Notes for CS 181: Probabilistic Regression

Last time: linear regression

$\hat{y} = w^T x$

Chose loss: least squares: $\underline{\mathcal{L}_D(w) = \frac{1}{2}\sum_n (y_n - w^T x_n)^2}$

▣ Interpretation 1: Projection View

$$\underbrace{Y^T}_{\substack{N\text{-dim}\\ \text{vect.}}} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = w^T X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ & & x_{DN} \end{bmatrix} \Big\} \begin{matrix} D \\ N\text{-dim} \\ \text{vecs} \end{matrix}$$

if $D < N$, then $x_1 \cdots x_D$ define some linear subspace in $\mathbb{R}^N$

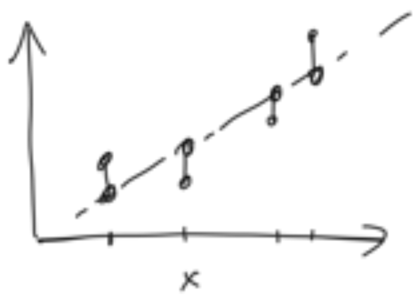
recall: compute projection: $\sum_d \underbrace{x_d}_{\substack{N\text{dim}\\ \text{vec.}}} \cdot \frac{\langle x_d, Y\rangle}{\langle x_d, x_d\rangle} \Rightarrow \underbrace{X^T(XX^T)^{-1}X}_{\substack{\text{projection}\\ \text{operator}}} Y$

note: linear algebra applies to random variables also.

▣ Probabilistic View: Idea: We're going to make a "story" for how our variables were created.

"generative model"

ex: $y_n = \underline{w^T x_n} + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$



we've made explicit:
- → form of noise (Gaussian)
- → indep of noise across samples

"graphical model"

$\underbrace{\Pr(\text{data} \mid \text{model})}_{\text{"likelihood" of model}} \doteq ?$

Quick Review of Gaussian distributions:

$\mathcal{N}(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(z-\mu)^2\}$

$= \prod_n \Pr(y_n \mid x_n, w, \sigma^2)$] because noise indep, y_n only depends on x_n, w, σ^2

First, let's take logs (monotonic):

$y_n \sim \mathcal{N}(w^T x_n, \sigma^2)$

$$\log \Pr(\text{data}|\text{model}) = \sum_n \log \Pr(y_n \mid x_n, w, \sigma^2)$$

$$= \sum_n \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - w^T x_n)^2\right\}$$

$$= \underbrace{N \log \frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{no dep on } w} + \underbrace{\sum_n \left(-\frac{1}{2\sigma^2}\right)(y_n - w^T x_n)^2}_{\text{if we min } \sum_n (y_n - w^T x_n)^2 \text{, we max pr(data|model)}}$$

for practice, we can do the same w/ matrices...

$P(Y \mid X, w, \sigma^2)$

$\Rightarrow$ multivariate Gaussian: $\underline{\mu} = w^T X$, $\Sigma = \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix} = I_N \sigma^2$

$Y \sim \mathcal{N}(w^T X, \Sigma)$

formula for multivariate Gaussian: $\frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{-\frac{1}{2}(z-\mu)^T \Sigma^{-1} (z-\mu)\right\}$

$$\underbrace{\log \frac{1}{\sqrt{2\pi|\Sigma|}}}_{-\frac{N}{2}\log 2\pi\sigma^2} + \underbrace{\left(-\frac{1}{2}\right)(Y - w^T X)\Sigma^{-1}(Y - w^T X)^T}_{-\frac{1}{2\sigma^2}(Y - w^T X)(Y - w^T X)^T}$$

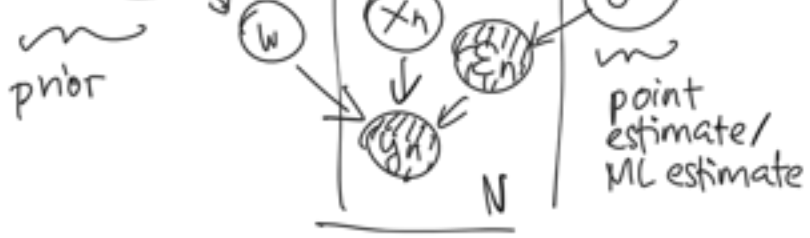
we solve (max of w) $\Rightarrow$ "maximum likelihood estimation"

we can do more: estimate $\underbrace{\sigma^2_{ML}}_{\text{max likelihood}}$

$$\frac{\partial}{\partial \sigma^2}: \quad -\frac{N}{2} \cdot \frac{1}{\sigma^2} - \frac{1}{2}(Y - w^T X)(Y - w^T X)^T \left(\frac{1}{\sigma^2}\right)^2 = 0$$

$$\sigma^2_{ML} = \frac{1}{N}(Y - w^T X)(Y - w^T X)^T$$ } empirical variance

Going further: we can <u>add</u> to story: what if w is a random variable?



