Announcements

- Homework 1 was due today.
- Homework 2 will be posted.
- Students on waiting list can register soon

Today's Lecture

- Notation (data; augmented and non-augmented vectors)
- Comment on definition of dataset D (y|x or y,x)
- Regression
  - Introduction
  - Based on MLE
  - Ridge regression (start)

Notation for augmented & unaugmented quantities

Non-augmented space	Augmented space
$\underline{w} = \underline{w}^{(0)} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$	$\underline{w} = \underline{w}^{(+)} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}$
$\underline{x} = \underline{x}^{(0)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$	$\underline{x} = \underline{x}^{(+)} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$
Linear $\hat{f}(\underline{x}) = w_0 + \underline{w}^T \underline{x}$	Linear $\hat{f}(\underline{x}) = \underline{w}^T \underline{x}$

(dropping superscripts)

Simarly for $\underline{\phi}^{(0)}(\underline{x})$, $\underline{\phi}^{(+)}(\underline{x})$, and $\underline{\phi}(\underline{x})$.

REGRESSION [MURPHY Ch. 7]

STOCK PRICE EX:

LET $d=2$ AND $\underline{w} = \underline{w}^{(t)}$ (also $\Rightarrow \underline{x} = \underline{x}^{(t)}$, $\underline{\phi} = \underline{\phi}^{(t)}$).

$$\hat{f}(\underline{x}) = w_0 + w_1 x + w_2 x^2$$

$$\text{LET } \underline{\phi}(x) = \begin{bmatrix} \phi_0(x) \\ \phi_1(x) \\ \phi_2(x) \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}$$

$$\hat{f}(x) = w_0 \phi_0(x) + w_1 \phi_1(x) + w_2 \phi_2(x)$$
$$= \underline{w}^T \underline{\phi}(x)$$

"BASIS ~~SET~~ FUNCTION EXPANSION"

"NONLINEAR MAPPING"

"ϕ MACHINE"

EX2: RENT OF APARTMENTS NEAR BEACH

LET $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \text{LIVING AREA} \\ \text{NO. OF ROOMS} \end{bmatrix}$

$\underline{w} = \underline{w}^{(+)}$ THROUGHOUT.

LINEAR MODEL:

$$\hat{f}(\underline{x}) = w_0 + w_1 x_1 + w_2 x_2 = \underline{w}^T \underline{x}$$

NONLINEAR MODEL: (NONLINEAR IN $\underline{x}$).

$$\hat{f}(\underline{x}) = \underline{w}^T \underline{\phi}(\underline{x})$$

EX: QUADRATIC: $\hat{f}(\underline{x}) = \underline{x}^T \underline{\underline{W}}\, \underline{x}$,

$\underline{\underline{W}}$ IS UPPER TRIANGULAR.

5

KEY ISSUE 1: HYPOTHESIS SET, OR MODEL [M 7.2]

MODEL IS STATISTICAL.

$\underline{w} = \underline{w}^{(+)}$ THROUGHOUT

MODEL y AS:

$$p(y \mid \underline{x}, \underline{\theta})$$

$\uparrow$ UNKNOWN PARAMETERS. TO BE ESTIMATED FROM $\mathcal{D}$.

→ MAKE <u>ASSUMPTION</u>

(i) HERE: $p(y \mid \underline{x}, \underline{\theta}) = N(y \mid \underline{w}^T \underline{x}, \sigma^2)$ [LINEAR IN $\underline{x}$]

(i)' OR $= N(y \mid \underline{w}^T \underline{\phi}(\underline{x}), \sigma^2)$ [NONLINEAR IN $\underline{x}$]

EQUIVALENT TO:

$$y(\underline{x}) = \underline{w}^T \underline{x} + n,$$

$$n \sim N(\delta y \mid 0, \sigma^2).$$

OR $\underline{\phi}(\underline{x})$

Regression using Maximum Likelihood Estimate (MLE)

$\underline{w} = \underline{w}^{(+)}$ throughout

[M 7.3]

$p(\mathcal{D}|\underline{\theta})$ = likelihood of $\underline{\theta}$.

