Unsupervised Learning

Chapter 9 Machine Learning

- Used when the training data has no labels, and we instead are interested in finding out patterns, trends and associations in data

K-means Clustering

- Groups a dataset into k clusters
    1. Randomly distribute k centre points across feature space
    2. Assign each point in the dataset to one of the centre points based on distance
    3. Create new cluster centres based on the mean value of the points in each cluster
    4. Repeat until convergence
- Note, this algo has a couple of fallacies/ things to be wary of
    - Doesn't guarantee global convergence as the minima are heavily dependent on initial starting points of cluster centres
    - K clusters to group into has to be pre-decided as there is no information to go by in and of the data itself as to how to categorise things
    - Numbering of clusters is (obviously) arbitrary
    - Sensitive to outliers
- An enhancement is to sample from many iterations of the algorithm using different starting points (and choose the centres which lead to the minimal distance from each point to it's centre)

Principal Component Analysis

- This aims to decompose the feature vectors into a linear combination of it's orthogonal vectors. As such, we are able to define these principle components that characterise our dataset.
- More formally for a specific point x, where w*i denote the principle components of x, and t_i denote the weights associated with each component $$ \boldsymbol{x}= \sum*{i=1}^n t_i\boldsymbol{w_i} $$
- Now with the key principle that w_i, the principle components come from a set of mutually orthogonal vectors, we can pre multiply knowing that for i /= j, the dot products of these vectors are 0 to give

$$\boldsymbol{w_j}\boldsymbol{x} = t_j$$

- The ability to do this simplification is why we necessitate orthogonality
- To actually find the principle components themselves, we look to sequentially maximise the variance of the largest principle components (as such, they capture in decreasing

order, the variability of the dataset)

- For the largest principle component, w*1, considering all points x_i (unlike just a specific point as earlier)$$argmax(\sum*{i=1}^m t^2*{1i})* = *argmax(\sum*{i=1}^m\boldsymbol{x*iw*{1}})

  =argmax(||\boldsymbol{X}w_1||^2

  = argmax((\boldsymbol{X}w_1)^T(\boldsymbol{X}w_1))

  = argmax(w_1^T \boldsymbol{X}^TXw_1

  = argmax(w_1^T\boldsymbol{C}w_1)$$ where we have assumed the feature space X has already been normalised.
- Remember: Each point x_i in the dataset will share the same components (hence 1 subscript for w), but have different weights for these components (duh) (hence 2 subscripts for t) so t_1i denotes the weight of the first principle component for point x_i
- The principle components are then given by the eigenvectors of C