Week 6 - Introduction to Classification

Introduction:

Welcome to this lecture on basic classification models used in statistical modelling, with a particular emphasis on the k-Nearest Neighbors (k-NN) algorithm. Classification is a fundamental task in statistical modelling, where the goal is to assign input data points to predefined categories or classes.

Classification algorithm is a Supervised Learning technique used to categorize new observations. In classification, a program uses the dataset or observations provided to learn how to categorize new observations into various classes or groups. For instance, 0 or 1, red or blue, yes or no, spam or not spam, etc. Targets, labels, or categories can all be used to describe classes. The Classification algorithm uses labeled input data because it is a supervised learning technique and comprises input and output information. A discrete output function (y) is transferred to an input variable in the classification process (x).

In simple words, classification is a type of pattern recognition in which classification algorithms are performed on training data to discover the same pattern in new data sets.

> 💡 Exercise
>
> a) When scrolling through your e-mail inbox, what factors do you take into account when you decide if an e-mail is spam or not?
>
> b) When a bank looks at a transaction, what factors do you think it takes into account when they decide if a transaction is fraudulent or not?

Applications of Classification models:

Classification models are a cornerstone of machine learning, used across a wide range of industries and applications to categorize data into predefined classes or categories. These models can handle both binary classification (two classes) and multi-class classification (more than two classes) problems. Here are some of the key applications of classification models:

Spam Detection: In email filtering systems, classification models are used to distinguish between spam and non-spam emails, helping to protect users from unwanted email content.

Fraud Detection: Financial institutions use classification models to identify potentially fraudulent transactions. These models analyze transaction patterns to flag unusual activities that could indicate fraud.

Customer Segmentation: Companies apply classification models to customer data to segment customers into different groups based on their behaviors, preferences, or demographic characteristics. This segmentation supports personalized marketing strategies.

Image Recognition: In computer vision, classification models identify and categorize objects within images. This technology is used in applications ranging from security surveillance to medical imaging.

Sentiment Analysis: Classification models analyze text data from sources like social media, reviews, and surveys to determine the sentiment expressed (positive, negative, neutral) towards products, services, or topics.

Disease Diagnosis: In healthcare, classification models assist in diagnosing diseases by analyzing medical images, patient data, and test results. These models can help in early detection and treatment planning.

Credit Scoring: Banks and lending institutions use classification models to assess the creditworthiness of borrowers. These models analyze financial histories and personal information to classify individuals into different risk categories.

Natural Language Processing (NLP): Classification models are used in NLP for tasks like language detection, topic classification, and identifying the parts of speech in text data.

Recommendation Systems: E-commerce and streaming platforms use classification to categorize users and items, helping to drive personalized recommendations based on user preferences and behavior.

Predictive Maintenance: In manufacturing and industrial settings, classification models predict equipment failures by classifying the state of machinery based on sensor data and historical records. This helps in scheduling maintenance to prevent unexpected downtimes.

Voice Recognition: Classification models are crucial in recognizing and categorizing spoken words into text, enabling voice-controlled applications and devices.

Environmental Monitoring: Classification models are used in monitoring and managing natural resources, such as classifying land use in satellite images or detecting pollutants in water sources.

For the purposes of this module, we will be focused on k-nearest neighbours.

k-Nearest Neighbours (k-NN):

k-NN is a simple yet effective classification algorithm that works based on the premise that the closer two data points are together, the more likely they are of the same class.

Lets consider a simple example. Consider we have a ball bearing factory, and we require all ball bearings to be below a threshold. If they meet this condition, we accept it. If they do not, then we don't accept it.

