Lecture 6/13 : Generative Models so far

train $P(X)$ = density / curve [param] fit to data X (Sep for each Y)

predict $P(Y|X) = \dfrac{P(X|Y) \cdot P(Y)}{P(X)}$

- For each Y $\quad P(X|Y) = \mathcal{N}(X|\mu, \underset{\text{D-dim}}{\Sigma}) \Rightarrow$ Gauss DA
- For each Y $\quad P(X|Y) = P(x^1\, x^2 \ldots x^D | Y)$
  $= P(x^1|Y) \cdot P(x^2|Y) \ldots P(x^D|Y) \Rightarrow$ Naive Bayes
- Not-So-Naive Bayes: → Belief Network
  group features that are very dependent
  indep

For each Y: $\quad P(X|Y) = P(x^1\, x^2 \ldots x^D | Y) =$

$$= P(\underbrace{\boxed{x^1, x^2}}_{\text{2dim}}|Y) \cdot P(\underbrace{\boxed{x^3}}_{\text{1-dim}}|Y) \cdot P(\underbrace{\boxed{x^4, x^5}}_{\text{2dim}}|Y, x^1) \cdot P(x^6, x^7|Y) \ldots$$

fit for Y, x^1

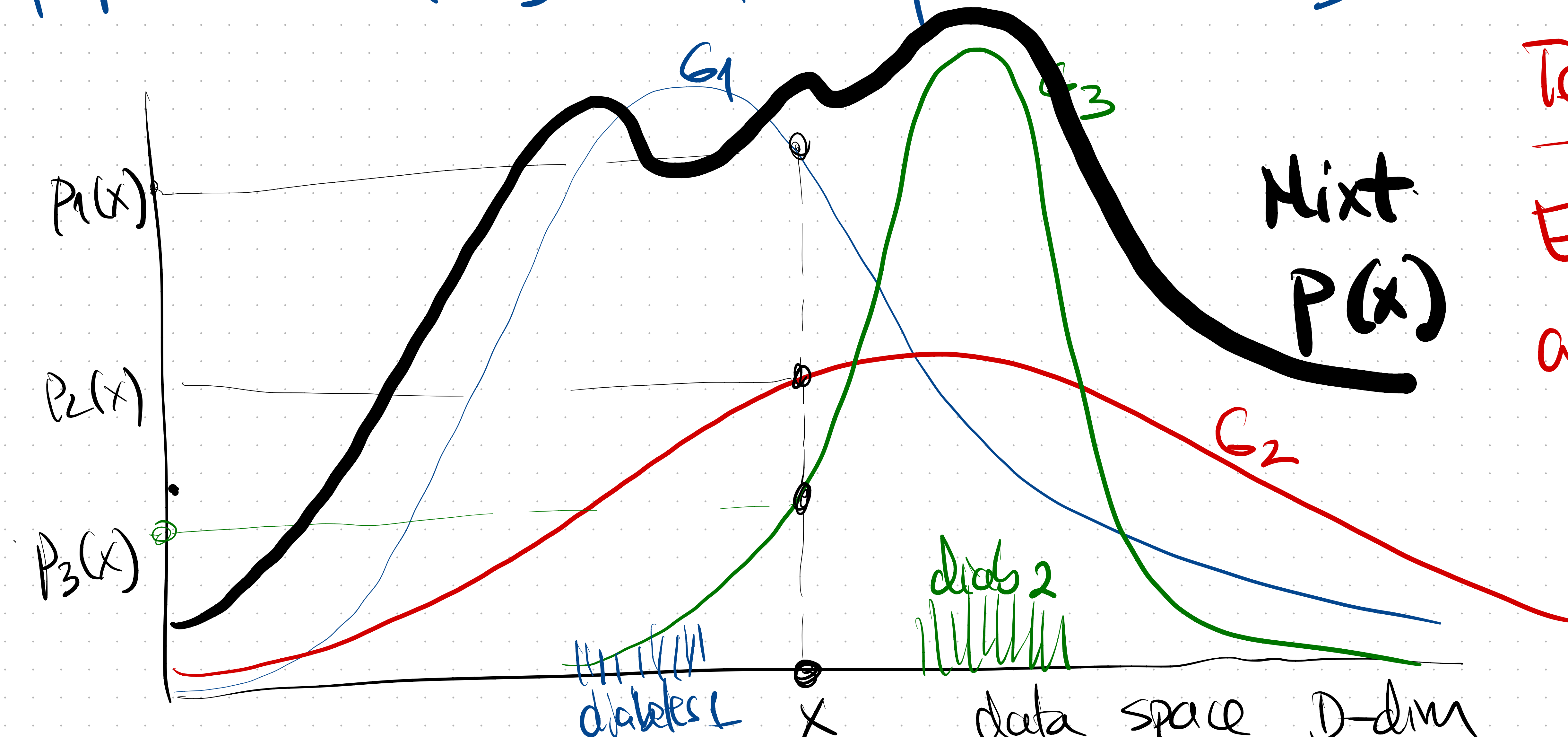
MIXT OF GAUSSIANS (GMM)

