Topic 5 - Probabilistic classifiers

Introduction to unsupervised Learning

- Supervised learning constitutes a set of approaches that's essentially the most straightforward way to apply machine learning algorithms.
- Nonetheless, supervised learning cannot be applied in all problems settings.
- Besides the fact that this process is costly, it's also prone to errors since it relies upon human judgment.
- Allows us to perform analysis on a set of data without labels
- **Clustering** is an unsupervised method of arranging data into subgroups in order to capture some interpretable qualities in the data.
- The second category of unsupervised learning is **dimensionality reduction**.

Clustering

We want to map our input observations X to our output values, which we call here Y. These should satisfy some constraints.

$f(ins, goat, ins, goat, goat) \rightarrow [0, 1, 0, 1, 1]$

$X = Input\ observations \qquad T = output\ that\ satisfies\ constraints$

$X \rightarrow Y \quad by\ applying\ f_{\theta}(X)$

Now the constraints could be some measure of similarity where we map observations that are similar in X to outputs that are close together in Y.

Alternatively, the constraint could be finding ways of reducing the size of X, the dimensionality of X while maximizing the variance between the data points once we get to our mapping in Y. This is an example of **dimensionality reduction**.

K-mean

Algorithm:

1. Initialise k centroids (random)
2. Calculate distance from every data point to centroid
3. Assign data points to nearest centroid to from clusters
4. Re-assign centroids as the mean of each cluster’s data
5. Repeat 2-4 unit **convergence** (centroids stop changing)

It requires good centroids to have good clusters
- Scaling is important

Dimensionality reduction

- Particularly useful when we're dealing with images

Data in high dimensions (image)

- 640 * 480 pixels
- Grayscale pixel intensity (8 bytes float)
- Each pixel os a dimension
    - 300,000 features / dimension

Solution

- map to a lower dimensional space
- capture the data's essence using appropriate latent factors

Principal Component Analysis (PCA)

- Works along the basis of looking for the maximum variability in the data
- If the data is represented in three dimensions, the algorithm creates two dimensions where the data fits the most variance. We can continue until creating one dimension