Consider the following diagram. The blue dots are rejected ball bearings and the orange dots are accepted.
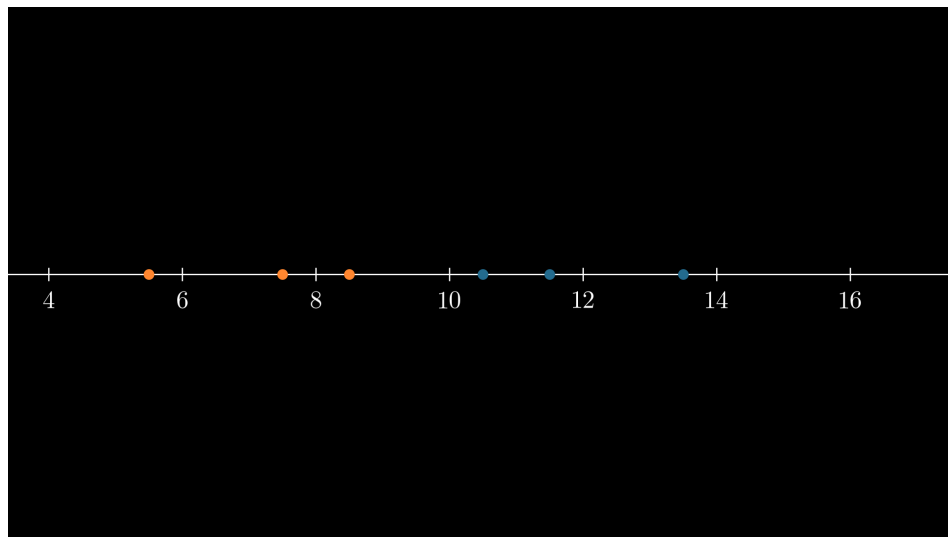

Figure 1: check

Is there a rule we can apply here?

What happens then, we have a new ballbearing (indicated by the yellow dot) that hasn't been classified? what would you classify this as using this rule?
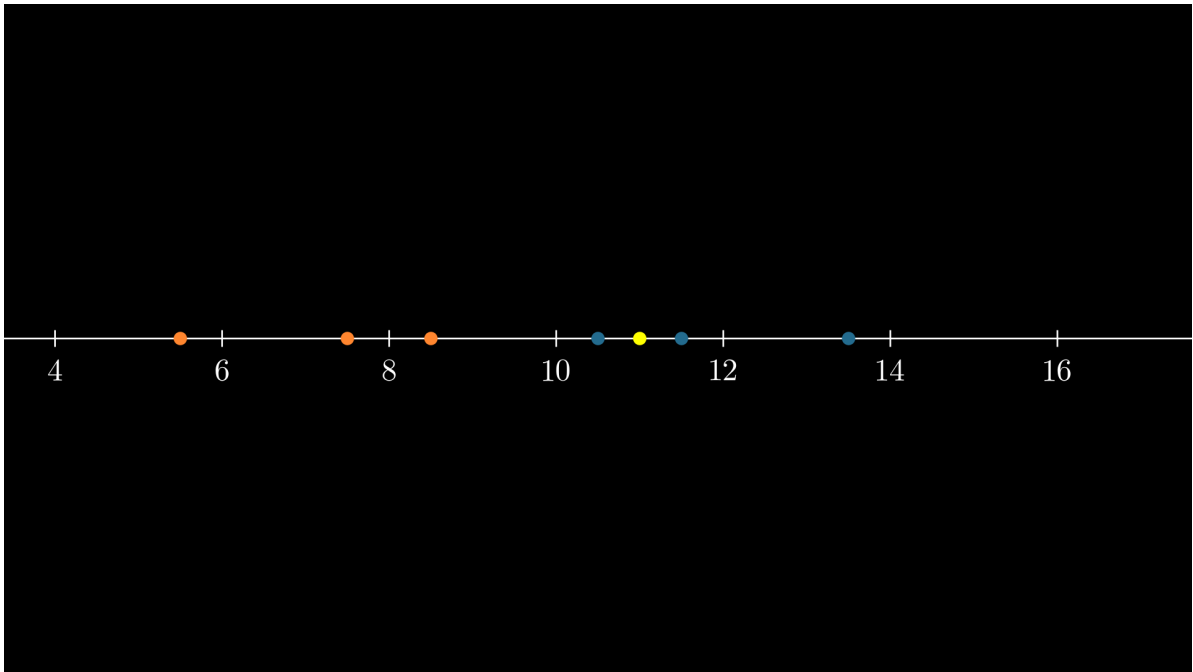
Figure 2: check

