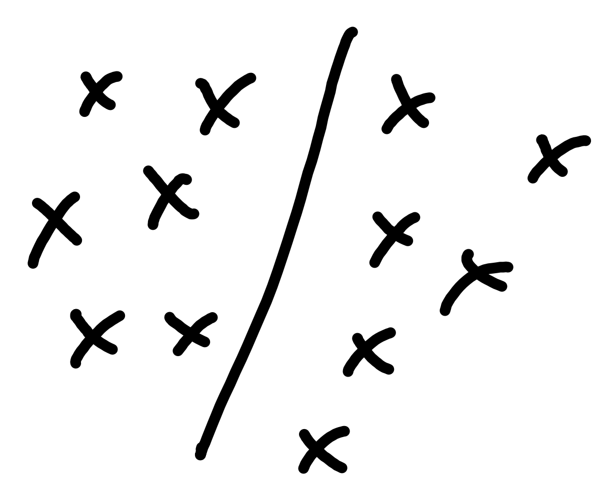
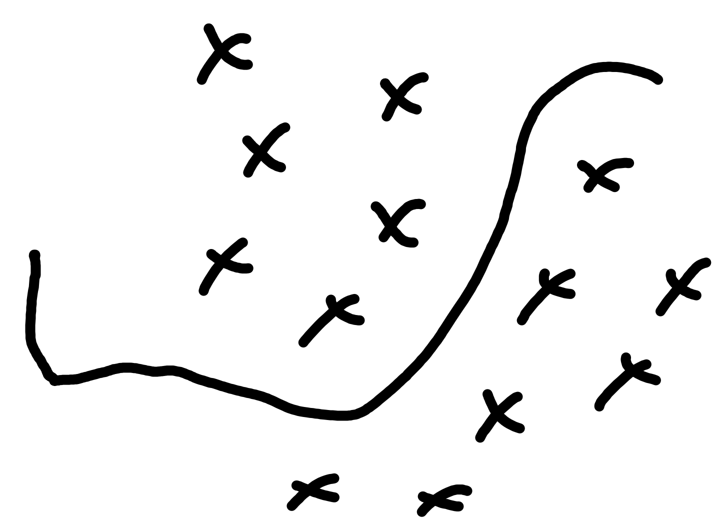
Clustering

In supervised learning (given colour of pts):



linear separ.



two-mean
(and bay cond.)

We can do some kind of such classification.

Next, we consider unsupervised learning and do some clustering algorithm.

The aim is to assign new test data pt $x \in \mathcal{X}$ to one of k clusters. i.e. Define: $\hat{j} : x \in \mathcal{X} \mapsto \{1, 2, \dots k\}$. So: $\mathcal{X} = \sum_1^k \hat{j}^{-1}(i)$.

We will use center-based clustering. i.e. fix k center pts $\{\mu_i\}_1^k \subset \mathcal{X}$. d is metric on $\mathcal{X}$. Then: $\hat{j}(x) = \arg\min_{j \le k} d(x, \mu_j)$.

Assume $\mathcal{X} = \mathbb{R}^d$. Consider Gaussian mixture model $\mathcal{H} = \{ dV(x) = \sum_1^k \xi_j f_j(x) dx .\ f_j = (\sqrt{2\pi}\sigma)^{-d} \cdot e^{-\frac{1}{2} \|(x - \mu_j)/\sigma\|_2^2}\ \mu_j \in [a, b]^d.\ \xi_j \in [0, 1].\ \sum_1^k \xi_j = 1 \}$.

And its EMF $\hat{\mathcal{L}}(\nu, \mathcal{X}_n) = -\sum_{t=1}^{n} \log(\sum_{j=1}^{k} f_j(x_t)\zeta_j)$

$= -\sum_{t=1}^{n} \log(\sum_{j=1}^{k} \zeta_j \exp(-\frac{1}{2}\|x_t - m_j\|_2^2)) + n\frac{d}{2}\log(2\pi\sigma^2)$

Consider $\Theta = \{(\zeta, m) \mid \zeta \in [\varepsilon, 1]^k, m \in \mathbb{R}^{d \times k}\}$.

But this model isn't identifiable by:

$d\nu_\pi = \sum_1^k \zeta_{\pi(j)} N(\cdot \mid m_{\pi(j)}, \sigma^2 I) = d\nu$. for $\forall$

$\pi \in S_k$. (k-permutation)

$J_1 : \theta \in \Theta \mapsto \nu_\theta \in \mathcal{M}$ isn't injective.

$\Rightarrow \Theta$ isn't real parametric model.

Rmk: We can use EM alg. to find EMR for $\hat{\mathcal{L}}_n$. But it's not unique here.

