Lecture 6/6 Generative Models

HW3: GDA, Naive Bayes etc

- Fit curve to data
- ≈ Fit density to data

	1		D
	X		
i	x_i^1	x_i^2	x_i^D
N			

Want prob. density over data space

$pr(x_i | \text{param}_\theta)$ = prob that x_i datapoint occurs

$$pr_\theta(x_i) = \frac{\#\text{datapoints } x_i \text{ or very similar}}{\#\text{total} = N}$$

curve = parametric with model for density

ex: curve = Normal = Gaussian

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

Next Wed 6/11
Demo for HW2B NNets

Project Proposal June 24
Project page
Smaller Project ≈ HW (20h)
- choices

x_i = patient
$pr(x_i)$ = proportions of patients like this

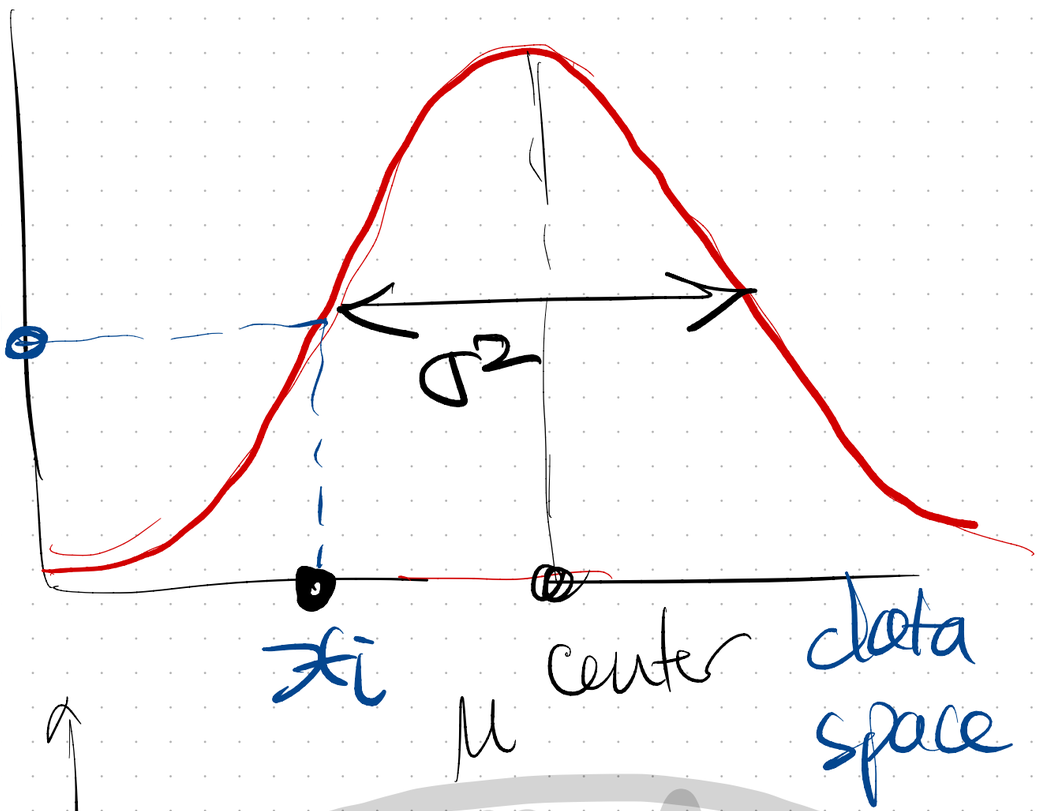
Fit curve = Gaussian to data X

(μ, σ^2)

⇓

Find the optimal **params μ, σ^2** to maximize Likelihood (data X)