What happens if they are a bit more mixed up? Is there a rule?
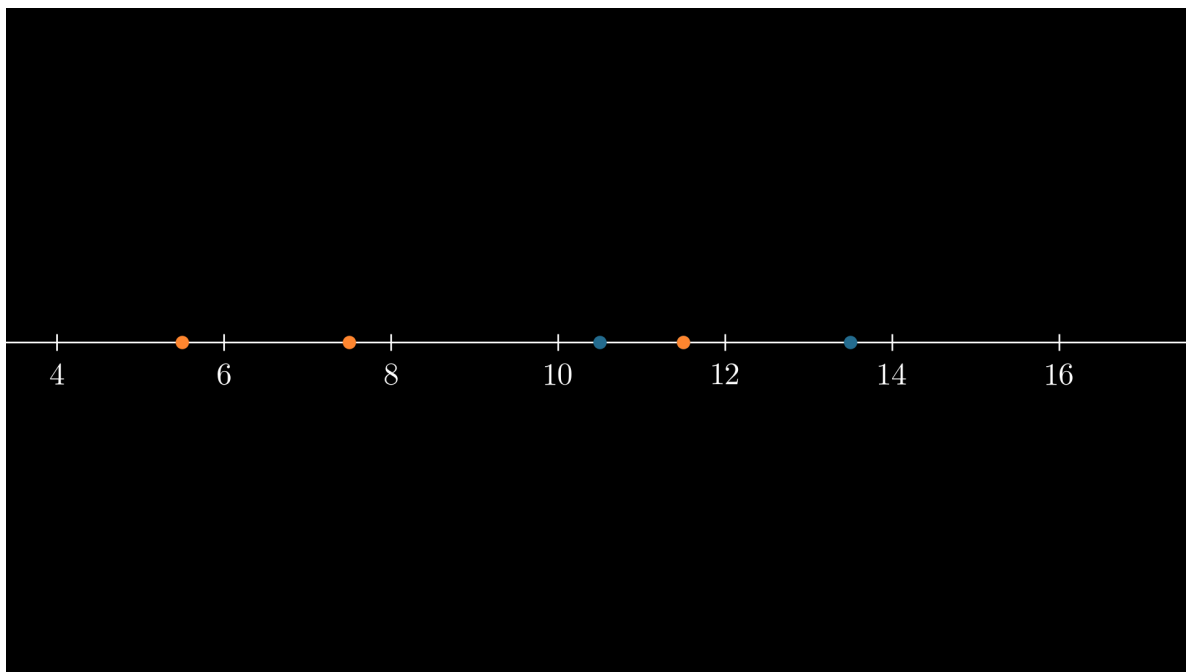
Figure 3: check

Now, we add a new ballbearing for you to identify. What would you class this as?
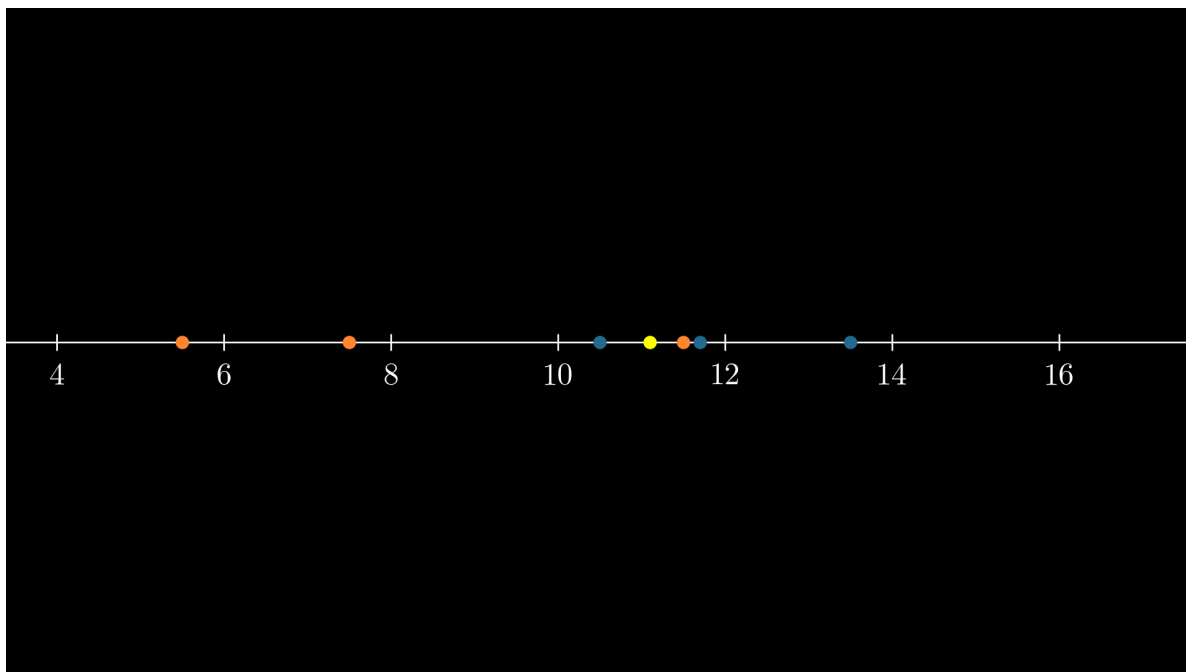
Figure 4: check

