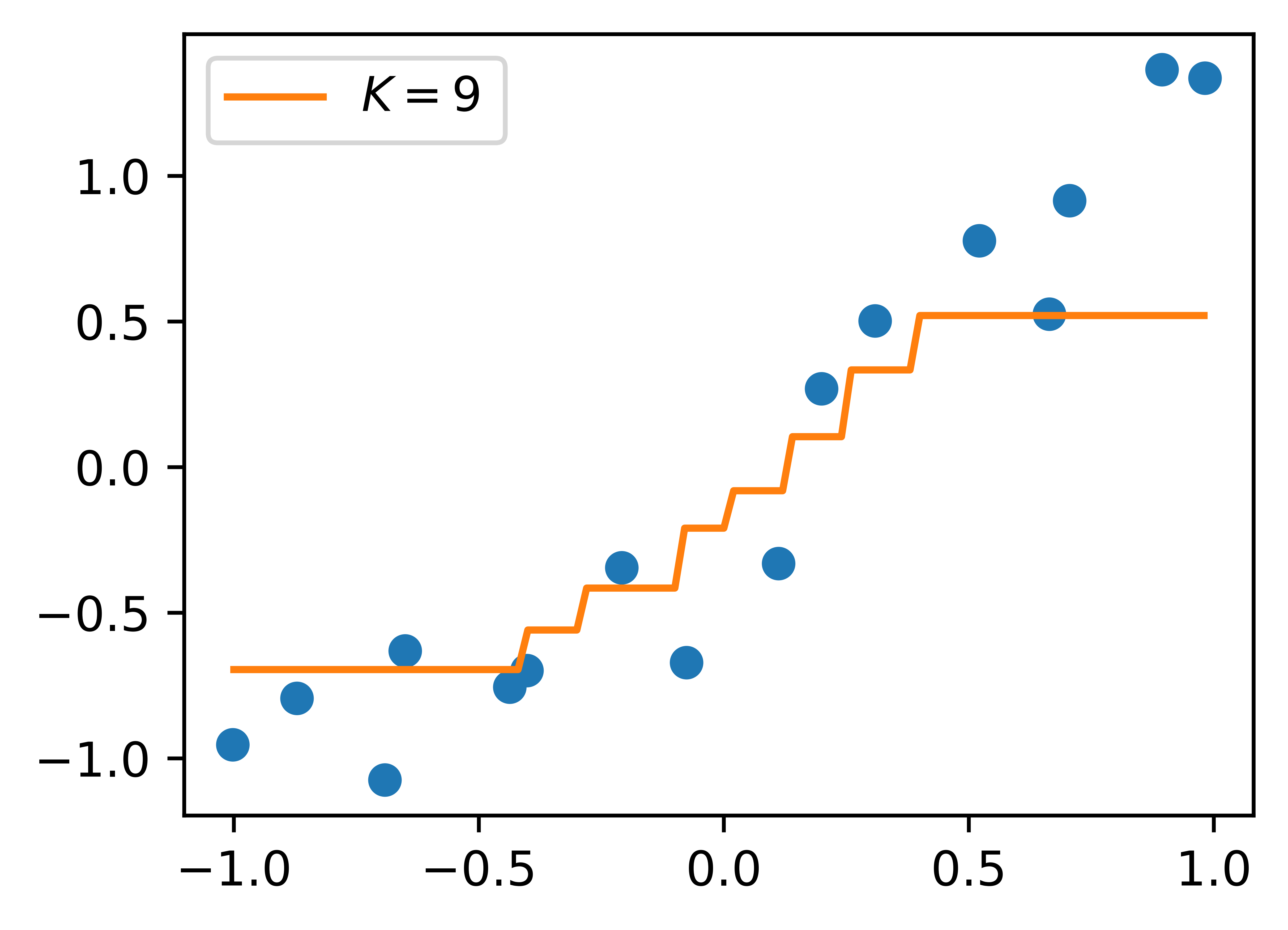


$K = 3$

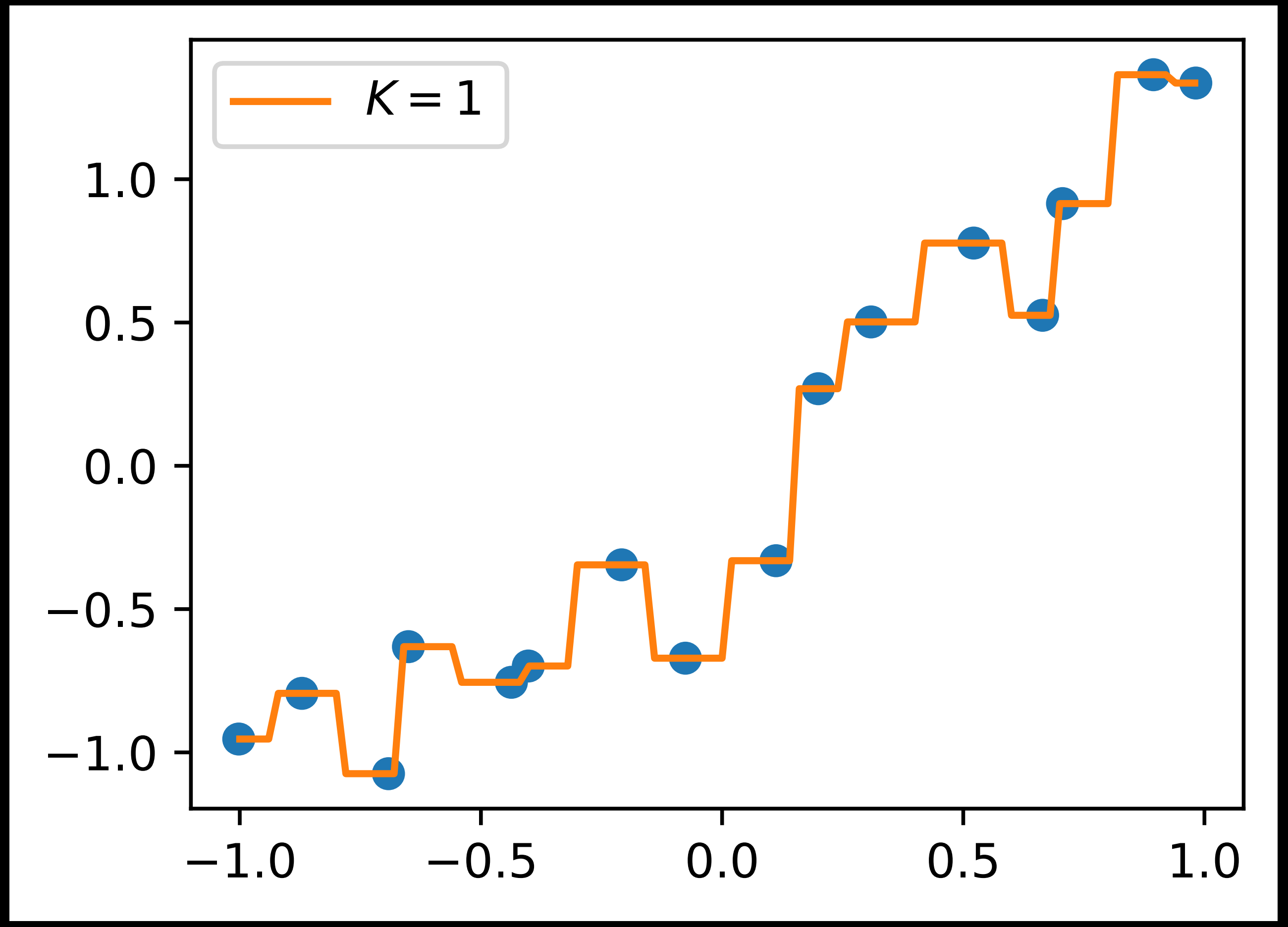


