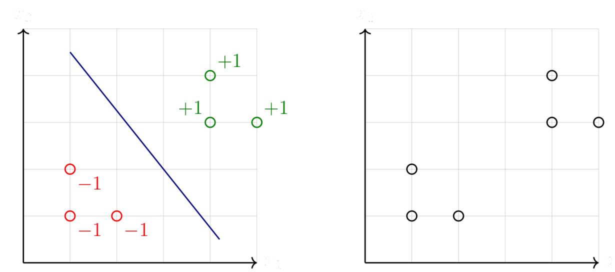
Clustering
K-means and GMM

Clustering



Supervised setting   Unsupervised setting

Automatically group data points into clusters

Lloyd's Algorithm

(a) Randomly initialize k cluster centers $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$.

(b) Reassign each point to the nearest center (for ℓ_2 / Euclidean distance):

$$C_j \;=\; \underset{i \in \{1,\ldots,k\}}{\arg\min} \left\| x_j - \mu_i \right\|_2 .$$

(c) Recompute each center as the mean of its assigned points:

$$\mu_i \;=\; \frac{\sum_{j=1}^{n} \mathbf{1}\{C_j = i\}\, x_j}{\sum_{j=1}^{n} \mathbf{1}\{C_j = i\}},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

Repeat steps (b) and (c) until assignments do not change; then k-means is said to have converged.

K-means Convergence

$$J(C, \mu) \;=\; \sum_{j=1}^{n} \left\| x_j - \mu_{C_j} \right\|_2^2$$

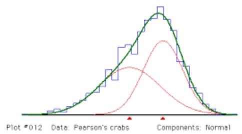
- Step (b) reduces above objective w.r.t choice of cluster assignments
- Step (a) reduces above objective w.r.t. choice of cluster centers the μ's
- Hence each step decreases objective (or at least doesnt increase)

How to choose K (no. of clusters)

- Elbow method:
  - plot Objective versus K, typically it monotonically decreases.
  - Pick point where there is a kink
  - Intuition: look at rate of change
- Add to objective penalty (+ pen(K)) and minimize, pen increases with K
  - intuition we prefer smaller number of clusters
  - Use prior knowledge to pick pen(K)
  - (AIC, BIC etc can been seen to be specific cases)

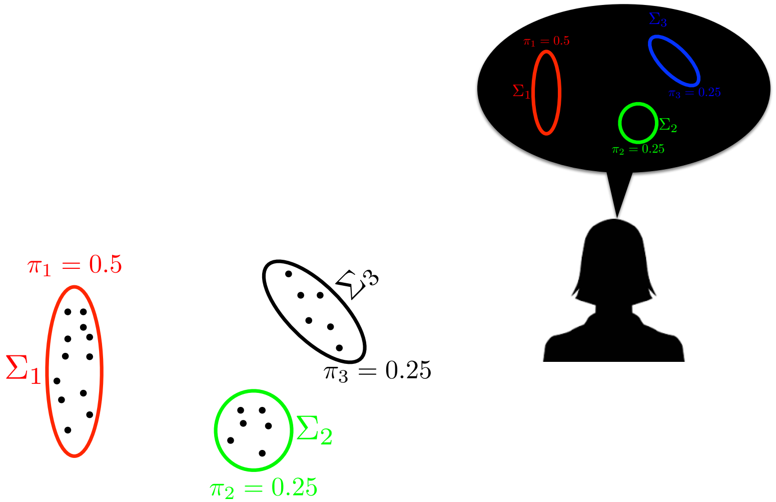
Weldon's Crab dataset

- 23 attributes, 1000 measurements
- All but one attribute were fit well by normal distribution
- One of them looked like…



Discovered that there were two species of crabs

Mixture of Gaussian



Algorithm

(a) Randomly initialize $\Theta = (\mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k, \pi)$.

(b) Compute posteriors (soft assignments) using Bayes' rule. For $i \in \{1, \ldots, K\}$,

$$P(Z_j = i \mid x_j; \Theta) = \frac{P(x_j \mid Z_j = i; \Theta)\, \pi_i}{\sum_{l=1}^{k} P(x_j \mid Z_j = l; \Theta)\, \pi_l}.$$

Using the Gaussian density:

$$P(x_j \mid Z_j = i; \Theta) = \frac{1}{\sqrt{2\pi |\Sigma_i|}} \exp\left(-\frac{1}{2}(x_j - \mu_i)^\top \Sigma_i^{-1} (x_j - \mu_i)\right).$$

(c) Update parameters using the soft assignments:

$$\pi_i = \frac{1}{n} \sum_{j=1}^{n} P(Z_j = i \mid x_j)$$

$$\mu_i = \frac{\sum_{j=1}^{n} P(Z_j = i \mid x_j)\, x_j}{\sum_{j=1}^{n} P(Z_j = i \mid x_j)},$$

$$\Sigma_i = \frac{\sum_{j=1}^{n} P(Z_j = 1 \mid x_j)\, (x_j - \mu_i)(x_j - \mu_i)^\top}{\sum_{j=1}^{n} P(Z_j = 1 \mid x_j)},$$

Repeat steps (b) and (c) until the cluster assignments (posteriors) no longer change appreciably. As with k-means, this resembles coordinate descent and can be susceptible to local optima.