$pr(x_i) = N(x_i)$



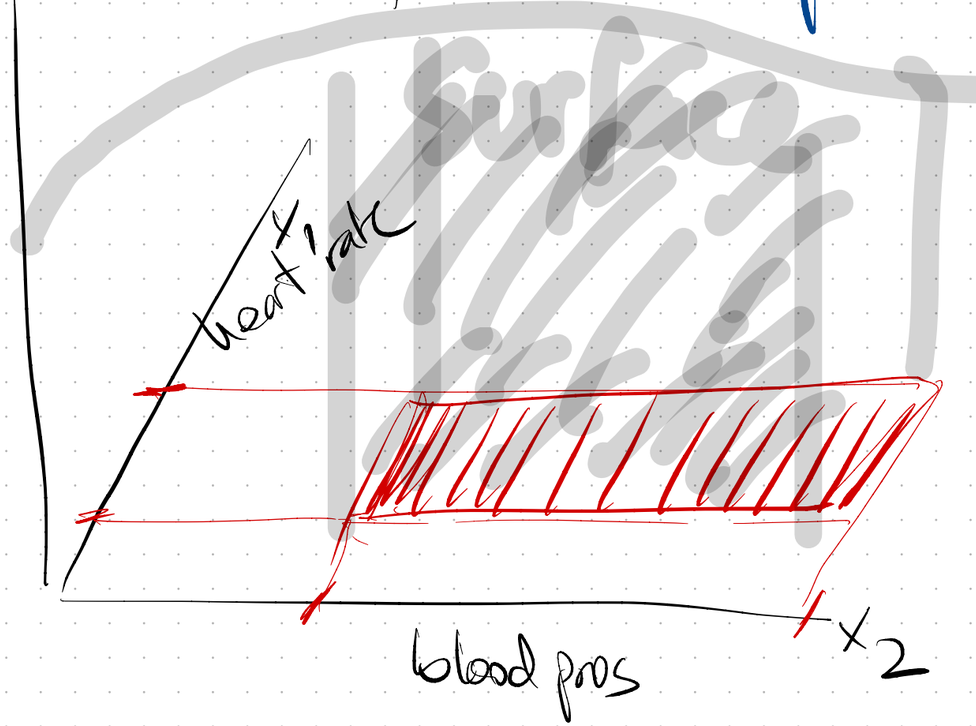
product ← because indep. datap

$$\prod_{i=1}^{N} pr(x_i)$$

Generative formulation:

Find curve = Gaussian (μ, Σ)

curve = generator of data ⇒ data X is most likely



1-dim Gaussian: $\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$

σ^2 = VAR

σ = std = $\sqrt{VAR}$

μ = data center most likely x

x_i = D-dim $\Rightarrow \mathcal{N}\left(\underset{D\text{-dim}}{x} \middle| \underset{D\text{-dim}}{\mu}, \Sigma\right)$

μ: D-dim vector

Σ: sigma (capital) DXD COVAR()

Σ = DXD COVAR. Matr

$\sum_{i=1}^{N}$ = summation

$$pr(x_i \mid \overset{\theta}{\mu, \Sigma}) = \frac{1}{(2\pi)^{-D/2}} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x_i-\mu)\Sigma^{-1}(x_i-\mu)^T\right\}$$

$|\Sigma|$: determinant (Σ)

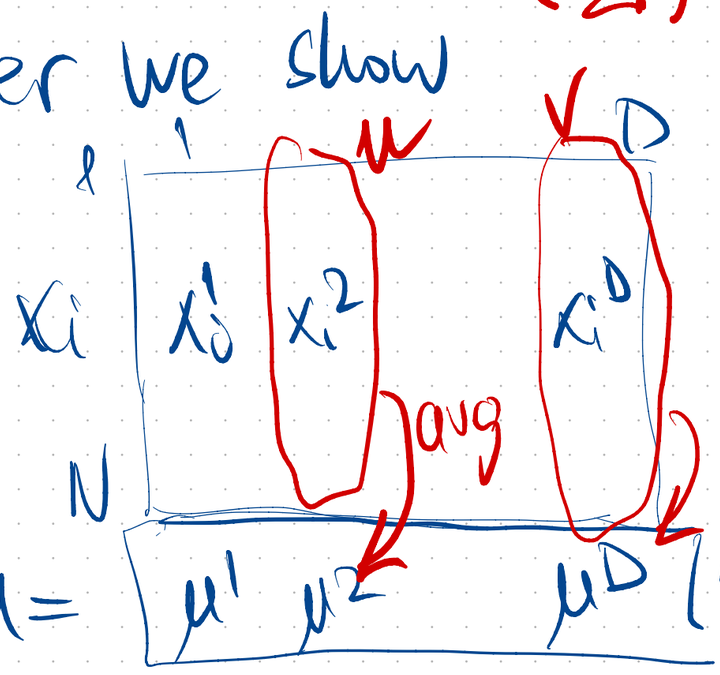
exp function

$(x_i-\mu)$: 1xD

Σ^{-1}: DXD, inverse of Σ

$(x_i-\mu)^T$: DX1

Later we show



μ = avg of each col = feature

if $\mu = 0$

$x^T \cdot x = \sum_{\text{COVAR}}$ DXD

$$= \begin{bmatrix} \langle col^1 \cdot col^1 \rangle & & \langle col^1 \cdot col^D \rangle \\ & & \\ \langle col^D \cdot col^1 \rangle & \cdots & \langle col^D \cdot col^D \rangle \end{bmatrix}$$

- Fit D-Dim Gaussian to data X (D-features)

$\Rightarrow \mu = \text{mean}(X)$ D-dim vector $(\mu^1 \mu^2 \ldots \mu^D)$ avg for each feat