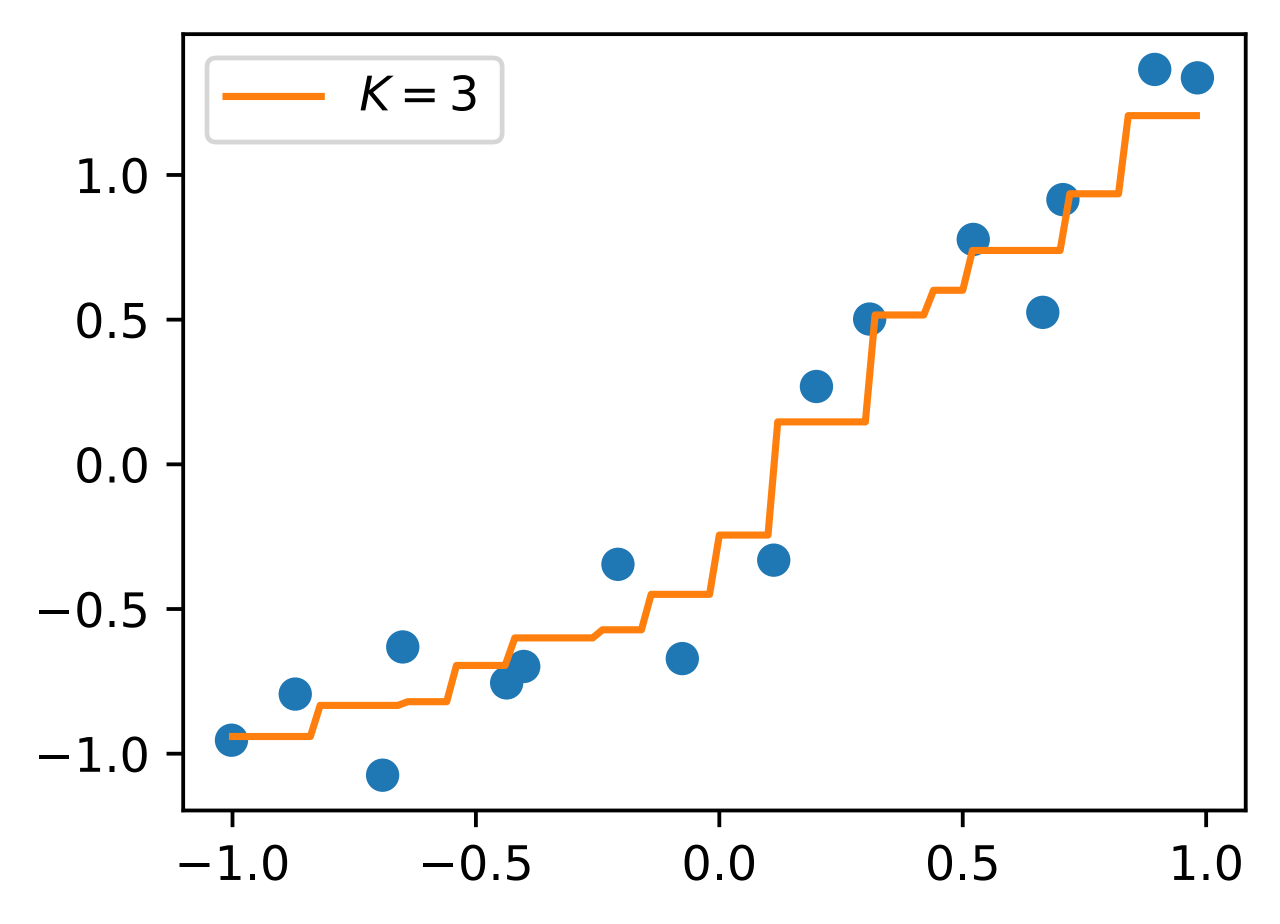
$K = 25$

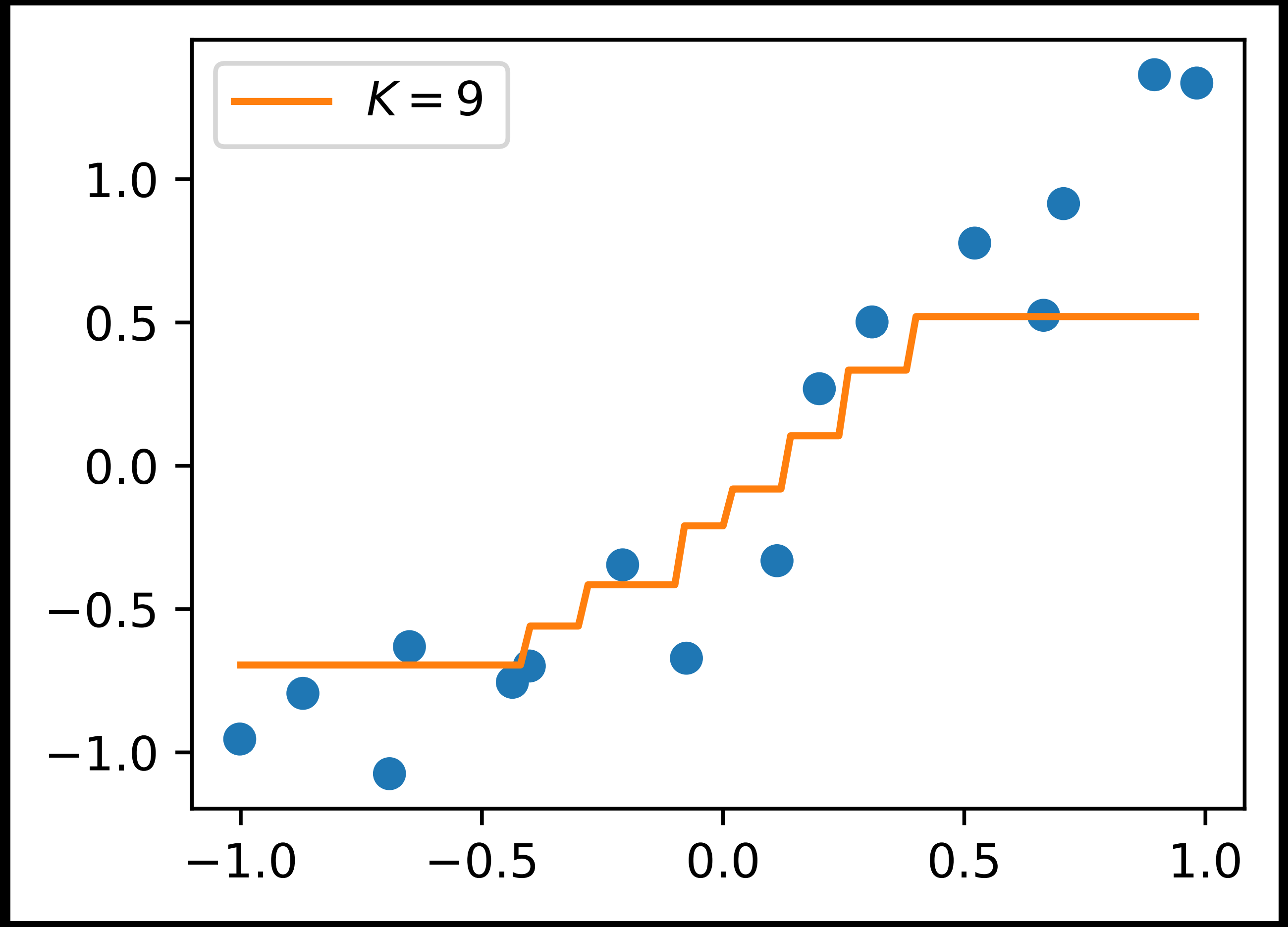
KNN for Regression (example: 1 feature)



