"Series" of topics connected

(1.) Probabilistic approach

2. Infer some structure from data

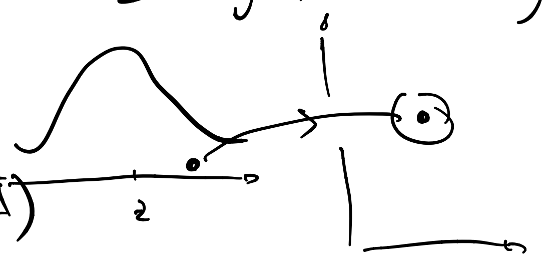
1. Tools

1. $\mathcal{N}(x \mid \underset{\text{covariance}}{\overset{\text{mean}}{\mu, \Sigma}}) = \underbrace{|2\pi\Sigma|^{-\frac{1}{2}}}_{\text{normilization constant} \sim \text{volume}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

2. $\log \mathcal{N}(x \mid \mu, \Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$

3. $p(z) \sim \mathcal{N}(z \mid \mu, I)$

$x = \underset{\downarrow \text{ linear}}{W} z + \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$



$x \sim \mathcal{N}(x \mid ?, ?)$

PCA: clever choice of $\boxed{W}$ $\downarrow$ MLE.

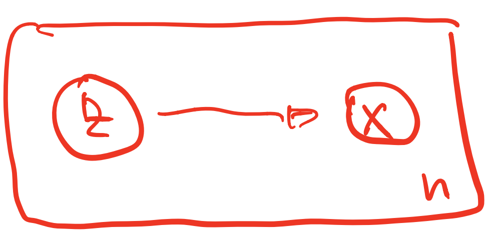
$E x = E_\varepsilon E_z \, Wz + \varepsilon = W\mu;$

$E xx^T = E_\varepsilon E_z (Wz+\varepsilon)(Wz+\varepsilon)^T =$

$= E_\varepsilon E_z \, W z z^T W^T + \underbrace{W z \varepsilon^T}_{=0} + \underbrace{\varepsilon (zW)^T}_{=0} + \varepsilon\varepsilon^T =$

$= W \underbrace{E_z z z^T}_{W(I+\mu\mu^T)W^T} W^T + \underbrace{E_\varepsilon \varepsilon\varepsilon^T}_{\sigma^2 I}$

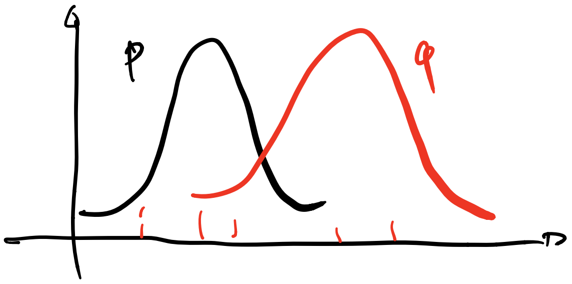
$\operatorname{cov} x = WW^T + \sigma^2 I$.



Wz

4. Distance between distributions?

= Total Variation



$$\int |p(x)-q(x)|\,dx$$

in each point $\Rightarrow$

$$\int q(x)\left[\log q(x) - \log p(x)\right]dx$$

$$\int q(x)\log\frac{q(x)}{p(x)}\,dx = KL\{q;p\}.$$

$= 1 \Rightarrow \log 1 = 0$ ✓

< 1 ✓

> 1 ✓

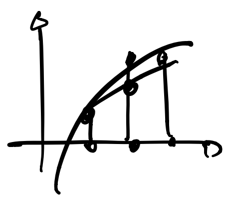
ratio is fine ✓

1. ? Is it positive ? Is it bounded from below

$$\int q(x)\log\frac{p(x)}{q(x)}\,dx \le \log\int q(x)\frac{p(x)}{q(x)}\,dx = \log\underbrace{\int p(x)dx}_{=1} = 0.$$

