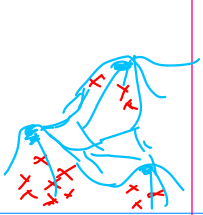
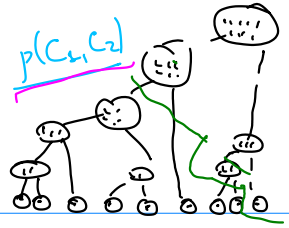


Clustering

$\rho(\bar{x}, \bar{y})$

ACGTGG
AAGTGG

$\rho(x, y)$



$\rho(C_1, C_2)$

$\rho(C_1, C_2) =$
- single-link: $\min_{x \in C_1, y \in C_2} \rho(x, y)$
- complete-link: $\max \rho(x, y)$
- $\rho(m_1, m_2)$
- $\rho(\text{center}_1, \text{center}_2)$
- avg $\rho(x, y)$

MST
minimal spanning tree

$\bar{x}_1 - \bar{x}_n$

$w_{ij} = \rho(\bar{x}_i, \bar{x}_j)$

DBSCAN

core sampled

BIRCH

μ, Σ

$\alpha N_1 + (1-\alpha) N_2$

σ_1 σ_2 μ_1 μ_2

$$p(x \mid \theta) = p(x \mid \mu_1, \mu_2, \sigma_1, \sigma_2, \alpha) = \alpha \cdot \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2} + (1-\alpha) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}(x-\mu_2)^2} \leftarrow p(x/\theta)$$

$\theta_{ML} = \arg\max p(D \mid \theta)$

$$p(D \mid \theta) = \prod_{x \in D} \left[\alpha \cdot e^{-\dots \mu_1} + (1-\alpha) \cdot e^{-\mu_2} \right] \rightarrow \max ?$$

process: $x_n, z_n = \begin{cases} 1, & x_n \in C_1 \\ 0, & x_n \in C_2 \end{cases}$ $\quad D = (x_n, z_n)$

$$p(X, z \mid \theta) = \prod_{(x_n, z_n)} \left[\alpha N(x \mid \mu_1, \sigma_1) \right]^{z_n} \left[(1-\alpha) N(x_n \mid \mu_2, \sigma_2) \right]^{1-z_n}$$

$\downarrow \max ?$

$$\ln p(X, z \mid \theta) = \sum_n \left[z_n \ln(\alpha N_1) + (1-z_n) \ln((1-\alpha) N_2) \right] \rightarrow \max$$

$$= \sum_n \left[z_n \ln \alpha + z_n \ln N_1 + (1-z_n) \ln(1-\alpha) + (1-z_n) \ln N_2 \right] \rightarrow \max$$

(α; μ_1, σ_1; α; μ_2, σ_2)

μ_1^*, σ_1^* : $\sum_n z_n \ln N(x_n \mid \mu_1, \sigma_1) = \sum_{z_n = 1} \ln N(x_n \mid \mu_1, \sigma_1) \rightarrow \max$

μ_2^*, σ_2^* $\quad \mu_1 = \frac{1}{|C_1|} \sum_{z_n = 1} x_n$, $\sigma_1 = \frac{1}{|C_1|} \sqrt{\sum (x_n - \bar{x})^2}$ $\quad p(C_1 \mid \mu_1, \sigma_1) \rightarrow \max$

$\alpha^* = \frac{|C_1|}{N}$

$\sum_n \left(z_n \ln \alpha + (1-z_n) \ln(1-\alpha) \right)$

$|C_1| \ln \alpha + |C_2| \ln(1-\alpha)$

$\frac{|C_1|}{\alpha} - \frac{|C_2|}{1-\alpha} = 0$

$\alpha = \frac{|C_1|}{|C_1| + |C_2|}$

X	$p(X \mid \theta) \rightarrow \max$	Задача	EM - expectation - maximization
X, Z	$p(X, z \mid \theta) \rightarrow \max$	метод	E-step - $z^{(n)} = \mathbb{E}_{p(z \mid \theta^{(n)}, X)} z$
			M-step - $\theta^{(n+1)} = \arg\max p(X, z^{(n)})$