KNN for Regression (example: 1 feature)

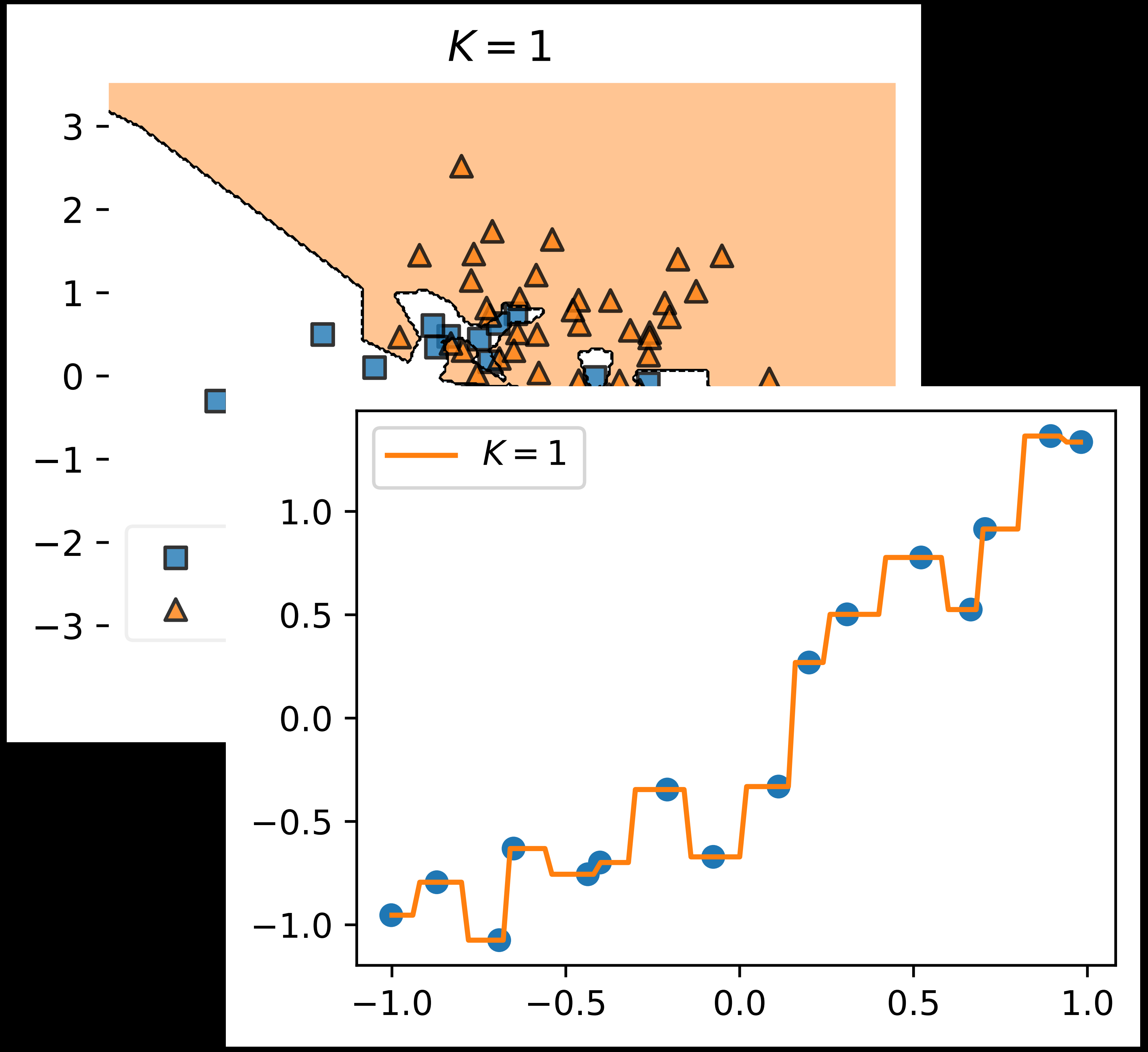


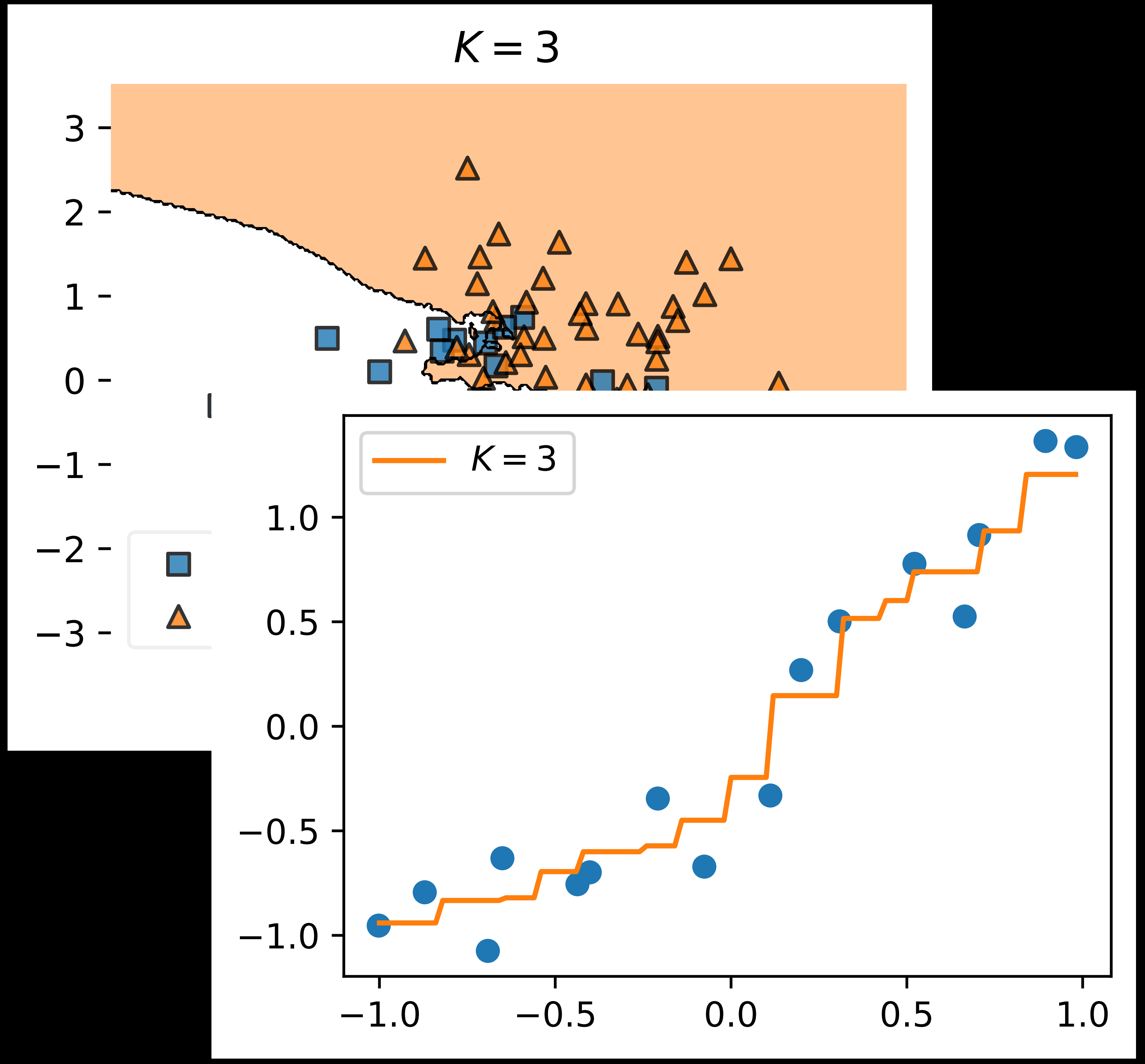


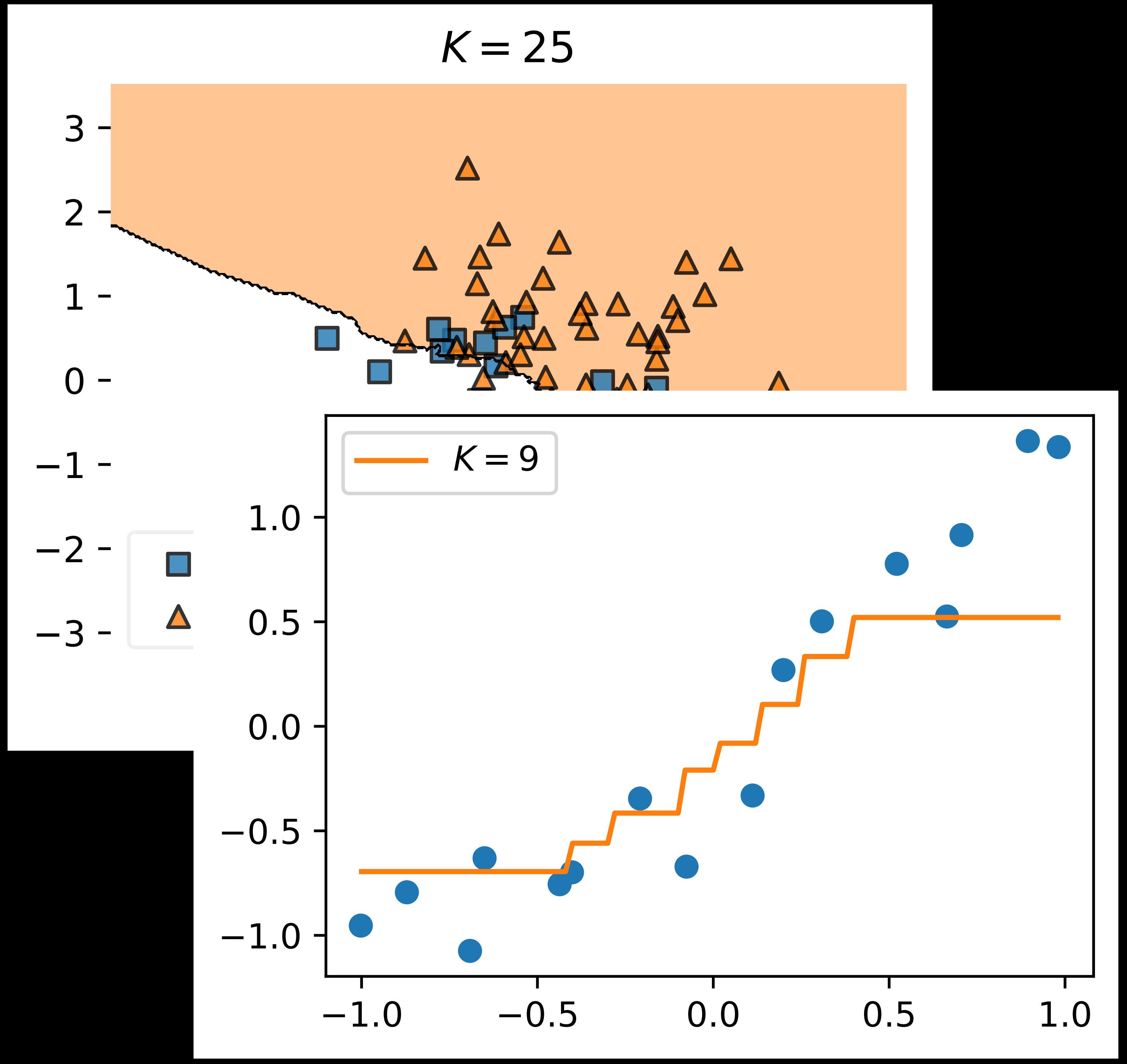


KNN Classifier and Regression

- Decision boundaries become smoother with larger k values
  - In both classification and regression







