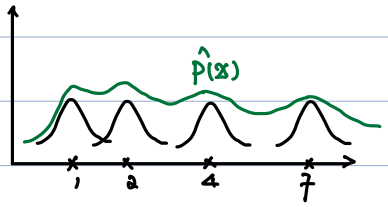
0 1 2 4 7 0 1 2 2.5 4 4.5 7

Observation: larger K, smoother outputs.

### 3) Parzen Window Estimation with Gaussian Kernels




→ Each sample has a Gaussian Distribution.

→ Add all the probability

→ Scale to have CDF = 1

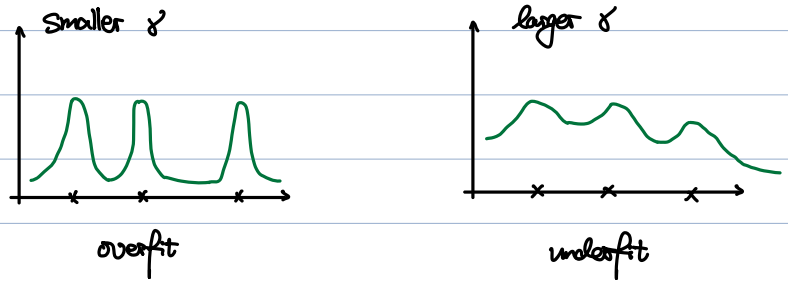
Details: Standard normal distribution (PDF)

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$\hat{p}(x) = \frac{1}{k} \sum_{i=1}^{N} \phi\left(\frac{\|x - x_i\|}{\gamma}\right)$$

$\gamma$: kernel width, width of each sample affected neighbor.

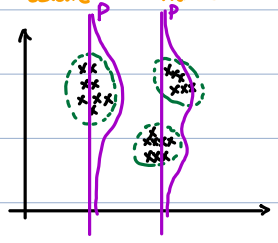
k: Normalizing constant




Note: If more points → complex computation.

## Lec 22.

### 4) Gaussian Mixture Model




Each cluster with its own Gaussian distribution

It suffices to use only k Gaussian distributions, one for each cluster.

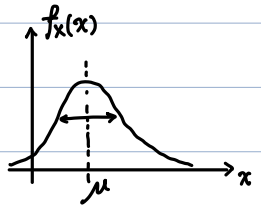
- mean: $\underline{\mu}_1, \underline{\mu}_2, \cdots, \underline{\mu}_k \in \mathbb{R}^d$

  covariance: $\Sigma_1, \Sigma_2, \cdots, \Sigma_k \in \mathbb{R}^d \times \mathbb{R}^d$

Review:

$d=1$: PDF of Gaussian R.V. X with mean $\mu$, and variance $\sigma^2$.

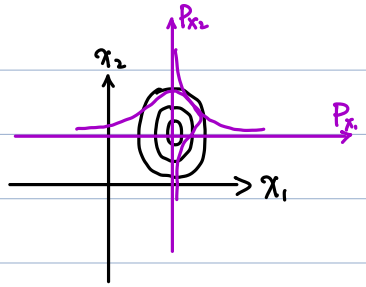
$$f_X(x) \equiv \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\left(\frac{x-\mu}{2\sigma^2}\right)^2}$$




$d=2$: PDF of jointed Gaussian R.V. $X_1$ and $X_2$, with mean $\mu_1$ & $\mu_2$, variances $\sigma_1^2$ & $\sigma_2^2$, and correlation coefficient $\rho$.

$$\rho = \frac{E[(X_1-\mu_1)(X_2-\mu_2)]}{\sigma_1 \cdot \sigma_2}$$

$$f_{X_1,X_2}(x_1,x_2) \equiv \mathcal{N}(x;\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$$
$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}$$




General $d>>1$:

Consider $d$ jointly Gaussian RVs, $X_1, X_2 \cdots X_d$, with $\mu_1, \mu_2 \cdots \mu_d$ and covariance matrix.

$$\Sigma \triangleq \begin{bmatrix} cov(X_1,X_1) & cov(X_1,X_2) & \cdots & cov(X_1,X_d) \\ \vdots & \ddots & \ddots & \vdots \\ cov(X_d,X_1) & \cdots & \cdots & cov(X_d,X_d) \end{bmatrix}$$

where $cov(X_i,X_j) = E[(X_i-\mu_i)(X_j-\mu_j)]$

$$\Leftrightarrow \Sigma = E[(X-\underline{\mu})(X-\underline{\mu})^T]$$

$$f_{\underline{X}}(\underline{x}) = \mathcal{N}(\underline{x},\underline{\mu},\Sigma)$$
$$= \frac{1}{(2\pi)^{d/2}\cdot \det(\Sigma)^{1/2}} \cdot e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})}$$

$$\frac{e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Data Generalization Model (Hypothesis)

(for each data point in $\mathcal{D}$)

First, pick a bump (cluster)

$j \in \{1, 2, \cdots, k\}$ according to probability distribution $\{w_1, w_2 \cdots, w_k\}$

$$\begin{cases} w_i > 0 \\ \sum_{i=1}^{k} w_i = 1 \end{cases}$$

Then, generate $\underline{x}$ according to probability density $P(\underline{x}|j) = \mathcal{N}(\underline{x}_j, \underline{\mu}_j, \Sigma_j)$

$$= \frac{1}{(2\pi)^{d/2}\cdot \det(\Sigma_j)^{1/2}} \cdot e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_j)^T\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j)}$$

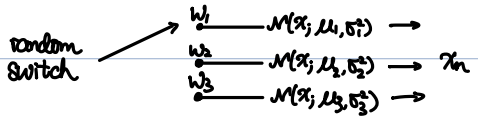
$\Rightarrow$ the overall PDF for $\underline{x}$ is

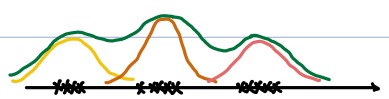
$$P(\underline{x}) = \sum_{j=1}^{k} w_j P(\underline{x}|j) = \sum_{j=1}^{k} w_j \cdot \mathcal{N}(\underline{x}_j, \underline{\mu}_j, \Sigma_j)$$

Example: $d=1$, $k=3$

$\mathcal{D} = \{x_1, x_2 \cdots, x_N\}$

For each index $n$,



Given $\mathcal{D}$, need to estimate $\mu_1, \mu_2, \mu_3$ & $\sigma_1, \sigma_2, \sigma_3$.

