ONLINE MASTERS IN **DATA SCIENCE**

DSC 257R - UNSUPERVISED LEARNING

BEYOND K-MEANS

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

K-Means: the Good and the Bad

The good:

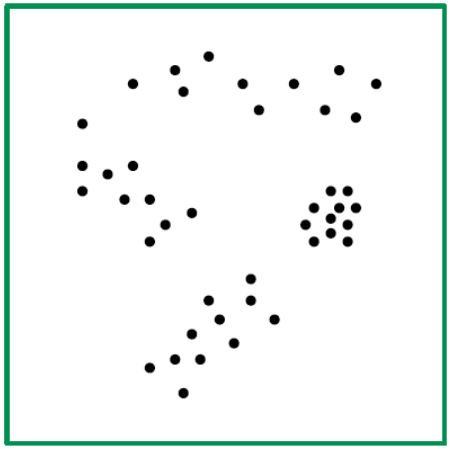
- Fast and easy.
- Effective in quantization.

The bad:

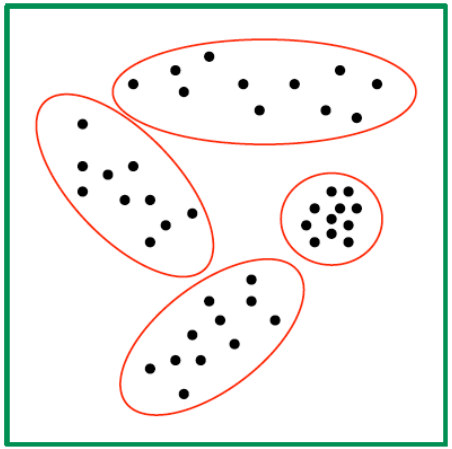
- Geared towards data in which the clusters are spherical, and of roughly the same radius.

Is there is a similarly-simple algorithm in which clusters of more general shape are accommodated?

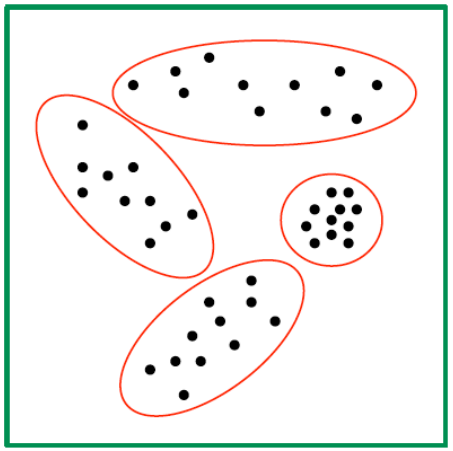
Preview: Mixtures of Gaussians



Preview: Mixtures of Gaussians



Preview: Mixtures of Gaussians



Each of the k clusters is specified by:

- a Gaussian distribution $P_j = N(\mu_j, \Sigma_j)$
- a mixing weight π_j

Overall distribution over $\mathbb{R}^d$: a **mixture of Gaussians**

$$\Pr(x) = \pi_1 P_1(x) + \cdots + \pi_k P_k(x)$$