Est. $\underline{\theta}$ using MLE:

$$\hat{\underline{\theta}}_{MLE} = \underset{\underline{\theta}}{\operatorname{argmax}} \ln p(\mathcal{D}|\underline{\theta})$$

If we assume (i) or (i)′, with σ, $\underline{\phi}$ known,

Then: $\underline{\theta} = \underline{w}$

And: $$\hat{\underline{w}}_{MLE} = \underset{\underline{w}}{\operatorname{argmax}} \underbrace{\ln p(\mathcal{D}|\underline{w})}_{\sum_{i=1}^{N} \ln p(y_i|\underline{x}_i, \underline{w})}$$

assuming points in $\mathcal{D}$ are independently distributed (i.d.).

2. OBJECTIVE FUNCTION

$$J(\underline{w}, \mathcal{D}) = -\ln p(\mathcal{D} \mid \underline{w}) = NLL(\underline{w}).$$

$$\left(\text{or } -p(\mathcal{D} \mid \underline{w})\right)$$

LET:

$$\underbrace{\underline{\underline{X}} = \begin{bmatrix} - & \underline{x}_1^T & - \\ - & \underline{x}_2^T & - \\ & \vdots & \\ - & \underline{x}_N^T & - \end{bmatrix} \text{ and } \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}$$

COMPRISE $\mathcal{D}$.

EE 660 **Notes on** $p\left(\mathcal{D}\middle|\underline{\theta}\right),\ p\left(\underline{y},\underline{\underline{X}}\middle|\underline{\theta}\right),\ p\left(\underline{y}\middle|\underline{\underline{X}},\underline{\theta}\right)$ **Fall 2018**

Using probability relations we have:

$$p\left(y,\underline{x}\middle|\underline{\theta}\right)=p\left(y\middle|\underline{x},\underline{\theta}\right)\,p\left(\underline{x}\middle|\underline{\theta}\right)=p\left(y\middle|\underline{x},\underline{\theta}\right)\,p\left(\underline{x}\right)$$

where for the last step we have dropped the last condition on $\underline{\theta}$ because it tells us nothing useful about $p\left(\underline{x}\right)$. If we are interested in maximizing (or minimizing) the likelihood, we will take:

$$\begin{aligned}
\underset{\underline{\theta}}{\arg\max}\ p\left(\mathcal{D}\middle|\underline{\theta}\right) &= \underset{\underline{\theta}}{\arg\max}\left\{\prod_{i=1}^{N} p\left(y_i,\underline{x}_i\middle|\underline{\theta}\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)p\left(\underline{x}_i\right)\right\} = \underset{\underline{\theta}}{\arg\max}\left\{\left(\cancel{\prod_{i=1}^{N} p\left(\underline{x}_i\right)}\right)\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)\right\}
\end{aligned}$$

and to obtain the last line, $\prod_{i=1}^{N} p\left(\underline{x}_i\right)$ was dropped because it is a positive multiplicative term that is a constant of $\underline{\theta}$. This can equivalently be seen by using the log likelihood instead:

$$\begin{aligned}
\underset{\underline{\theta}}{\arg\max}\ p\left(\mathcal{D}\middle|\underline{\theta}\right) &= \underset{\underline{\theta}}{\arg\max}\left\{\ln\prod_{i=1}^{N} p\left(y_i,\underline{x}_i\middle|\underline{\theta}\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\ln\prod_{i=1}^{N} p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)p\left(\underline{x}_i\right)\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\sum_{i=1}^{N} \ln\left[p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)p\left(\underline{x}_i\right)\right]\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\sum_{i=1}^{N} \left[\ln p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)+\cancel{\ln p\left(\underline{x}_i\right)}\right]\right\} \\
&= \underset{\underline{\theta}}{\arg\max}\left\{\sum_{i=1}^{N} \ln p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)\right\}
\end{aligned}$$

and to obtain the last line, the additive terms that don't depend on $\underline{\theta}$ have been dropped.

So, when the goal of using the likelihood is to find its argmax or argmin w.r.t. $\underline{\theta}$, we can replace $p\left(y_i,\underline{x}_i\middle|\underline{\theta}\right)$ directly with $p\left(y_i\middle|\underline{x}_i,\underline{\theta}\right)$.