Today + Next: $P(x) = \underbrace{\boxed{w_1}}_{\text{fixed}} \mathcal{N}_1(x \mid \underbrace{\mu_1, \Sigma_1}_{\text{D-dim}}) + \underbrace{\boxed{w_2}}_{\text{fixed}} \mathcal{N}_2(x \mid \underbrace{\mu_2, \Sigma_2}_{\text{D-dim}}) + \underbrace{\boxed{w_3}}_{\text{fixed}} \mathcal{N}_3(x \mid \underbrace{\mu_3, \Sigma_3}_{\text{D-dim}})$

$w_1 + w_2 + w_3 = 1$

proportions

$K = 3$ # of "components" D = # data dim



Technique

E(xp) M(ax) algorithm

Mixture adv: ability to fit data with multiple "hills"

"bumps"

Recap for intuition: Kmeans clustering algorithm.

X D-dim data $\Longrightarrow$ cluster into $\cancel{K}$ 6 groups
$i = 1:N$ $\qquad k = 1:6$

- $\boxed{\pi_{ik} = \text{membership}}$
  $\begin{cases} 1 \text{ if } x_i \rightarrow \text{cluster } k \\ 0 \text{ if not} \end{cases}$
- μ_k = centroid for cluster k. $\boxed{\text{param}}$