$\Rightarrow \sum = \text{covar}(X) = (x-\mu)^T (x-\mu)$

$(\mu, \sum)$ params that makes generator Gaussian best for data X

when used to generate data $\Longrightarrow$ max-chance to obtain $X_{1:N}$

Data likelihood (as generated by Gaussian with param μ, Σ)

$$L(x) = \prod_{i=1}^{N} pr(x) = \prod_{i=1}^{N} \mathcal{N}(x_i | \mu, \Sigma)$$

all datap $i = 1 \ldots N$

$$\prod_{i=1}^{N} \left[(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (x_i - \mu) \Sigma^{-1} (x_i - \mu)^T \right\} \right]$$

$$\text{loglikelihood} = LL(x) = \log\left[\prod_{i=1}^{N} \mathcal{N}(x_i | \mu, \Sigma) \right]$$

$$= \sum_{i=1}^{N} \left[\log(2\pi)^{-d/2} + \frac{1}{2} \log|\Sigma^{-1}| - \frac{1}{2} (x_i - \mu) \Sigma^{-1} (x_i - \mu)^T \right]$$

want $\max LL(x) \Rightarrow \dfrac{\partial LL()}{\partial \mu}$, $\dfrac{\partial LL()}{\partial \Sigma^{-1}}$ μ, Σ are indep. (no constraints)

$$\frac{\partial \log L}{\partial \mu} = \sum_{i=1}^{N} \frac{1}{2} \Sigma^{-1} (x_i - \mu) \overset{\text{want}}{=\!=} 0 \Rightarrow \sum_{i=1}^{N} (x_i - \mu) = 0$$

$$\Rightarrow \boxed{\mu = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

$$\frac{\partial \log L}{\partial \Sigma^{-1}} = \sum_{i=1}^{N} \left[\frac{1}{2} \Sigma - \frac{1}{2} (x_i - \mu)^T (x_i - \mu) \right] \overset{\text{want}}{=\!=} 0$$

$$N \cdot \Sigma = \sum_{i=1}^{N} (x_i - \mu)^T (x_i - \mu)$$

$$\Sigma = \boxed{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^T (x_i - \mu)}$$

COVAR

$D \times 1 \quad 1 \times D$

SUM OF N

$D \times D$ mat

$$\frac{1}{N} \begin{bmatrix} \sum (x_i^1 - \mu^1)(x_i^1 - \mu^1) & \cdots & \sum (x_i^1 - \mu^1)(x_i^D - \mu^D) \\ \vdots & \ddots & \vdots \\ \sum (x_i^D - \mu^D)(x_i^1 - \mu^1) & \cdots & \sum (x_i^D - \mu^D)(x_i^D - \mu^D) \end{bmatrix}$$

$$\sum_{i=1}^{N} (x_i^1 - \mu^1)^2 \quad \cdots \quad \sum_{i=1}^{N} (x_i^1 - \mu^1)(x_i^D - \mu^D)$$

$$\vdots \qquad \ddots \qquad \vdots$$

$$\sum_{i=1}^{N} (x_i^D - \mu^D)(x_i^1 - \mu^1) \quad \cdots \quad \sum_{i=1}^{N} (x_i^D - \mu^D)^2$$

$$\mu = 0 \Leftrightarrow X_{centered} = X - \mu$$

$$\text{COVAR } \Sigma = \frac{1}{N} X^T X$$

Classification with gen models (Gaussian fit to data)

wout

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

- label
- datapoint
- $P(X|Y)$: density of x in certain class $y = y_e$
- $P(Y)$: prior for each y=class $Y_1, Y_2 \dots Y_L$ priordist over $Y_1 \dots Y_L$
- $P(X)$: normalization of numerators For L classes

---

y binary $\in \{1, 0\}$

x = test point

fit curve to y=1 points

fit curve to y=0 points

priors

$$P(y=1|X) = P(X|y=1) \cdot P(y=1) / P(X)$$

$$P(y=0|X) = P(X|y=0) \cdot P(y=0) / P(X)$$

ignore

GDA = Gauss Discriminant Analysis

- $\mathcal{N}_1(\mu_1, \Sigma_1)$ fit on X data subset with $y=+1$
- separate $\mathcal{N}_0(\mu_0, \Sigma_0)$ fit on X data subset $y=0$
- predict: compute test point z
  - $P(y=1|z) = \mathcal{N}_1(z|\mu_1,\Sigma_1) \cdot P(y=1) / P(z)$
  - $P(y=0|z) = \mathcal{N}_0(z|\mu_0,\Sigma_0) \cdot P(y=0) / P(z)$
  - predict $y = \text{argmax}$

ignore

- can do this for $y=1, y=2, \ldots y=L$ (L classes)
- can use common Σ for all classes.