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

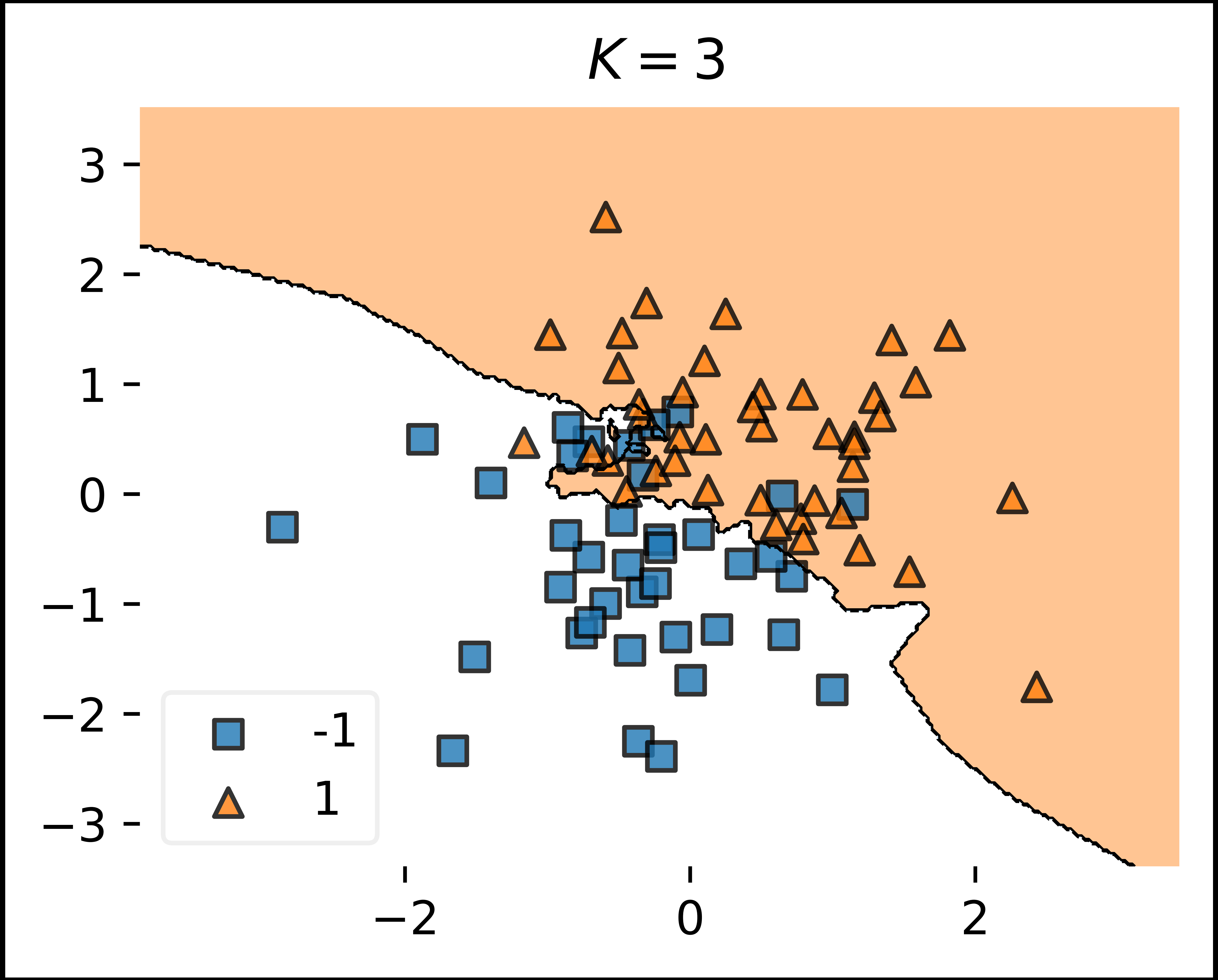
Differentiate binary, multi-class, and multi-label classification

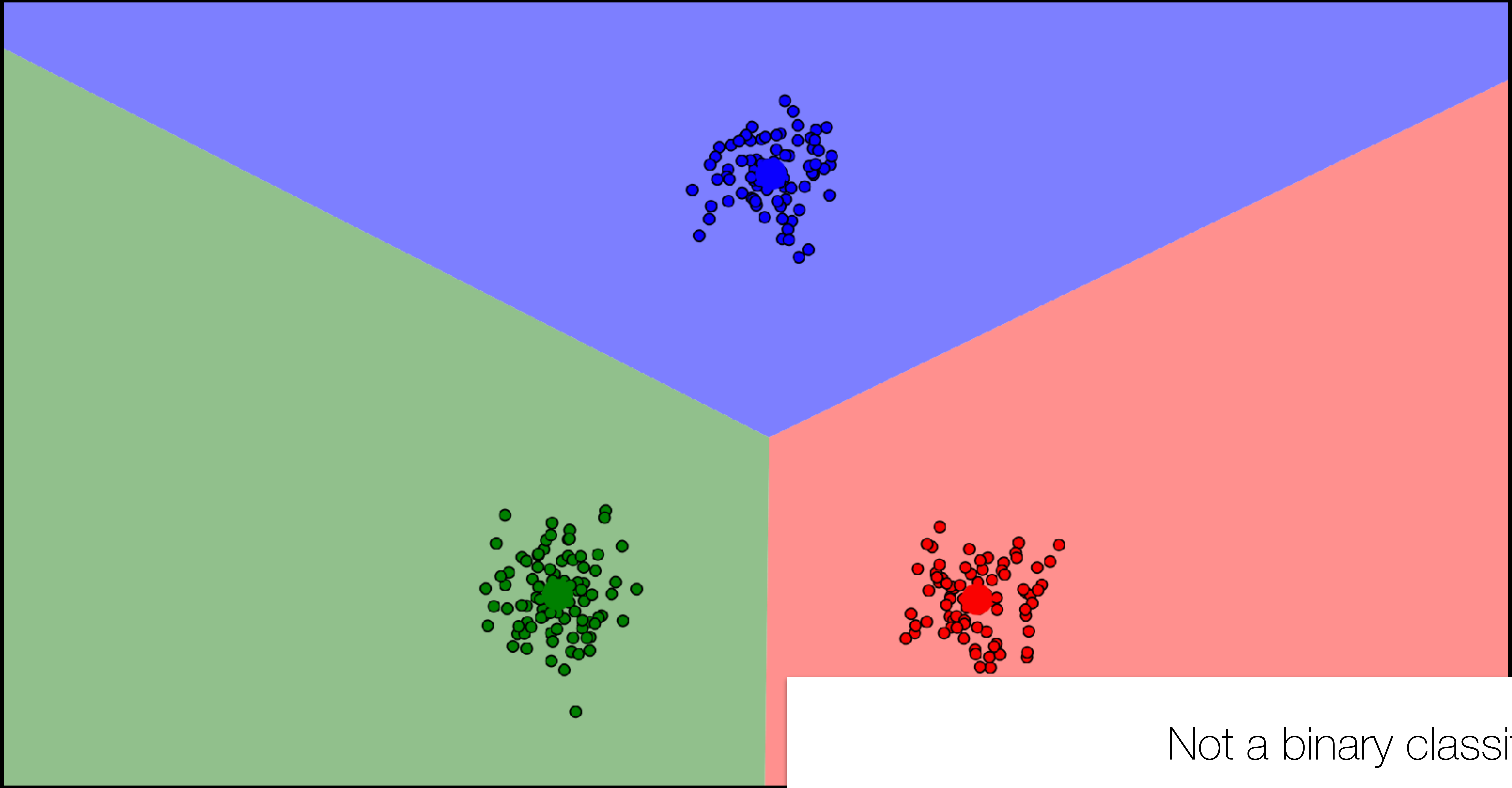
Define overfitting and describe its impact on generalizability