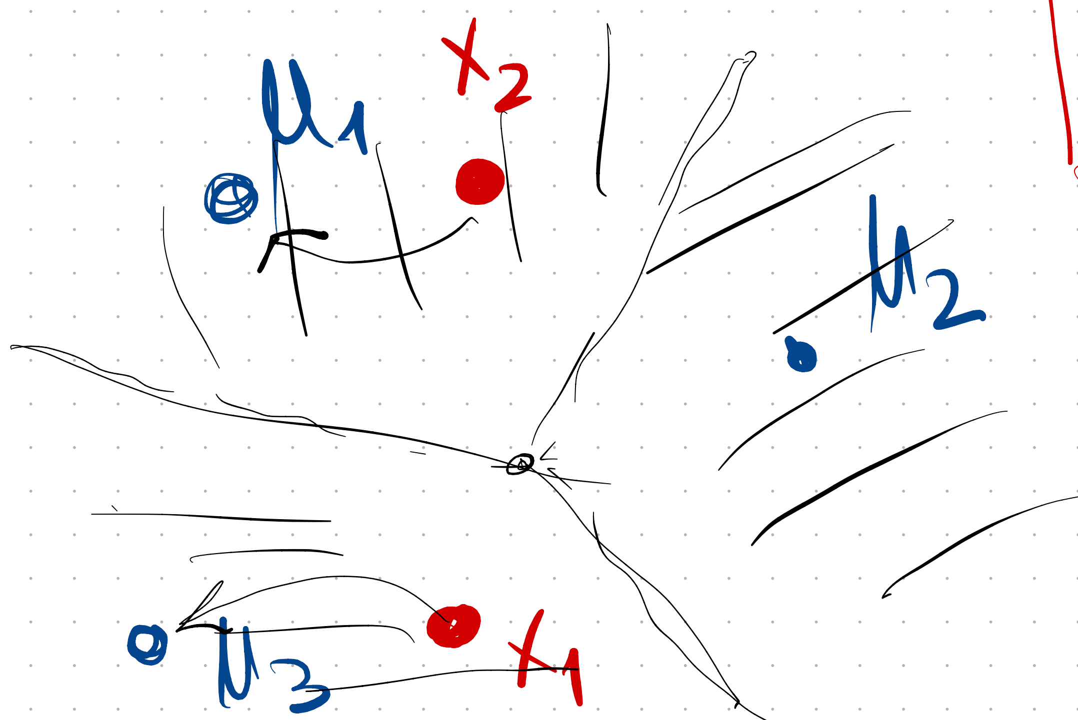
E step

calculate π_{ik}, given μ_k

$\pi_{ik} = 1$ for closest μ_k

$= \arg\min_k \|x_i - \mu_k\|^2$

HARD: π_{ik}



M-Step

calculate μ_k, given π_{ik}

μ_k = avg of (x points) in cluster k

$$= \frac{\sum_{i=1}^{N} x_i \cdot \boxed{\pi_{ik}}}{\boxed{\sum_{i=1}^{N} \pi_{ik}}}$$