We will rebuild the parameter space:

$\bar{\Theta} := \{(\zeta, m) \in ([\varepsilon, 1]^k \cap \{\mathbb{1}\zeta^T = 1\}) \times [a, b]^{d \times k} :$

$m_{1,1} \le m_{2,1} - \varepsilon \le m_{3,1} - \varepsilon \le \cdots \le m_{k,1} - \varepsilon\}$. i.e. adding some order into Θ. $\Rightarrow \bar{\Theta}$ is cpt and $\theta \in \bar{\Theta} \mapsto \nu_\theta$ is injective

prop. For para. model $\bar{\Theta}$ induced above. For $\nu = \nu_{\theta_0} \in \mathcal{I}_m(\nu)$, $\theta_0 \in \bar{\Theta}$. If $\hat{\theta}_n$ is $\hat{\mathcal{L}}_n$-MLE i.e. $\hat{\theta}_n = \arg\min_{\bar{\Theta}} \tilde{I}_n(V_\theta))$ Then: $\hat{\theta}_n \xrightarrow{pr} \theta_0$

Pf: Since $\bar{\Theta}$ is cpt. We want to apply uniform LLN before.

Note $l(x|\theta) = -\log(\sum \xi_j e^{-\frac{1}{2}\|\frac{x-\mu_j}{\sigma}\|^2}) + \text{Const.}$

is conti. in (ξ, μ).

It remains to prove: $|l(x|\theta)| \le k(x) \in L^1(V_{\theta_0})$

i) $\sum_1^k \xi_j e^{-\frac{1}{2}\square} \le \sum \xi_j = 1.$

2) $\sum_1^k \xi_j e^{-\frac{1}{2}\square} \ge \sum \exp(-\frac{1}{2}\max_j \|\frac{x-\mu_j}{\sigma}\|^2)$

$\ge \sum \exp(-\frac{1}{2}\max_{[a,b]^d} \|x-c\|^2/\sigma^2)$

$\max_{[a,b]^d} \|x-c\|^2 \ge \|x-x_0\|^2 - \text{Const}. \quad x_0 \in [a,b]^d.$

So: $\exists k(x) \in L^1(V_{\theta_0})$.

Next, we see how the data space $\mathcal{X} = \mathbb{R}^d$ be partitioned basing on GMM:

We define decision rule by maximum a-posterior (MAP) principle:

i) View data X as $\mathcal{X}$-valued r.v. following c.d.f. $p(X|\theta)$. θ is $\mathcal{K}$-valued r.v.

ii) Model r.v. $J \in [1,\cdots k]$ which encodes the affiliation of dist. of X. (label)

$$\Rightarrow p(X|\theta) = \sum_1^k p(X,j|\theta) = \sum_1^k p(X|j,\theta)p(J=j|\theta)$$
$$= \sum_1^k f_j(X|\theta)\, \xi_j$$

So: $p(j|X,\theta) = p(x|j,\theta)\cdot p(j|\theta) / p(x|\theta)$

$$= f_j(x|\theta)\,\xi_j / \sum_1^k \xi_j f_j(x|\theta).$$

MAP principle demands $x \in \mathcal{X}$ to be assigned $\hat{j}$ if $\hat{j} = \underset{j=1\cdots k}{\arg\max}\ \xi_j f_j(x|\theta)$.

<u>Rmk:</u> The MAP decision rule will converge for $X_k \overset{i.i.d}{\sim} f_j$ if $\theta \mapsto p(\cdot|\theta)$ is conti. by prop. above. $\forall j$.

For GMM we maximize $\log(\xi_j f_j(x|\theta)) = \log(\xi_j (\sqrt{2\pi}\sigma^2)^{-k}) - \|x-\mu_j\|_2^2/2$.

When $\xi_j = 1/k$. $\forall j$. equal frequency. $\Rightarrow$ MAP maximize $-\|x-\mu_j\|_2^2/2$ by choosing $j \in [1,\cdots k]$.

But in practical, we process as below:

a) Pick random center pts m_j. $j \le k$.

b) Apply MAP rule to get clusters of data

c) Update centers by minimizing:

$$-\log \prod_{\hat{j}(t)=q}^{k} f(X_t \mid m) = -\frac{1}{2} \sum_{t=1,\, \hat{j}(t)=q}^{k} \left\| \frac{X_t - m}{\sigma} \right\|_2^2$$

$\Rightarrow$ Choose $\hat{m}_q = N_q^{-1} \sum_{t=1,\, \hat{j}(t)=q}^{n} X_t$. where $N_q :=$

$|\{t : \hat{j}(t) = q,\ t = 1, 2, \cdots n\}|$

d) Iterate by applying MAP basing on new centers $\{\hat{m}_q\}_1^k$.

e) Stop when no association change clusters.

Rmk: It's called k-means Algorithm. But it sometimes fails to converge.

(It also depends on initialization of m_j)