Python Data Products
Course 2: Design thinking and predictive pipelines

Lecture: Classification: Nearest Neighbor
Learning objectives

In this lecture we will...
• Introduce a simple classification algorithm, before we proceed to more complex alternatives in later lectures
• Demonstrate a “non-learning” solution to classification problems
Suppose we have some data we wish to classify, belonging to one of two classes (positive or negative)
What is the simplest algorithm we could come up with to classify a new data point?

positive examples

new data point – positive or negative?

negative examples
The **nearest neighbor** classification algorithm assigns the point the **label of the nearest point**.

- **positive examples**
- **new data point** – positive or negative?
- **nearest labeled point**
- **negative examples**

→ **Label = negative**
Precisely speaking, if we have a collection of points $\mathbf{X}$ (really a collection of feature vectors) and labels $\mathbf{y}$, and we see a new point (that we wish to label) $\mathbf{z}$, then:

$$\text{label}(\mathbf{z}) = y_{\arg \min_i ||\mathbf{z} - \mathbf{X}_i||}$$
Summary of concepts

- Introduced **nearest neighbor** classification
- Introduced the notation used do describe classifiers for the rest of this course