The answer is not clear cut. Luckily we can use k-Nearest Neighbours to figure this out.

The k-NN Algorithm:

Here is the k-nearest neighbours process, or algorithm

1. Choose the value of k, the number of nearest neighbours to consider (**an odd number is often chosen to avoid ties**).
2. Compute the distance between the input data point and all other data points in the dataset.
3. Select the k-nearest neighbours based on the computed distances.
4. Count the class labels of the k-nearest neighbours and assign the majority class label to the input data point.

Back to our example

If we compared our new ballbearing (yellow) to the closest neighbour (k = 1). We would classify this as accepted as the closest dot to the yellow dot in figure 5 is orange. Lets say that we want to use three neighbours (k = 3), the three dots closest to the yellow dot are orange,

blue and blue. Blue wins (meaning we class the yellow dot as accepted) because there are more blue dots (2) than orange dots (1) to the yellow dot.
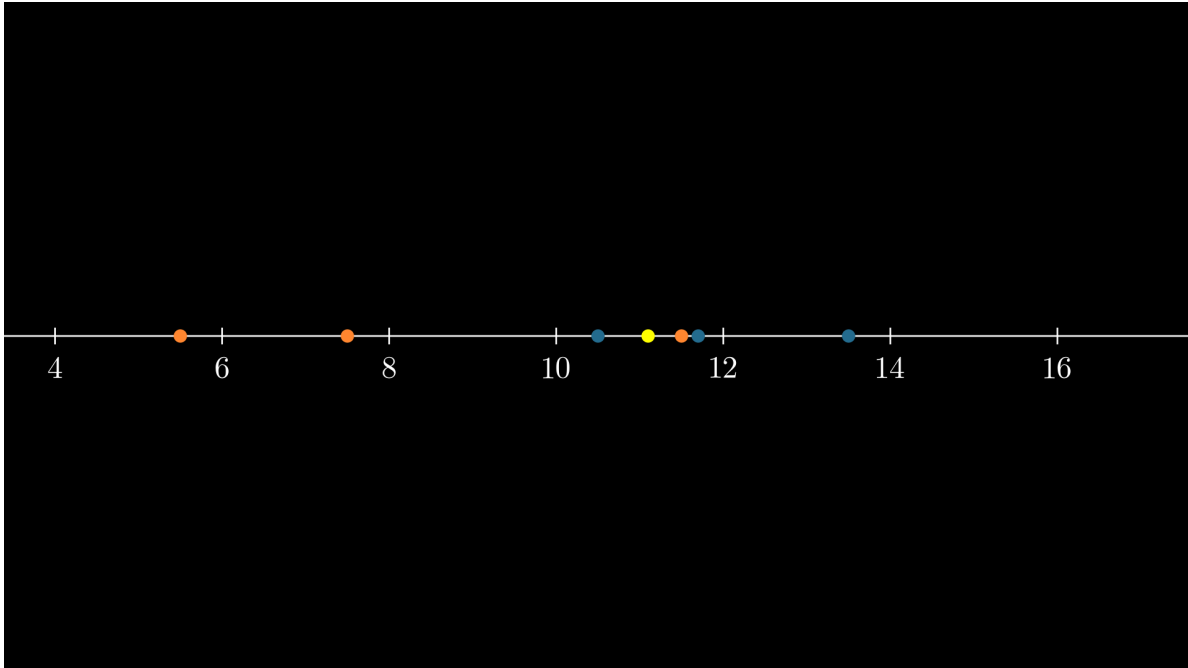
Figure 5: check

The intuition behind the k-nearest-neighbours model.