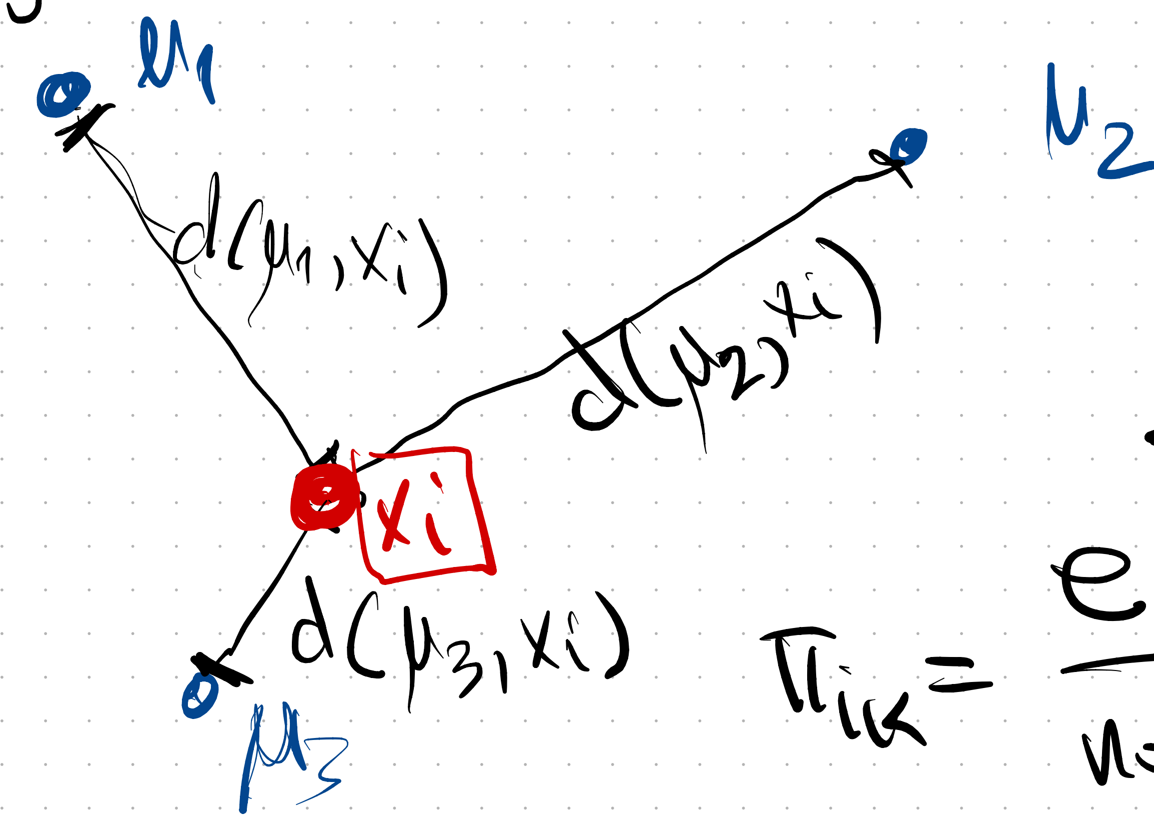
filter only points in cluster k

$\sum_{i=1}^{N} \pi_{ik} = N_k$ = # of points in cluster k

soFT : $\pi_{ik} = \text{prob}(x_i \leftrightarrow \mu_k)$

E step

- score
  dist
  reg output $\Big\} \Rightarrow$ probability



$$\pi_{ik} = \frac{e^{-1/\beta \overbrace{\|x_i - \mu_k\|^2}^{d(x_i,\, \mu_k)}}}{\text{normalized over all } k}$$

small dist $\|x_i - \mu_k\|^2 \equiv$ large π_{ik}

fix datapoint x_i

$[\pi_{i1} \quad \pi_{i2} \quad \cdots \quad \pi_{ik}]$ distribution over clusters

- parts of x_i distributed $\rightarrow x_i \cdot \pi_{ik}$
- which cluster/group x_i $\boxed{\text{is in}}$ $\Rightarrow$ which generator μ_k produced datap. x_i

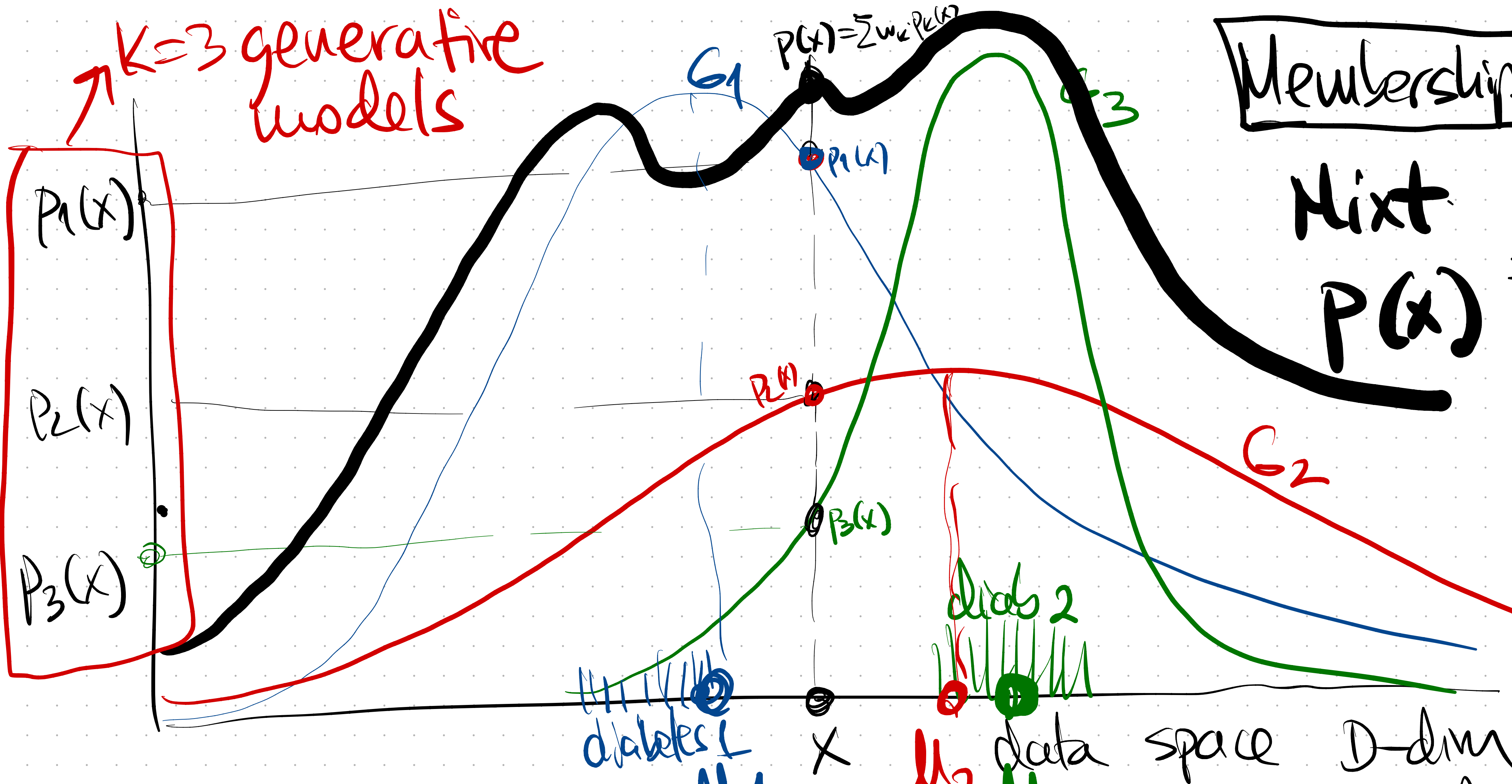
M step (same)