Given $\mathcal{D} = \{\underline{x}_1, \cdots \underline{x}_N\}$, and $k$. find a GMM $\Omega = \{w_j, \underline{\mu}_j, \Sigma_j\}_{j=1}^{k}$ that is the best fit for $\mathcal{D}$.

Best-fit: MLE

$\hat{P}_\Omega(\mathcal{D}) = \prod_{n=1}^{N} \hat{P}(x_n)$ is maximized.

Minimize $E_n(\Omega) \triangleq -\log \hat{P}_\Omega(\mathcal{D})$

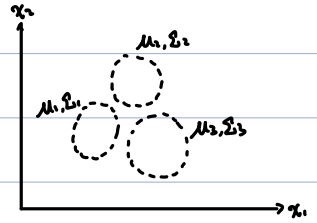
$$= -\sum_{n=1}^{N} \log \hat{P}(x_n)$$

$$e_n(\Omega) = -\log \hat{P}(x_n)$$

How to minimize $E_{in}(\Omega)$ ?

SGD?

## Lecture 23




$$D = \{x_1, x_2, \cdots, x_N\}$$

$$P(x) = \sum_{j=1}^{K} w_j \cdot \mathcal{N}(x_j, \mu_j, \Sigma_j)$$

Given set $D$ and $k$, find the GMM $\Omega = \{w_j, \mu_j, \Sigma_j\}_{j=1}^{k}$ that is the "best fit" for $D$.

Maximize the likelihood

$$\hat{P}_{\Omega}(D) = \prod_{n=1}^{N} \hat{P}(x_n)$$

$\equiv$ Minimize $E_{in}(\Omega) = -\log \hat{P}_{\Omega}(D)$

$$= \sum_{n=1}^{N} -\log \hat{P}(x_n)$$

$$= \sum_{n=1}^{N} \underbrace{-\log \sum_{j=1}^{k} w_j \cdot \mathcal{N}(x_n; \mu_j, \Sigma_j)}_{e_n(\Omega)}$$

Can use SGD ? No

- Complicated $e_n$
- $\sum_j w_j = 1$ . constrained optimization

### The Expectation Maximization (EM) Algorithm

An alternating optimization approach with 2 subproblems.

#### Subproblem 1. (M step)

Given $D = \{x_1, \cdots, x_N\}$, suppose for each $x_i \in D$, we know the bump (Gaussian distribution) it belongs to.

Let $B_j \in D$ denotes the $j^{th}$ bump.

i.e. all points in $B_j$ are sampled from

$$P(x|j) = \mathcal{N}(x; \mu_j, \Sigma_j)$$

Estimate $w_j, \mu_j, \Sigma_j$ for $j = 1, 2 \cdots k$

$$w_j = \frac{N_j}{N} \quad . \ N_j : \text{\# of points in } B_j$$

$$\mu_j = \frac{1}{N_j} \sum_{x_n \in B_j} x_n \quad \text{(sample mean)}$$

$$\Sigma_j = \frac{1}{N_j} \sum_{x_n \in B_j} (x_n - \mu_j)(x_n - \mu_j)^T$$

(sample covariance)

#### Subproblem 2. (E step)

Given $\Omega = \{w_j, \underline{\mu}_j, \Sigma_j\}_{j=1}^{K}$, estimate bump membership. ie. for each $\underline{x}_n$, find bump $j$ that is most likely to produce $\underline{x}_n$.

$$\underline{x}_n \in B_{j^*} \quad \text{if} \quad j^* = \underset{j \in \{1,2,\cdots k\}}{\operatorname{argmax}} \; P_r(j|\underline{x}_n)$$

Recall: Bayes' Thm

$$P_r\{j|\underline{x}_n\} = \frac{P_r\{\underline{x}_n|j\} \cdot P_r\{j\}}{P_r(\underline{x}_n)} = \frac{P_r\{\underline{x}_n|j\} \cdot P_r\{j\}}{\sum_{i=1}^{k} P(\underline{x}_n|i) \cdot P_r(i)} = \frac{\mathcal{N}(\underline{x}_n; \underline{\mu}_j, \Sigma_j) \cdot w_j}{\sum_{i=1}^{k} \mathcal{N}(\underline{x}_n; \underline{\mu}_i, \Sigma_i) \cdot w_i}$$

$$\Rightarrow j^* = \underset{j}{\operatorname{argmax}} \; \mathcal{N}(\underline{x}_n; \underline{\mu}_j, \Sigma_j) \cdot w_j$$

EM algo. Summary: (Hard decisions)

1. Initialization: start with arbitary bump membership for each $\underline{x}_n$.
2. Estimate $\Omega = \{w_j, \underline{\mu}_j, \Sigma_j\}_{j=1}^{K}$ given bump membership $\{B_1, \cdots B_k\}$ (subproblem 1)
3. Estimate bump membership $\{B_1, \cdots B_k\}$ given $\Omega$ (subproblem 2)

Repeat step 2 & 3 until convergence.

Note: ① Convergence is guaranteed.

② $B_1, \cdots, B_k$ are auxiliary ("hidden" variables).