Lets think about the k-NN model intuitively. What we are basically saying is, if two objects share similar characteristics, they may be the same class of object. Now, of course that may also not be true for many reasons and we can be critical of this as a model later.

Worked Example

Plot these ballbearings on an x-axis (as above). Hint, use X to represent accepted and O to represent rejected.

Give worked example

Change the context of the example

Ballbearing Size	Classification
4.3	Rejected
2.2	Accepted
3.5	Rejected
1.5	Accepted
3.3	Rejected
2.6	Rejected
2.5	Rejected
3.2	Accepted
2.9	Accepted

Now, think about the following ballbearings with no classification.

Using K-NN of 1, 3 and 5, predict the classification of the following ballbearings

1. 2.8
2. 3.9
3. 0.7

Answers

Measure the distance between each ballbearing and the other ballbearings.

Ballbearing Size	Classification	Abs. Distance to 1.	Abs. Distance to 2.	Abs. Distance to 3.
4.3	Rejected	1.5	0.4	3.6
2.2	Accepted	0.6	1.7	1.5
3.5	Rejected	0.7	0.4	2.8
1.5	Accepted	1.3	2.4	0.8
3.3	Rejected	0.5	0.6	2.6
2.6	Rejected	0.2	1.3	1.9
2.5	Rejected	0.3	1.4	1.8
3.2	Accepted	0.4	0.7	2.5
2.9	Accepted	0.1	1.0	2.2

Order by Absolute Distance to 1. from least distance to most distance.

Ballbearing Size	Classification	Abs. Distance to 1.
2.9	Accepted	0.1
2.6	Rejected	0.2
2.5	Rejected	0.3
3.2	Accepted	0.4
3.3	Rejected	0.5
2.2	Accepted	0.6
3.5	Rejected	0.7
1.5	Accepted	1.3
4.3	Rejected	1.5

K-NN = 1, is the closest. It's classification is Accepted. K-NN = 3, is the closest 3. The classifications are 1 Accepted, 2 Rejected. Therefore it's classification is Rejected. K-NN = 5, is the closest 5. The classifications are 2 Accepted, 3 Rejected. Therefore it's classification is Rejected.

Order by Absolute Distance to 2. from least distance to most distance.

Ballbearing Size	Classification	Abs. Distance to 2.
4.3	Rejected	0.4
3.5	Rejected	0.4
3.3	Rejected	0.6
3.2	Accepted	0.7
2.9	Accepted	1.0
2.6	Rejected	1.3
2.5	Rejected	1.4
2.2	Accepted	1.7
1.5	Accepted	2.4

K-NN = 1, is the closest. It's classification is Rejected. K-NN = 3, is the closest 3. The classifications are 3 Rejected. Therefore it's classification is Rejected. K-NN = 5, is the closest 5. The classificaitons are 3 Rejected, 2 Accepted. Therefore it's classification is Rejected.

Order by Absolute Distance to 3. from least distance to most distance.

Ballbearing Size	Classification	Abs. Distance to 3.
1.5	Accepted	0.8

Ballbearing Size	Classification	Abs. Distance to 3.
2.2	Accepted	1.5
2.5	Rejected	1.8
2.6	Rejected	1.9
2.9	Accepted	2.2
3.2	Accepted	2.5
3.3	Rejected	2.6
3.5	Rejected	2.8
4.3	Rejected	3.6

What do you notice? Are there any ballbearings where the classification changes when you use different values of k?

Parameters of k-NN:

Two main parameters in k-NN are:

k: The number of nearest neighbors to consider (a hyperparameter that needs to be tuned).
Distance metric: The measure used to calculate the distance between data points.

Distance Metrics

When we have one value to use as a classification, the distance calculation is quite easy. We can just use the absolute difference between the two values as our distance. However when we have two values, there are several ways to measure distance. We are going to go through two of them.