$X = \{x_1 \dots x_N\}$

$z = \{z_1 \dots z_N\}$

$\theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$

$p(x \mid \theta) = \prod \left(\alpha \cdot \alpha_1 + (1-\alpha) N_2 \right) \rightarrow \max$

$p(x, z \mid \theta) = \prod (\cdot)^{z_n} (\cdot)^{(1-z_n)}$

- init $\theta^{(0)}$ randomly
- E-step: $\mathbb{E} z_n = p(x_n \in C_1 \mid \theta^{(n)}) = p(C_1 \mid x_n, \theta^{(n)}) = \frac{p(C_1) p(x_n \mid C_1)}{p(C_1) p(x_n \mid C_1) + p(C_2) p(x_n \mid \theta)}$

$$\mathbb{E} z_n = \frac{\alpha N(x_n \mid \mu_1, \sigma_1)}{\alpha N(x_n \mid \mu_1, \sigma_1) + (1-\alpha) N(x_n \mid \mu_2, \sigma_2)}$$

C_1 C_2 x_n

- M-step: $p(x, z^{(n)} \mid \theta) =$

$\mathbb{E} z_n$ $\quad (1 - \mathbb{E} z_n)$ $\quad p(C_1) p(x_n \mid C_1)$ $\quad p(C_2) p(x_n \mid C_2)$

$$p(x, z^{(n)} \mid \theta) = \prod_n \left(\alpha \cdot N(x_n \mid \mu_1, \sigma_1) \right) \cdot \left((1-\alpha) N(x_n \mid \mu_2, \sigma_2) \right) \rightarrow \max$$

$$\ln p = \sum_n \left[\mathbb{E} z_n \cdot \ln \alpha + (1 - \mathbb{E} z_n) \ln(1-\alpha) + \mathbb{E} z_n N(x_n \mid \mu_1, \sigma_1) + (1 - \mathbb{E} z_n) \cdot N(\dots_2) \right]$$

$\rightarrow \max$

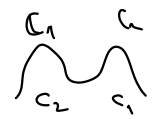
EM:

$\theta^{(0)} \xrightarrow{E} z^{(1)} \xrightarrow{M} \theta^{(1)} \xrightarrow{E} z^{(2)} \xrightarrow{M} \theta^{(2)} \xrightarrow{E} \dots$

$\alpha^{(n+1)} = \frac{\sum \mathbb{E} z_n}{N}$

$\mu_1^{(n+1)} = \frac{1}{\sum \mathbb{E} z_n} \cdot \sum_n \mathbb{E} z_n \cdot x_n$

$\sigma_1^{(n+1)} = \dots \sqrt{\sum (\dots)^2}$



(x_n, z_n) $\quad p(X,Z|\theta) \to \max$

$p(X|\theta) = \int p(X,Z|\theta)\,dz \to \max$

$\log p(x,z|\theta) = \sum_n \left[z_n \log \alpha N_1 + (1-z_n) \log (1-\alpha) N_2 \right] \to \max$

$Q = \mathbb{E}_{\theta_n} \log \ldots = \mathbb{E}_{z|x,\theta_n} \left[z_n \log \alpha N_1 + (1-z_n) \log (1-\alpha) N_2 \right]$

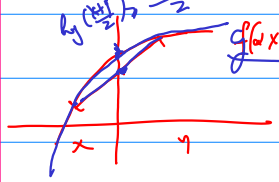
$Q(\theta,\theta_n) = \sum_k \left[\mathbb{E}_{\theta_n} z_k \right] \cdot \log \alpha N_1 + (1 - \mathbb{E}_{\theta_n} z_k) \log (1-\alpha) N_2 \to \max_\theta$

