# Linear Discriminant Analysis

We want to reduce the dimensions of the data that produces some separability in the data.

LDA uses the data from the dataset to create a new axis that it projects the data onto the new axis to maximize the separation.

Chart, scatter chart

Description automatically generated

Tries to maximize the distance between means of the different groups. Also tries to minimize the scatter within a group

Diagram, box and whisker chart

Description automatically generated

We want to maximize this ratio, this means a large difference between means but a very small scatter between points.

# PCA

# TSNE

This reduces the dimensionality of data to something that is easier to understand, the goal is to preserve individual clustering of data. Projects data in a low dimensional space so clustering in the higher dimensional space is preserved. TSNE has a strong attraction to similar cluster pints, and a push from points it should not be close to in the 2D scatter plot.

Step 1 is determining the similarity of all points in the scatter plot; this can be distance between two points. We then plot the distance on a normal curve that is centered on the point we are comparing all other point distances to. The distance from the other point, or the point in comparison, to the normal curve is measured. This distance is the unscaled similarity.

Step 2 is to scale the values, this is because the density of points in a cluster impact the width of the normal curve. Scaling the values means that they will be equal no matter the width of the normal curve.

Chart

Description automatically generated

Scaled score = score / sum of all scores

This means no matter the width of the curve we will get the same similarity scores. This means that the point of interest has less impact on the creation of the similarity values.

Step 3 is to repeat steps 1 and 2 for all points. We then take the average of the similarity score from A to B and B to A ( since every node was considered ) and this tells us the average similarity score. These values wont be exactly the same due to the fact that the width of the distribution is related to the density of surrounding points.

TSNE defines similarity to itself as 0 in the similarity matrix we generate for each point to another.

We can randomly project all our points into the lower dimension that we want. We then use a t-distribution to calculate scaled similarity scores like we did before. T-SNE then iterates over points, moving them in the direction that makes their scaled similarity values more similar to those calculated in the higher dimension

Chart, bubble chart

Description automatically generated

The t-distribution is used to ensure the clusters don’t tend towards the middle as much, this allows for better seperability.