Week 9
Clustering and EM

From supervised to unsupervised

Supervised: given labels

Unsupervised: the labels are not given

Semi-supervised: some have labels, some not.

Clustering

- **Deterministic**: hard assignment to each cluster (K-means)

- **Probabilistic:** model assignment as a discrete latent variable (Mixtures of Gaussians, Dirichlet process)

KNN vs. K-means?

K-means: unsupervised / clustering
KNN: supervised / classification

Algorithm of K-means:
step 1: decide the number of clusters (k) and randomly choose k seeds.
step 2: compute distances between all data to the k means, and cluster the points that share the same closet mean
step 3: compute a new mean based on the new cluster
step 4: repeat step 1 - 3 until converge

Algorithm of KNN:
step 1: Compute distance to other training records (euclidean distance)
step 2: Identity k nearest neighbours
step 3: Use class labels of nearest neighbours to determine the class label of unknown record (majority vote, weight the vote...)

KNN vs. K-means?

notice:
if k is too small => sensitive to noise points
if k is too large => neighbourhood may include points from other classes.

KNN tries to classify an unlabelled observation based on it's k surrounding neighbours. It is also know as lay learner because it involves minimal training of model. Hence, it doesn't use training data make generalisation on unseen data set.
K-means tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.

Problem with KNN:

- scaling issues
- high dimensional data (problem with euclidean measure)
- produce counter-intuitive results (normalise the vectors to unit length)

EM

Expectation - Maximisation (EM algorithm)

Expectation Step: calculate parameters using the previous estimation
Maximisation Step: maximise (or minimise) the objective function using result from E-step

EM in K-means

E-step: assign each data point either to the red cluster or blue cluster, according to which cluster is nearer. (b)

M-step: each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster. (c)

(d) – (i): successive E and M steps through to final convergence