$\mathbb{E}_{\theta_n} z_k$ — $\alpha_n N(x_k | \mu_1^n, \Sigma_1^n)$

$\alpha_n N(\ldots \mu_1^n) + (1-\alpha_n) N(\ldots \mu_2^n)$

$\int f(z) q(z|x,\theta_n)\,dz = \mathbb{E}_{z|x,\theta_n} f(z)$

$\log \mathbb{E}_z f(z) \geq \mathbb{E}_z \log f(z)$ — Jensen's inequality



$g(\alpha x + (1-\alpha) y) \geq \alpha g(x) + (1-\alpha) g(y)$

$g(\sum \alpha_k x_k) \geq \sum_k \alpha_k g(x_k), \quad \sum \alpha_k = 1$

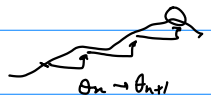
$g(\mathbb{E}x) \geq \mathbb{E}\, g(x)$

EM:

$Q(\theta,\theta_n) = \mathbb{E}_{p(z|x,\theta_n)} \log p(x,z|\theta)$

$\theta_{n+1} := \arg\max_\theta Q(\theta,\theta_n)$

Means: $\log p(x|\theta_n) < \log p(x|\theta_{n+1})$



$\log p(x|\theta) - \log p(x|\theta_n) =$

$= \log \int p(x,z|\theta)\,dz - \log p(x|\theta_n) =$

$= \log \int p(x,z|\theta) \cdot \frac{p(z|x,\theta_n)}{p(z|x,\theta_n)}\,dz - \log p(x|\theta_n)$

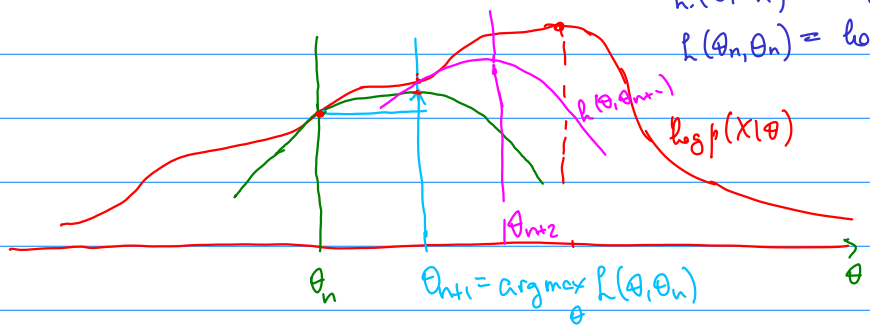
$= \log \int \frac{p(x,z|\theta)}{p(z|x,\theta_n)} \cdot p(z|x,\theta_n)\,dz - \log p(x|\theta_n) \geq$

$\geq \int \log \frac{p(x,z|\theta)}{p(z|x,\theta_n)} p(z|x,\theta_n)\,dz - \log p(x|\theta_n) =$ — $h(\theta,\theta_n)$

$= \int \log \frac{p(z|x,\theta)\,p(x|\theta)}{p(z|x,\theta_n)\,p(x|\theta_n)} \cdot p(z|x,\theta_n)\,dz$

$h(\theta,\theta_n) \leq \log p(x|\theta)$

$h(\theta_n,\theta_n) = \log p(x|\theta_n)$



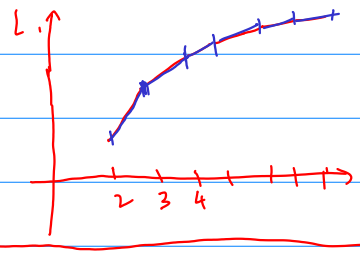
$\int \log \frac{p(x,z|\theta)}{p(z|x,\theta_n)} \cdot p(z|x,\theta_n)\,dz$ — $\theta \to \max$