Euclidean Distance:

Definition: Euclidean distance is the straight-line distance between two points in a Euclidean space. It is the length of the line segment connecting the two points.
Formula (for 2D space):

$$\sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

In higher-dimensional spaces, the formula extends accordingly. Example: In a 2D plane, the Euclidean distance between points (3, 4) and (6, 8) is calculated as

$$\sqrt{(6 - 3)^2 + (8 - 4)^2} = \sqrt{9 + 16} = 5$$

Manhattan Distance:

Definition: Manhattan distance, also known as taxicab distance or city block distance, is the sum of the absolute differences of the coordinates of two points. It measures the distance a vehicle would travel along a grid-like street network to reach the destination.

Formula (for 2D space):

$$|x2 - x1| + |y2 - y1|$$

In higher-dimensional spaces, the formula extends accordingly.

Example: In a 2D plane, the Manhattan distance between points (3, 4) and (6, 8) is calculated as

$$|6 - 3| + |8 - 4| = 3 + 4 = 7$$

2d Example.

Plot these ballbearings on an x, y axis. Hint, use X to represent accepted and O to represent rejected.

Ballbearing Mass	Ballbearing Size	Classification
3.1	4.3	Rejected
0.9	2.2	Accepted
3.2	3.5	Rejected
1.2	1.5	Accepted
3.2	3.3	Rejected
4.6	2.6	Rejected
4.3	2.5	Rejected
2.0	3.2	Accepted
2.4	2.9	Accepted

Now, think about the following tumors with no classification.

Using K-NN of 1, 3 and 5, predict the classification of the following tumours

1. Ballbearing Mass 2.8, Ballbearing Size = 2.3
2. Ballbearing Mass 3.9, Ballbearing Size = 2.1
3. Ballbearing Mass 0.7, Ballbearing Size = 4.6

Measure the distance between each ballbearing and the first ballbearing.

Ballbearing Mass	Ballbearing Size	Classification	Mass Distance	Size Distance	Total Distance
3.1	4.3	Rejected	$(3.1 - 2.8)^2$	$(4.3 - 2.3)^2$	4.09
0.9	2.2	Accepted	$(0.9 - 2.8)^2$	$(2.2 - 2.3)^2$	3.62
3.2	3.5	Rejected	$(3.2 - 2.8)^2$	$(3.5 - 2.3)^2$	1.6
1.2	1.5	Accepted	$(1.2 - 2.8)^2$	$(1.5 - 2.3)^2$	3.2
3.2	3.3	Rejected	$(3.2 - 2.8)^2$	$(3.3 - 2.3)^2$	1.16
4.6	2.6	Rejected	$(4.6 - 2.8)^2$	$(2.6 - 2.3)^2$	3.33
4.3	2.5	Rejected	$(4.3 - 2.8)^2$	$(2.5 - 2.3)^2$	2.29
2.0	3.2	Accepted	$(2.0 - 2.8)^2$	$(3.2 - 2.3)^2$	1.45
2.4	2.9	Accepted	$(2.4 - 2.8)^2$	$(2.9 - 2.3)^2$	0.52

Order by Absolute Distance to 1. from least distance to most distance.

Ballbearing Mass	Ballbearing Size	Classification	Total Distance
2.4	2.9	Accepted	0.52
3.2	3.3	Rejected	1.16
2.0	3.2	Accepted	1.45
3.2	3.5	Rejected	1.6
4.3	2.5	Rejected	2.29
1.2	1.5	Accepted	3.2
4.6	2.6	Rejected	3.33
0.9	2.2	Accepted	3.62
3.1	4.3	Rejected	4.09

K-NN = 1, is the closest. It's classification is Accepted K-NN = 3, is the closest 3. The classifications are 2 Accepted, 1 Rejected. Therefore it's classification is Accepted K-NN = 5, is the closest 5. The classificaitons are 3 Rejected, 2 Accepted. Therefore it's classification is Rejected.

Measure the distance between each ballbearing and the second ballbearing.