$w_1 f(\cdot) + w_2 f(\cdot)$

$w_1 + w_2 = 1$



$$w_1\log(x_1) + w_2\log(x_2) \le \log(w_1x_1 + w_2x_2)$$

$w_1, w_2 > 0$

$w_1 + w_2 = 1$

$$-KL\{q|p\} \le 0 \Rightarrow KL\{q|p\} \ge 0.$$

$$KL\{q|p\} > 0, \; = 0 \text{ iff } q \overset{?}{=} p.$$

2. KL convex on both q, p.

3. EM algorithm

(1) Optimization: Minimization-Maximization algorithms

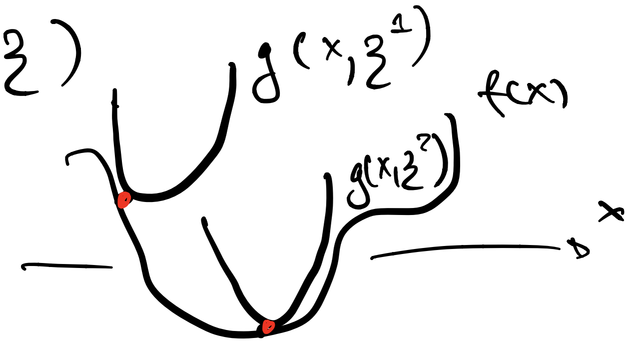
MM

$$\min_x \underbrace{f(x)}$$

not suitable for optimization

1) non-convex
2) non-smooth
3) computationally costly

"variational bound" $g(x, \xi)$



$$\forall \xi \quad f(x) \le g(x, \xi)$$

$$f(x^0) = \min_{\xi} g(x^0, \xi).$$

(2) Probabilistic Approach $\Rightarrow$ Build Such Bounds

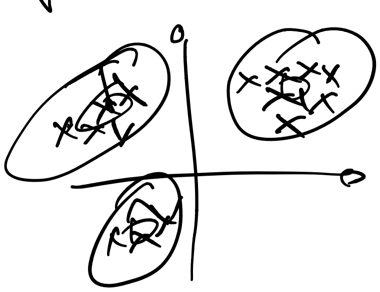
"Expectation".

dat points

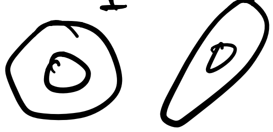
"K-Means"

$$p(X) = \prod_{n=1}^{N} p(x_n) \qquad p(x_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

MLE-problem $\prod_{n=1}^{N} \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \right] \sim K^N$ $\propto 1000$ 10 very big



I



fit :)

= covariance

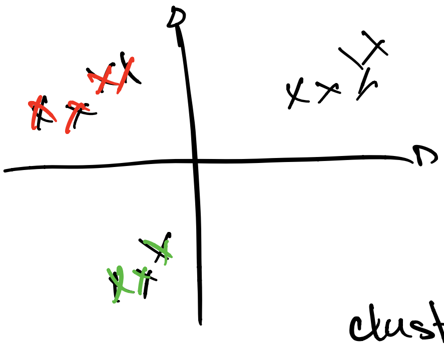
k-Means

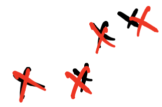
1. If we have clusters?
2. We know assigns to clusters? [E M :))

$$p(X) = \prod_{n=1}^{N} \Big[\sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \underbrace{\mu_k}, \underbrace{\Sigma_k}) \Big]$$

1. We now, each point cluster label

usuall MLE.





$$\hat{\mu} = \frac{1}{N} \left(x\, x^{\,x\,x} \right)$$

$$\Sigma = \sum_{n=1}^{N} (x_n - \hat{\mu})(x_n - \hat{\mu})^T$$

clusters => MLE.

2. Know parameters, how to find clusters?

x_n * → label z_{nk} $\{0, \dots, 1, \dots, 0\}$ one-hot encode of K cluster

coordinates — kth

$$p(z_{nk} = 1) = \pi_k$$

1. Flip coin (K-dim) $p(z_{nk} = 1) = \pi_k$ $\pi_1, \dots, \pi_K$
2. => choose cluster
3. generate point with μ_k, Σ_k

$$\underbrace{p(x_n)} = \int \underbrace{p(x_n \mid z)}\, p(z)\, dz$$

$$\sum_{K} \left[\mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right]^{z_{nk}} \{\pi_k\}^{z_{nk}} =$$

$$= \sum_{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \square$$

$$\underline{p(z_{nk}=1|x_n)} = \frac{p(x_n|z_{nk}=1)\,p(z_{nk}=1)}{p(x_n)} =$$

$$= \{\text{Bayes theorem } p(A|B) = \frac{p(B|A)p(A)}{p(A)}\} =$$

$$= \frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)\,\pi_k}{\sum_{k=1}^{K}\pi_k\,\mathcal{N}(x_n|\mu_k,\Sigma_k)}$$