Classifying into two labels (yes/no, class 1/2, etc) is super common

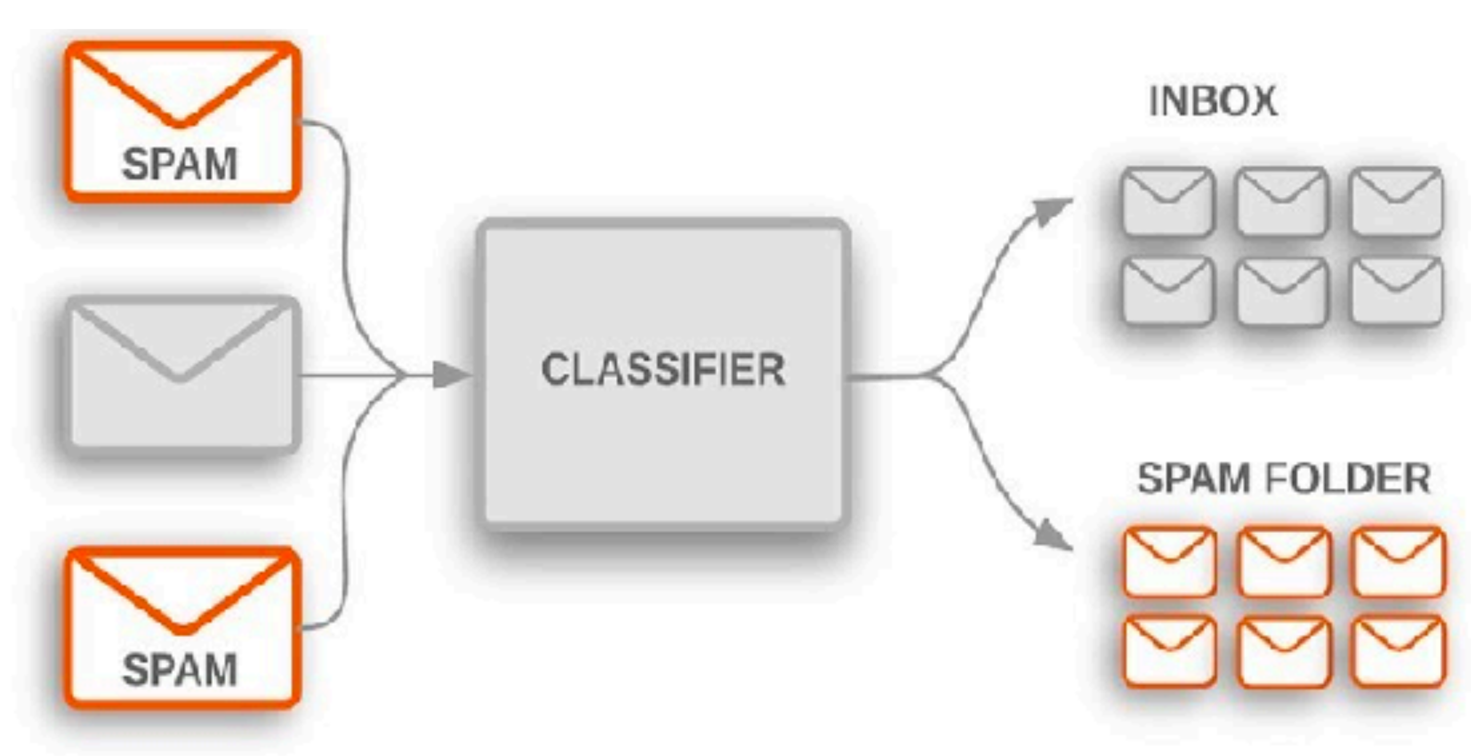
But not the only task

Binary Classification

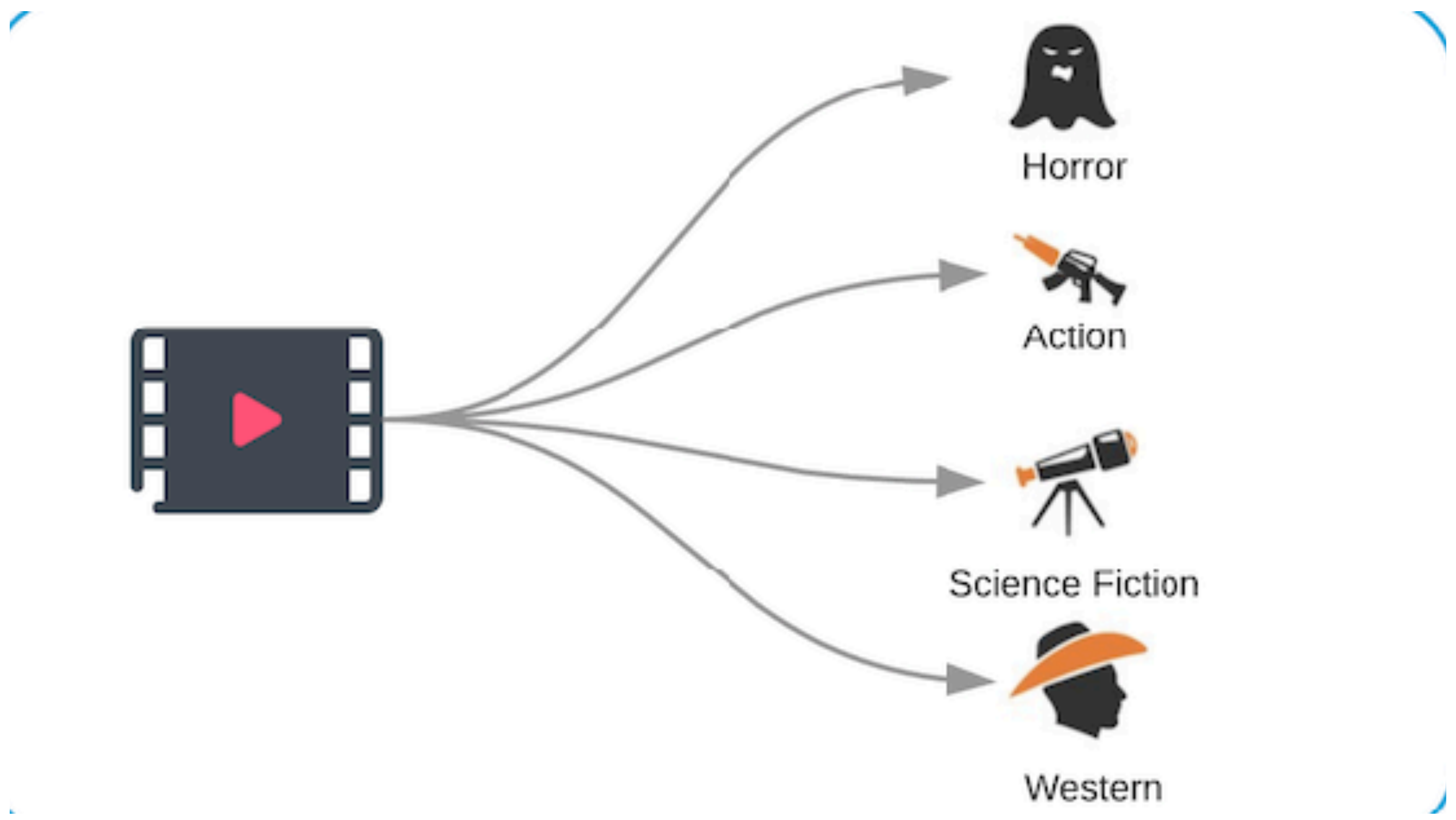




Not a binary classification



Binary Classification
e.g., Spam vs. Not-Spam



Multi-Class Classification
e.g., Movie Genre

Accuracy in Multi-Class Classification

Can still calculate accuracy here

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

|X| = # of samples

Was the prediction the same as the actual label?