write down as generative model:

$w \sim p(w)$
for n in $1 \ldots N$
$\varepsilon_n \sim \mathcal{N}(0, \sigma_n^2)$
$y_n = w^T x_n + \varepsilon_n$

and now, we can write joint prob: $pr(\{y_n\}, w \mid X, \sigma_n^2)$

$$= \underbrace{p(\{y_n\} \mid \{x_n\}, w, \sigma_n^2)}_{\text{likelihood from before}} \; \underbrace{p(w \mid X, \sigma_n^2)}_{p(w)} \quad = p(\{y_n\}, w \mid X, \sigma_n^2)$$
$$\propto p(w \mid X, \sigma_n^2, \{y_n\})$$

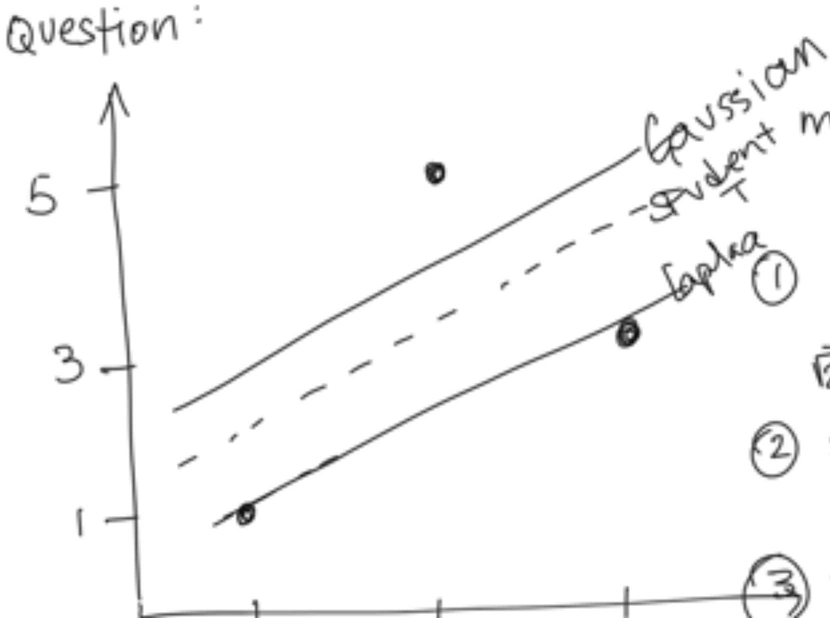
taking logs:
log JOINT = $\underbrace{\text{log LIKELIHOOD}}$ + log PRIOR   ex. $w_d \sim \mathcal{N}(0, \sigma_w^2)$

$$N \log \frac{1}{\sqrt{2\pi\sigma_n^2}} + \sum_n \left(-\frac{1}{2\sigma_n^2}\right)(y_n - w^T x_n)^2 + D \log \frac{1}{\sqrt{2\pi\sigma_w^2}} + \sum_d \left(\frac{-1}{2\sigma_w^2}\right)(w_d)^2$$

We can ask same kinds of questions:
what is the w that maximizes JOINT prob? (best σ_n^2, best σ_w^2)

before "maximum likelihood" solution (ML)
Now: "maximum aposteriori" solution (MAP)

Concept Question:



model: linear model
$y_n = w^T x_n + \varepsilon_n$

① $\varepsilon_n \sim$ Gaussian
$\frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2} \varepsilon^2\}$

② $\varepsilon_n \sim$ Laplace
$\propto \exp\{-\lambda |\varepsilon|\}$

③ $\varepsilon_n \sim$ Student T

1 2 3 $\propto (1+\varepsilon^2)^{-1}$

1. Convert all three options into log losses to minimize
2. Think about what effect diff losses will have on fitted ω?

Gaussian: $\varepsilon^2 \rightarrow \sum_n (y_n - \omega^T x_n)^2$

Laplace: $\rightarrow \sum_n |y_n - \omega^T x_n|$

Student T: $\rightarrow \sum_n \log(1 + (y_n - \omega^T x_n)^2)$