$p(z_{n2}=1|x_n)$ 2 $\quad$ 1 $\;p(z_{n1}=1|x_n)$

EM.

$$\log p(X) = \{\log p(X)\}\cdot 1 =$$

$$= \int q(Z)\log \underbrace{p(X)}\, dZ =$$

$$\{\text{" } p(X,Z) = \underbrace{p(X|Z)}_{\text{"our model}}\,p(Z) \text{ "}\}$$

$$\{\ p(X,Z) = p(Z|X)\,\underbrace{p(X)}\ \}$$

$$= \int q(Z)\log\frac{p(X,Z)}{p(Z|X)}\,dZ =$$

$$= \int q(Z)\log\frac{p(X|Z)\,p(Z)}{p(Z|X)}\,dZ =$$

$$= \int q(Z)\log\frac{p(X|Z)\,p(Z)}{q(Z)}\,dZ + \underbrace{\int q(Z)\log\frac{q(Z)}{p(Z|X)}\,dZ}_{KL\geq 0}.$$

$$\log p(x) \geqslant \underbrace{\int q(z) \log \frac{p(x|z)\,p(z)}{q(z)}\, dz}$$

$$\max_\theta \log p_\theta(x) \;?$$

$$\max_\theta \log \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)$$

$$\theta = \{\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K\}$$

Variational Bound

1. $\log p_\theta(x) \geqslant \int q(z|x) \log \dfrac{p_\theta(x|z)\,p_\theta(z)}{q(z|x)}\, dz$

$\forall \theta$

2. $\forall \theta, \exists\, q^*(z|x)$

$$\log p_\theta(x) = \int q^*(z|x) \log \frac{p_\theta(x|z)\,p_\theta(z)}{q^*(z|x)}\, dz$$

$$q^*(z|x) = p(z|x).$$

$\underline{\log p_\theta(x)}$

$$= \underline{\int q(z) \log \frac{p(x|z)p(z)}{q(z)}\, dz} + \underbrace{\int q(z) \log \frac{q(z)}{p(z|x)}\, dz}.$$

$$q(z) = p(z|x) = 0.$$

How to optimize? GMM (Gaussian Mixture Model)

1. $q^*(Z|X) = P_\theta(Z|X)$

3. $p(z_{nk}=1|x_n) = \dfrac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}$

fixed!

q_{nk}

2. $\max_\theta \int q(Z) \log p(X|Z)p(Z)\,dZ$

1. q_{nk} $\forall n, \forall k$

$$\max_\theta \int q(Z) \log p_\theta(X|Z)\,p(Z)\,dZ =$$

$$= \max_\theta \int q(Z) \log \prod_{n=1}^{N}\prod_{k=1}^{K} \left[\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right]^{z_{nk}} dZ$$

$\in \{0,1\}$

$$= \int \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\log \pi_k + \log \mathcal{N}(x_n|\mu_k,\Sigma_k)\} \times q(Z)\,dZ$$

$E\, z_{nk} = q_{nk}$

fixed

$$\max_{\theta=\{\mu_k,\pi_k,\Sigma_k\}} \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\{\log \pi_k + \log \mathcal{N}(x_n|\mu_k,\Sigma_k)\}$$

$$\nabla_{\mu_k}: \sum_{n=1}^{N} q_{nk} \{ \nabla_{\mu_k} \log \mathcal{N}(x_n | \mu_k, \Sigma_k) \} = 0$$

$$\nabla_{\mu_k} \log \mathcal{N}(x_n | \mu_k, \Sigma_k) = \nabla_{\mu_k} -\frac{1}{2} (x-\mu_k)^T \Sigma^{-1} (x-\mu_k) =$$

$$= +\frac{2}{2} \Sigma^{-1} (x-\mu_k)$$

$$\sum_{n=1}^{N} q_{nk} \Sigma_k^{-1} (x_n - \mu_k) = 0$$

$$\Sigma_k^{-1} \Big(\sum_{n=1}^{N} q_{nk} \cdot x_n \Big) = \Sigma_k^{-1} \Big(\sum_{n=1}^{N} q_{nk} \cdot \mu_k \Big)$$