But can't a movie be in more than one genre?

Can assume only one genre for standard multi-class classification

BUT! We can handle multiple labels

Multi-Label Classification: Samples can have multiple labels associated with them

“Cabin in the Woods”:
Horror, Comedy

“Thankskilling”:
Horror

“Happy Death Day”:
Horror, Comedy

“Evil Dead”:
Horror

This Module’s Learning Objectives

Part 1

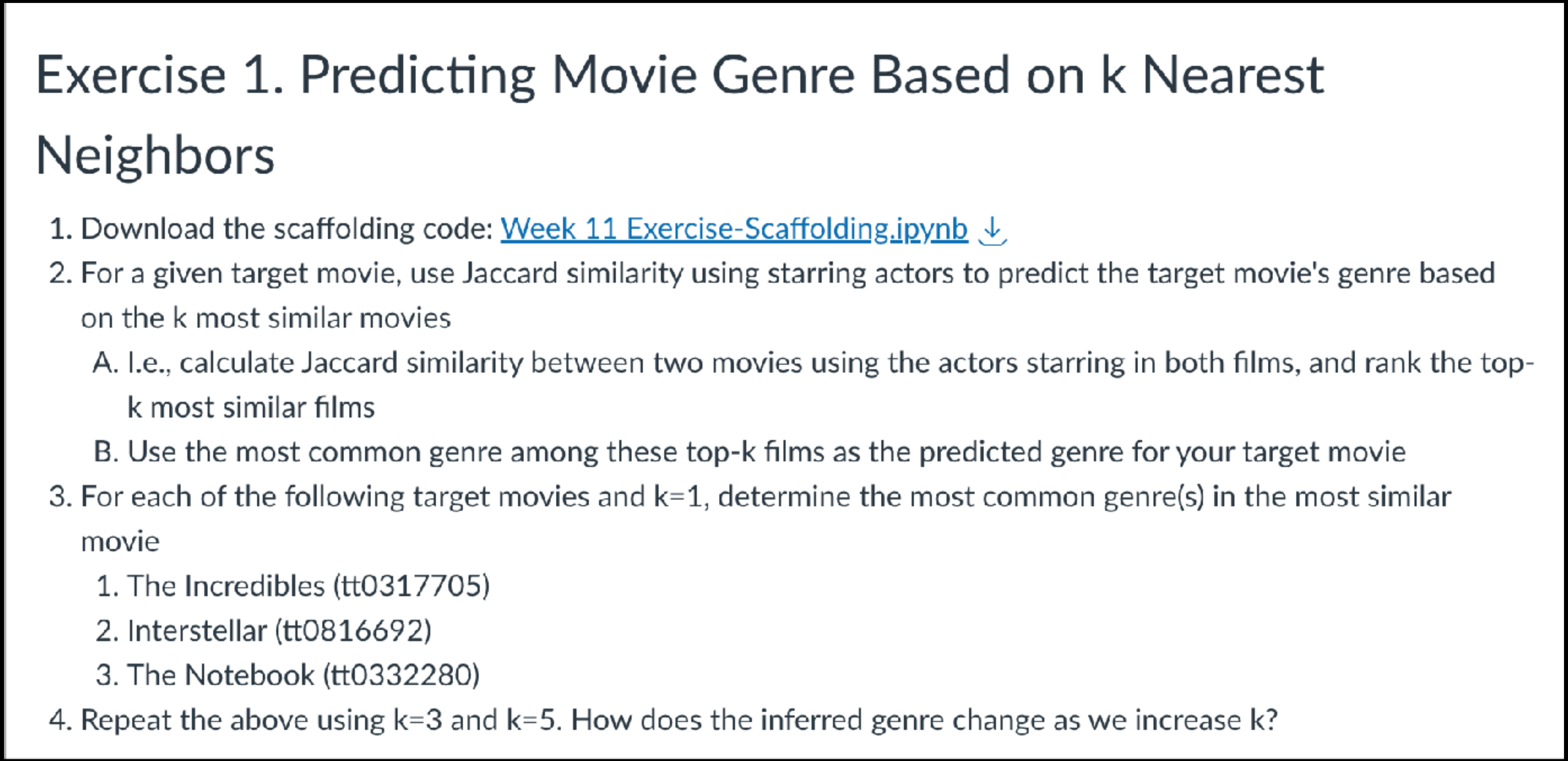
- Differentiate between classification and regression in supervised learning
- Describe how voting is used in for k-nearest neighbors classification
- Differentiate binary, multi-class, and multi-label classification
- Define overfitting and describe its impact on generalizability

Today's Exercises

Exercise 1: Predict Genre Based on k Nearest Neighbors

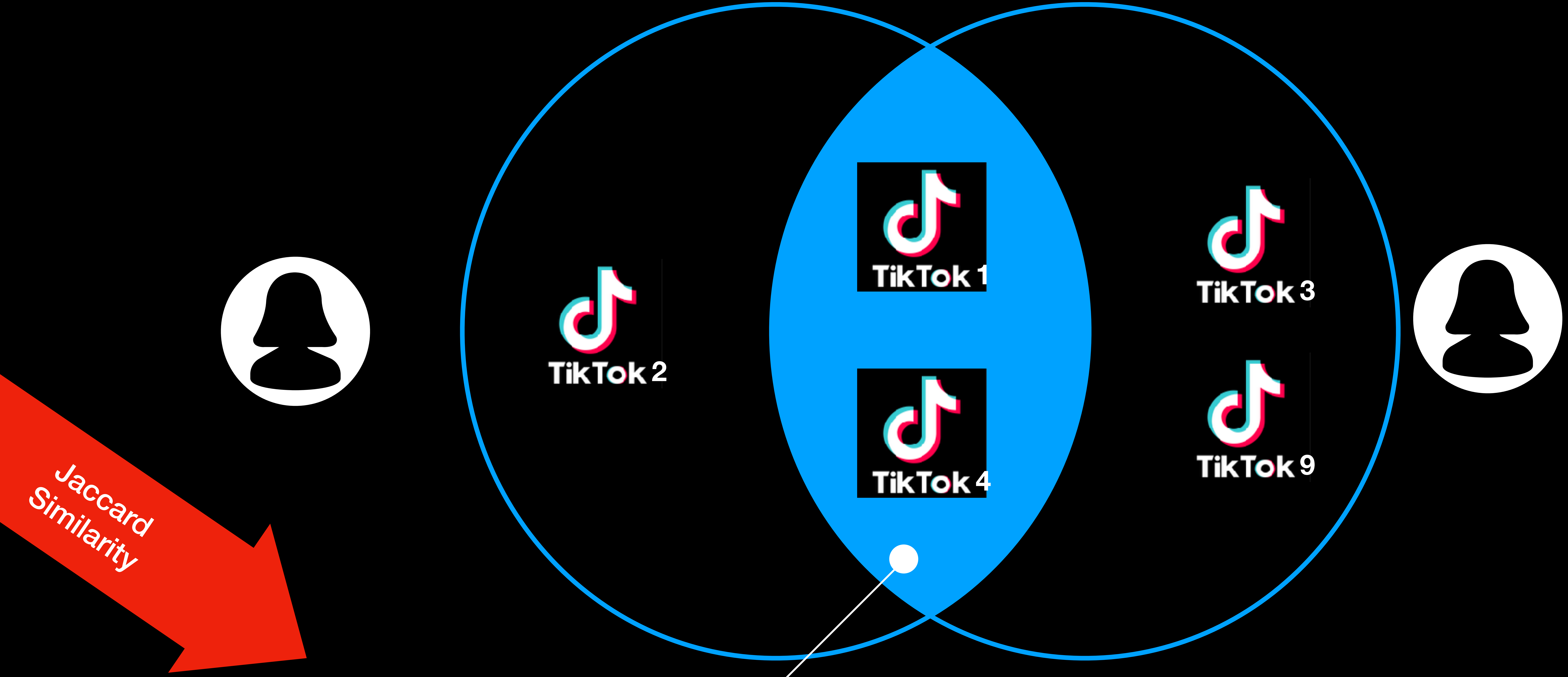
(Extra) Exercise 2: Predict Rating Based on k Nearest Neighbors