Ballbearing Mass	Ballbearing Size	Classification	Mass Distance	Size Distance	Total Distance
3.1	4.3	Rejected	$(3.1 - 3.9)^2$	$(4.3 - 2.1)^2$	5.48
0.9	2.2	Accepted	$(0.9 - 3.9)^2$	$(2.2 - 2.1)^2$	9.01
3.2	3.5	Rejected	$(3.2 - 3.9)^2$	$(3.5 - 2.1)^2$	2.45
1.2	1.5	Accepted	$(1.2 - 3.9)^2$	$(1.5 - 2.1)^2$	7.65
3.2	3.3	Rejected	$(3.2 - 3.9)^2$	$(3.3 - 2.1)^2$	1.93
4.6	2.6	Rejected	$(4.6 - 3.9)^2$	$(2.6 - 2.1)^2$	0.74
4.3	2.5	Rejected	$(4.3 - 3.9)^2$	$(2.5 - 2.1)^2$	0.32
2.0	3.2	Accepted	$(2.0 - 3.9)^2$	$(3.2 - 2.1)^2$	4.82
2.4	2.9	Accepted	$(2.4 - 3.9)^2$	$(2.9 - 2.1)^2$	2.89

Order by Absolute Distance to 2. from least distance to most distance.

Ballbearing Mass	Ballbearing Size	Classification	Mass Distance	Size Distance	Total Distance
4.3	2.5	Rejected	0.32		
4.6	2.6	Rejected	0.74		
3.2	3.3	Rejected	1.93		
3.2	3.5	Rejected	2.45		
2.4	2.9	Accepted	2.89		

Ballbearing Mass	Ballbearing Size	Classification	Mass Distance	Size Distance	Total Distance
2.0	3.2	Accepted	4.82		
3.1	4.3	Rejected	5.48		
1.2	1.5	Accepted	7.65		
0.9	2.2	Accepted	9.01		

K-NN = 1, is the closest. It's classification is Rejected K-NN = 3, is the closest 3. The classifications are 3 Rejected. Therefore it's classification is Rejected K-NN = 5, is the closest 5. The classificaitons are 4 Rejected, 1 Accepted. Therefore it's classification is Rejected.

Measure the distance between each ballbearing and the third ballbearing.

Ballbearing Mass	Ballbearing Size	Classification	Mass Distance	Size Distance	Total Distance
3.1	4.3	Rejected	$(3.1 - 0.7)^2$	$(4.3 - 4.6)^2$	5.85
0.9	2.2	Accepted	$(0.9 - 0.7)^2$	$(2.2 - 4.6)^2$	5.8
3.2	3.5	Rejected	$(3.2 - 0.7)^2$	$(3.5 - 4.6)^2$	7.46
1.2	1.5	Accepted	$(1.2 - 0.7)^2$	$(1.5 - 4.6)^2$	9.86
3.2	3.3	Rejected	$(3.2 - 0.7)^2$	$(3.3 - 4.6)^2$	7.94
4.6	2.6	Rejected	$(4.6 - 0.7)^2$	$(2.6 - 4.6)^2$	19.21
4.3	2.5	Rejected	$(4.3 - 0.7)^2$	$(2.5 - 4.6)^2$	17.37
2.0	3.2	Accepted	$(2.0 - 0.7)^2$	$(3.2 - 4.6)^2$	3.65
2.4	2.9	Accepted	$(2.4 - 0.7)^2$	$(2.9 - 4.6)^2$	5.78

Order by Absolute Distance to 1. from least distance to most distance.

Ballbearing Mass	Ballbearing Size	Classification	Mass Distance	Size Distance	Total Distance
2.0	3.2	Accepted	3.65		
2.4	2.9	Accepted	5.78		
0.9	2.2	Accepted	5.8		
3.1	4.3	Rejected	5.85		
3.2	3.5	Rejected	7.46		
3.2	3.3	Rejected	7.94		
1.2	1.5	Accepted	9.86		
4.3	2.5	Rejected	17.37		
4.6	2.6	Rejected	19.21		

K-NN = 1, is the closest. It's classification is Accepted K-NN = 3, is the closest 3. The classifications are 3 Accepted. Therefore it's classification is Accepted K-NN = 5, is the closest 5. The classificaitons are 3 Accepted, 2 Rejected. Therefore it's classification is Rejected.

What do you notice? Are there any ballbearings where the classification changes when you use different values of k?

Applications of k-NN:

k-NN is used in various applications, including: - Image classification.

- Recommender systems.
- Anomaly detection.
- Medical diagnosis.
- Text categorization.

Conclusion:

Classification is a fundamental task in machine learning, and there are several basic models to choose from. k-NN is an intuitive and straightforward classification algorithm that relies on proximity in the feature space. Understanding the strengths, weaknesses, and parameters of k-NN is essential for effective model selection and application.