$= \int \log p(x,z|\theta)\, p(z|x,\theta_n)\,dz -$

$Q(\theta,\theta_n)$

$- \int \log p(z|x,\theta_n)\, p(z|x,\theta_n)\,dz$

$H(z|\theta_n,x)$



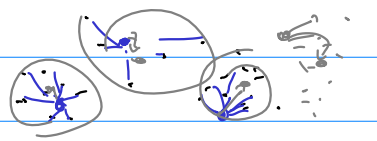
Model selection

$BIC(M) = \ln L - \frac{1}{2} M \cdot \ln N$

$d = 2 - \mu, \sigma \quad - 2$

$d = 10 - \bar{\mu} \cdot 10, \ \Sigma - \frac{10 \cdot 11}{2}$ — 56

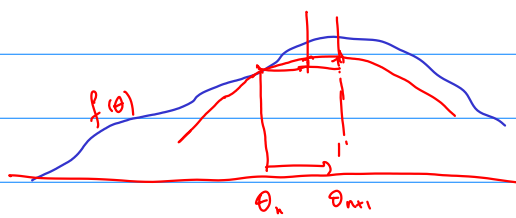
k-means



$\mu_1 - \mu_{10} \to z_n \to \mu_1 \to z_n$

$\mathbb{E}_z f(z) \approx \frac{1}{R} \sum_{r=1}^{R} f(z_r)$

$z_r \sim p(z|x,\theta_n)$



$L(\theta,\theta_n) \leq f(\theta)$

$L(\theta_n,\theta_n) = f(\theta_n)$

$\theta_{n+1}: \ L(\theta_{n+1},\theta_n) > L(\theta_n,\theta_n)$

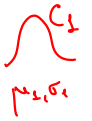
Generalized EM

$\theta_{n+1} = \arg\max_\theta \mathbb{E}_{z|x,\theta_n} \log p(x,z|\theta)$

$= \int \log p(x,z|\theta)\, p(z|x,\theta_n)\,dz \approx \frac{1}{R} \sum_{z=1}^{R} \log p(x,z_z|\theta)$ — (x,z)

$\theta \to \max$

Stochastic EM



x_i — x_n

m

ACGTACGT
AGGTCCGT — C_k
AGCTCCGT
CGAACGGA

x_n

$$\rho(x,y) = \frac{\#\{x_i \neq y_i\}}{|x| \geq m}$$

C_1 : 1 ◇ ◇ ... ◇ m (3m)

p_{mA} p_{mC} p_{mT} — — — i

$p_{mT} = 1 - \sum p_{m\ldots}$

- init θ $C_1 \ldots C_K$

- $\mathbb{E} z_{nk} = \dfrac{\alpha_k p(x_n|C_k)}{\sum \alpha_j p(x_n|C_j)} = \alpha_k \prod_{s=1}^{m} \prod_{l \in ACGT} p_{klm}^{x_{nml}} \Big/ \sum \ldots$

- $p_{klm} := \dfrac{\#\{\text{букв } l \text{ на поз. } m \text{ в } C_k\}}{\#\{\text{букв на поз. } m \text{ в } C_k\} = |C_k|}$

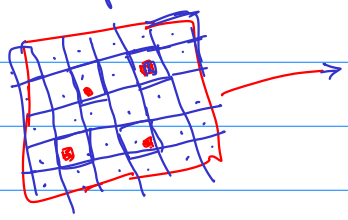
$$p_{klm} := \frac{\sum_{n:\, x_{nm}=l} \mathbb{E} z_{nk}}{\sum_{n} \mathbb{E} z_{nk}}$$

$$\mathbb{E}_{z|x,p^{(m)}} \log p(X, Z \mid p_{klm}) = \mathbb{E}_z \log \prod_n \prod_k \prod_m [\ldots]^{z_{nk}} z$$

$$z \mathbb{E}\left[\sum_{nk} z_{nk} \cdot \sum_m [\ldots] \right] \to \max$$

$$\sum \mathbb{E} z_{nk} \cdot [\ldots] \to \max$$

Multiple instance learning (MIL)



Yes No

$X = \underbrace{x_{11} \ldots x_{1n}}_{=0} \underbrace{x_{21} \ldots x_{2n}}_{=0} \ldots x_{m1} \ldots x_{mn}$

Yes No — Yes

x_{ij} z_{ij}

z_{ij}

$$z_{ij} := \frac{p(x_{ij} = \text{Yes})}{\sum \ldots}$$