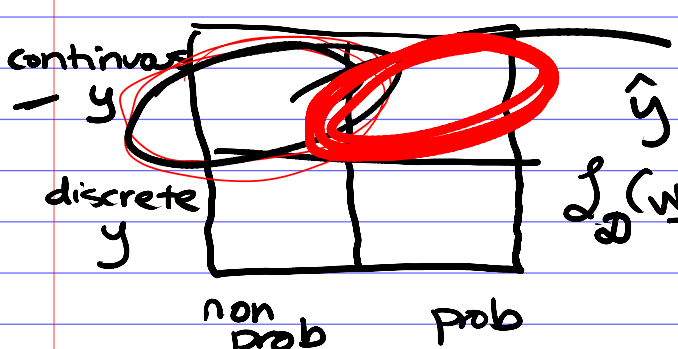
CS181: Probabilistic Regression

Housekeeping
- HW1: due Thurs
  T1_Q4.py
- Notes, Notation

Supervised Learning ($x \to y$)



Last time:

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

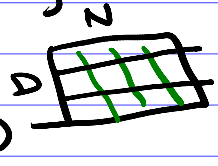
$\mathcal{L}_{\mathcal{D}}(\underline{w}) = \frac{1}{2}\sum_n (y_n - \hat{y}_n)^2$ (loss)

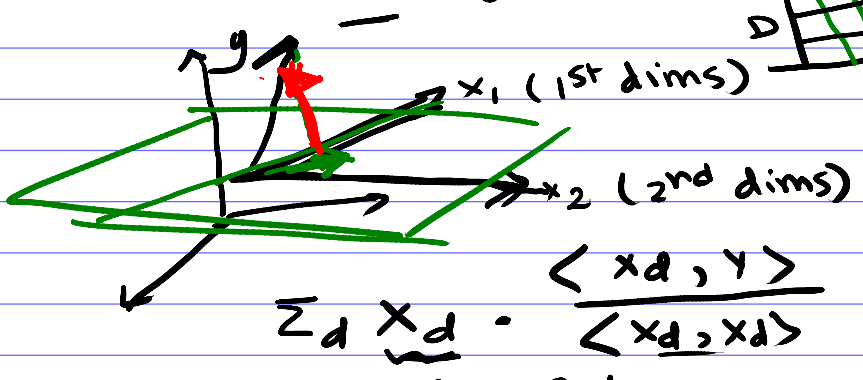
$\underline{w}^* = \underbrace{(X^TX)^{-1}}_{\text{looks like var}} \underbrace{(X^Ty)}_{\text{looks like cov}}$

This time:

1) provide geometric view of linear regression.

$\underset{N\text{-dim}}{\underline{y}}^T \approx \underset{D\text{ dim}}{\underline{w}^T} \underset{D \times N\text{ dim}}{\underline{X}^T}$ } D vectors of length N





$\sum_d \underset{N\text{-dim vector that is all elements of dth dim}}{\underline{x}_d} \cdot \frac{\langle x_d, y \rangle}{\langle x_d, x_d \rangle}$

Book 2.6.3

$X(X^TX)^{-1}X^Ty$ (X; our weight vector)

2) Probabilistic Methods.

Key Idea: Generative Model.
"Story for how the data came to be"

Story / Generative Model:

1. we some input $\underline{x}_n$
2. multiply $\underline{x}_n$ by $\underline{w}$: $\underline{w}^T\underline{x}_n$
3. output is the product + noise, $\underline{\varepsilon} \overset{iid}{\sim} \mathcal{N}(\underline{0}, \sigma^2)$

Book 2.6.2

$y_n = \underline{w}^T\underline{x}_n + \varepsilon_n$

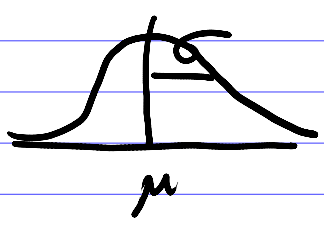
Probability of the data given w, σ^2 → likelihood of the model

$Pr(\text{data} \mid \text{model}) = \prod_n^N pr(y_n \mid \underline{x}_n, \underline{w}, \sigma^2)$

$\log pr(\text{data} \mid \text{model}) = \sum_n \log pr(y_n \mid \underline{x}_n, \underline{w}, \sigma^2)$

detour

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



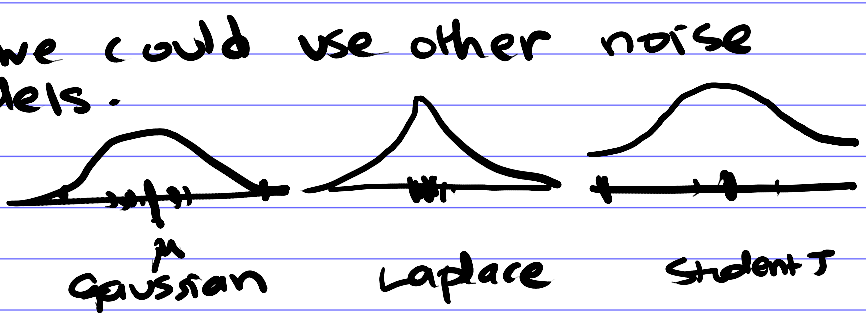
$\sum_n \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \underline{w}^T\underline{x}_n)^2\right\}$

$= \sum_n \underbrace{\log \frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{no dep. on } w} - \underbrace{\frac{1}{2\sigma^2}(y_n - \underline{w}^T\underline{x}_n)^2}_{\text{has dep on } w,\ \text{looks just like our least squares loss!!}}$

So now, if we maximize prob of data, we will get the same solution as least squares'.

- solution is known as the "maximum likelihood solution" (for this generative model)

Note: we could use other noise models.



Given this generative model, we can also find σ^2_{MLE}.

First, just for practice, let's write our probability of data w/ vectors

$\log \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{-\frac{1}{2}\underbrace{(\underline{y} - \underline{w}^T\underline{X}^T)}_{\text{vectors of size N}}\Sigma^{-1}(\underline{y} - \underline{w}^T\underline{X}^T)^T\right\}$

$\Sigma = \sigma^2 I$

$\begin{bmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \sigma^2 \\ & & & \ddots \end{bmatrix}$

$= -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\underline{y} - \underline{w}^T\underline{X}^T)(\underline{y} - \underline{w}^T\underline{X}^T)^T$

Take gradient w.r.t. σ^2 (note: grad w.r.t σ^2 not σ)

$\frac{\partial}{\partial\sigma^2}$ is: $-\frac{N}{2}\cdot\frac{1}{\sigma^2} + \frac{2}{2}(\underline{y} - \underline{w}^T\underline{X}^T)(\underline{y} - \underline{w}^T\underline{X}^T)^T\left(\frac{1}{\sigma^2}\right)^2$

$0 =$

Solve by multiplying through by σ^4:

$0 = -N\sigma^2 + (\underline{y} - \underline{w}^T\underline{X}^T)(\underline{y} - \underline{w}^T\underline{X}^T)^T$

$\hat{\sigma}^2_{MLE} = \frac{(y - w^TX^T)(y - w^TX^T)^T}{N}$

vec. of length N that contains diff btw true y and our predicted y → product: $\sum_n (err)^2$

$\hat{\sigma}^2_{MLE} = \frac{\sum_n (err)^2}{N}$

What we did: prob approach to regression 1) Generative model 2) Max likelihood of model w.r.t params.

Note: in a few lectures, we will return to this model, put priors over e.g. w

$w \sim p(w)$, then $y = \underbrace{w^Tx + \varepsilon}_{\text{today}}$