Exercise 1: Predict Genre



Exercise 1. Predicting Movie Genre Based on k Nearest Neighbors

1. Download the scaffolding code: Week 11 Exercise-Scaffolding.ipynb ↓
2. For a given target movie, use Jaccard similarity using starring actors to predict the target movie's genre based on the k most similar movies
   A. I.e., calculate Jaccard similarity between two movies using the actors starring in both films, and rank the top-k most similar films
   B. Use the most common genre among these top-k films as the predicted genre for your target movie
3. For each of the following target movies and k=1, determine the most common genre(s) in the most similar movie
   1. The Incredibles (tt0317705)
   2. Interstellar (tt0816692)
   3. The Notebook (tt0332280)
4. Repeat the above using k=3 and k=5. How does the inferred genre change as we increase k?

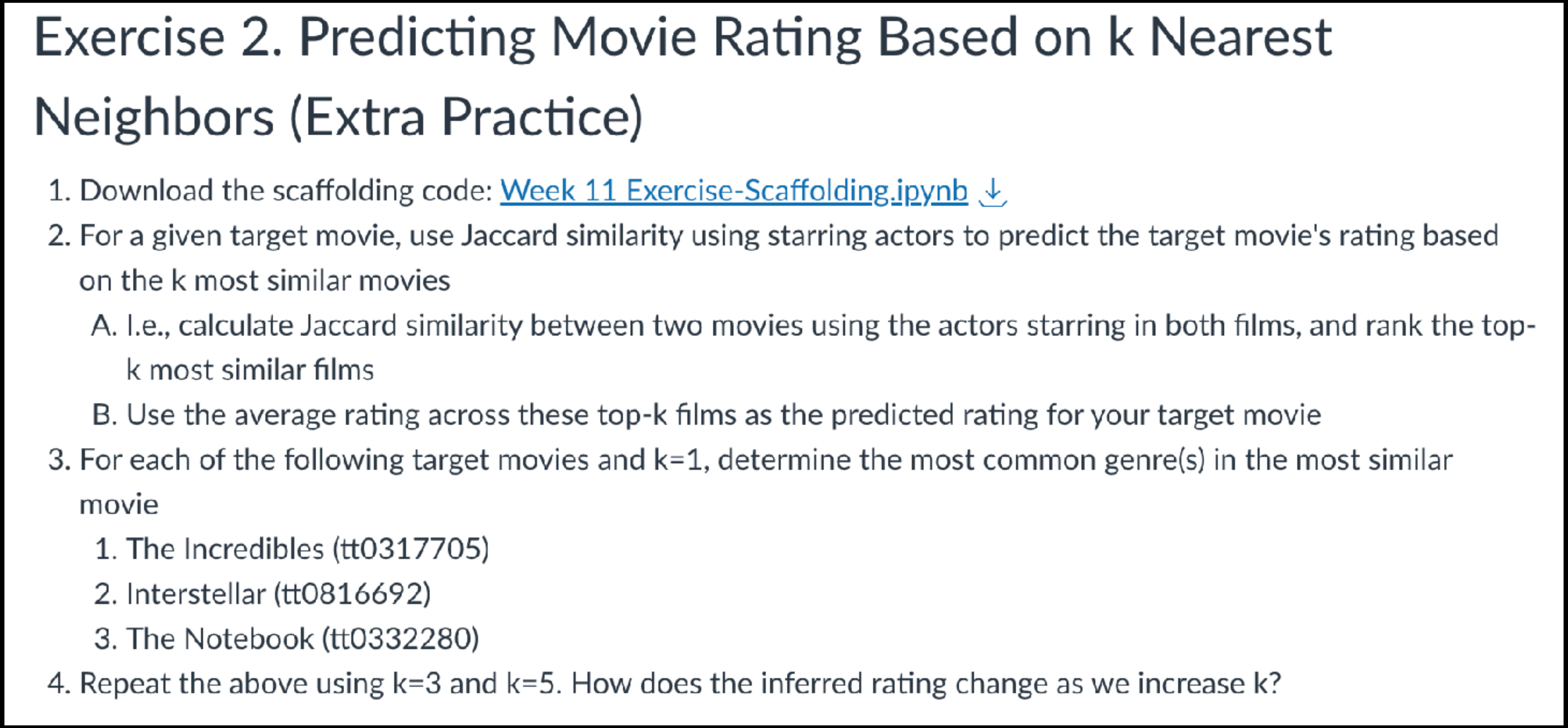


Overlap = 2

All Watched Videos = 5

Similarity = 2/5 = 0.4

(Extra) Exercise 2: Predict Rating



Exercise 2. Predicting Movie Rating Based on k Nearest Neighbors (Extra Practice)

1. Download the scaffolding code: Week 11 Exercise-Scaffolding.ipynb ↓
2. For a given target movie, use Jaccard similarity using starring actors to predict the target movie's rating based on the k most similar movies
   A. I.e., calculate Jaccard similarity between two movies using the actors starring in both films, and rank the top-k most similar films
   B. Use the average rating across these top-k films as the predicted rating for your target movie
3. For each of the following target movies and k=1, determine the most common genre(s) in the most similar movie
   1. The Incredibles (tt0317705)
   2. Interstellar (tt0816692)
   3. The Notebook (tt0332280)
4. Repeat the above using k=3 and k=5. How does the inferred rating change as we increase k?

What are your questions?

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