$$\mu_k = \frac{\sum_{i=1}^{N} x_i \cdot \boxed{\pi_{ik}}}{\boxed{\sum_{i=1}^{N} \pi_{ik}}}$$

proportion (weight)

$\rightarrow$ weighted avg

$N_k = E[\#\text{points}]$ in cluster k



Membership π_{ik} = prob that x_i was generated by Gauss k

Mixt $P(x)$

$= pr(k\ gen \mid x_i) = w_k \cdot pr(x_i \mid \boxed{\mu_k \Sigma_k})$

param$_k$

normalized over all$_k$

const: w_k = size of cluster/source k

E[#datap] from k $= \sum_{i=1}^{N} \pi_{ik}$

Mixture adv: ability to fit data with multiple "hills" "bumps"

$$\boxed{p(x)} = \sum_{k=1}^{3} \boxed{w_k} \cdot \boxed{N_k(x \mid \mu_k \Sigma_k)}$$

D-dim

prob to observe patient/email/car x_i overall sources $k=1, k=2, \ldots k=K$

π_{ik} = probab of source/generator K for $\boxed{\text{observed}}$ datap x_i (model the evidence)

w_k = prob of source/gen K in general (prior) → given x_i

E step

calculate π_{ik} given params $(\mu_k, \Sigma_{ik}, w_k)_{k=1:K}$

$$\pi_{ik} = pr(K \text{ source generated } x_i)$$

$$= pr(K \mid x_i) = \frac{Pr(x_i \mid K) \cdot pr(K)}{Pr(x_i)}$$

$$= \frac{\mathcal{N}_k(x_i \mid \mu_k, \Sigma_{ik}) \cdot w_k}{\text{normalize over } K \text{ generators}}$$

M step

calculate param $(\mu_k, \Sigma_{ik}, w_k)$ $k=1:K$

given memberships π_{ik}

- Log likelihood (data) = $LL(x)$
- take expectation w.r.t π_{ik}

  $E[LL(x)] \simeq LL(x)$
- take differentials

Constrain $w_1 + w_2 + w_3 = 1$

$\frac{\partial LL(x)}{\partial \mu_k} = 0$ | $\frac{\partial LL(x)}{\partial \Sigma_{ik}^{-1}} = 0$ | $\boxed{\text{Lagrangian}}$ $\frac{\partial LL(x)}{\partial w_k} \Rightarrow ?$