"weights of assigning" points

$$\mu_k = \frac{\sum_{n=1}^{N} q_{nk} \cdot x_n}{\sum_{n=1}^{N} q_{nk}} .$$

1. $$q_{nk} = \frac{\pi_k^{old} \mathcal{N}(x_n | \mu_k^{old}, \Sigma_k^{old})}{\sum_k \text{———} " \text{———}}$$

2. $$\mu_k^{new} = \frac{\sum q_{nk} \cdot x_n}{\sum q_{nk}}$$

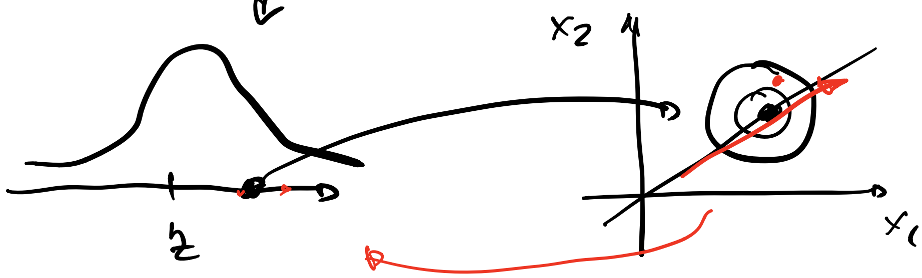
update q_{nk}.

PCA model.

$p(z) = \mathcal{N}(z|0, I_d)$ $\quad W \; z$ $\quad D \gg d.$

$\quad\quad\quad\quad\quad\quad\quad\quad\quad D\times d \;\; d\times 1$

↓ small dim



$p(x|z) = \mathcal{N}(x|Wz + \mu, \sigma^2 I).$

$\theta = \{W, \mu, \sigma\}$

E-step $\quad p(z_n|x_n) = \dfrac{p(x_n|z_n)\,p(z_n)}{p(x_n)} =$

$\propto \mathcal{N}(x_n|W z_n + \mu, \sigma^2 I)\,\mathcal{N}(z_n|0, I)$?

$\log p(z_n|x_n) = -\frac{\sigma^2}{2}(x_n - W z_n - \mu)^T(x_n - W z_n - \mu) - \frac{1}{2} z_n^T z_n.$

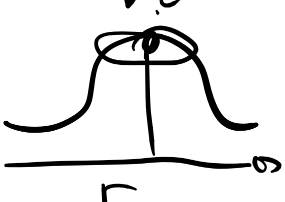
$\Rightarrow$ quadratic form with respect to z

var(Hessian$)^{-1}$ $\Rightarrow$ $p(z_n|x_n)$ again Normal

vert.

$\Rightarrow$ mean, $\quad$ covariance

$= \nabla = 0 \quad\quad = [\nabla^2]^{-1}$.



$\nabla_z \quad -\frac{\sigma^2}{2} \times (-2)\, W^T(x_n - W z_n - \mu) - z_n = 0$

$\sigma^2 W^T x_n - W^T W z_n - W^T \mu - z_n = 0$

$(I + W^T W) z_n = \sigma^2 W^T (x_n - \mu)$

$$z_n^* = \underbrace{(I+W^TW)^{-1}}_{\text{covariance.}} \sigma^{?} W^T(x_n-\mu) \Rightarrow \text{mean.}$$

$$p(z_n|x_n) = \mathcal{N}(z_n | \text{mean} \quad \text{covariance}^{-1}).$$

<u>M-step</u>

$$\int q(Z|X) \log p(X|Z)p(Z)\, dZ$$

$$\mathbb{E}_q \sum_{n=1}^{N} \log p(x_n|z_n)p(z_n) =$$

$$= \mathbb{E}_q \sum_{n=1}^{N} -\frac{\sigma^2}{2}(x_n - Wz_n - \mu)^T(x_n - Wz_n - \mu) - \frac{1}{2} z_n^T z_n \rightarrow \max_{W,\mu,\sigma}$$

→ do this home.

$D\mu$:

$$\sum_{n=1}^{N} \mathbb{E}_q\, \sigma^2(x_n - Wz_n - \mu) = 0.$$

$$\frac{1}{N}\sum_{n=1}^{N} x_n - W\, \mathbb{E}_q z_n = \mu.$$

$$\frac{1}{N}\sum_{n=1}^{N} x_n - W \text{ input}$$

<u>W.</u> 1. Take D
2. Don't forget $\mathbb{E}_{q(z)}$.