Using $p(\mathcal{D}|\underline{\theta}) = \prod_{i=1}^{N} p(y_i | \underline{x}_i, \underline{\theta})$

$$= \prod_{i=1}^{N} N(y_i | \underline{w}^T \underline{x}_i, \sigma^2)$$

plug in

$$\Rightarrow J_1(\underline{w}) = \frac{1}{\underbrace{2\sigma^2}_{\text{Drop}}} \sum_{i=1}^{N} (y_i - \underline{w}^T \underline{x}_i)^2 + \underbrace{\frac{N}{2} \log(2\pi\sigma^2)}_{\text{const. of } \underline{w}}$$

Let $J(\underline{w}) = \frac{1}{2} RSS(\underline{w})$

$$= \frac{1}{2} \sum_{i=1}^{N} (y_i - \underline{w}^T \underline{x}_i)^2$$

$$= \frac{1}{2} \| \underline{y} - \underline{\underline{X}} \, \underline{w} \|_2^2$$

$$= \frac{1}{2} (\underline{y} - \underline{\underline{X}} \, \underline{w})^T (\underline{y} - \underline{\underline{X}} \, \underline{w})$$

3. OPTIMIZATION METHOD

$\nabla_{\underline{w}} J(\underline{w}, \mathcal{D}) = \underline{0}$ & SOLVE ALGEBRAICALLY.

SOLVING GIVES $\hat{\underline{w}}$:

$$\underline{\underline{X}}^T \underline{\underline{X}} \, \hat{\underline{w}} = \underline{\underline{X}}^T \underline{y}$$

IF $(\underline{\underline{X}}^T \underline{\underline{X}})$ IS INVERTABLE, THEN

$$\boxed{\hat{\underline{w}}_{OLS} = \underline{\underline{X}}^{-} \underline{y} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{y}}$$

= ~~ORID~~ ORDINARY LEAST SQUARES SOLUTION.

$\underline{\underline{X}}^{-}$ = PSEUDOINVERSE OF $\underline{\underline{X}}$.

$\hat{\underline{w}}_{OLS}$ FOLLOWS FROM THE NORMAL DENSITY ASSUMPTION (AND i.d.)

Ridge Regression

$(\underline{w} = \underline{w}^{(0)}$ THROUGHOUT$)$

→ USE MAP ESTIMATE

$$\hat{\underline{\theta}}_{MAP} = \arg\max_{\underline{\theta}} \, p(\underline{\theta}|\mathcal{D})$$

$$= \arg\max_{\underline{\theta}} \left\{ \frac{p(\mathcal{D}|\underline{\theta})\, p(\underline{\theta})}{p(\mathcal{D})} \right\}$$

DROP.

$$= \arg\max \left\{ p(\mathcal{D}|\underline{\theta})\, p(\underline{\theta}) \right\}$$

$$(1) \quad \hat{\underline{\theta}}_{MAP} = \arg\max_{\underline{\theta}} \left\{ \underbrace{\ln p(\mathcal{D}|\underline{\theta})}_{\text{likelihood of } \underline{\theta}} + \underbrace{\ln p(\underline{\theta})}_{\text{prior}} \right\}$$

ASSUME $(\underline{x}_i, y_i)$ ARE i.d., FROM $\mathcal{N}(y \,|\, w_0 + \underline{w}^T \underline{x}, \sigma^2)$

$$(2) \quad \text{so: } \ln p(\mathcal{D}|\underline{\theta}) = \ln \prod_i \mathcal{N}(y_i \,|\, w_0 + \underline{w}^T \underline{x}_i, \sigma^2)$$

→ MLE TERM

FOR PRIOR, CHOOSE:

$$p(\underline{\theta}) = p(\underline{w}) = \prod_{j=1}^{D} N(w_j \mid 0, \tau^2)$$

$$(3) \quad \ln p(\underline{\theta}) = \sum_{j=1}^{D} \ln N(w_j \mid 0, \tau^2)$$

WHY A <u>GAUSSIAN</u> PRIOR?

1. ALGEBRAICALLY CONVENIENT
2. CAN CHOOSE τ (NARROW OR WIDE GAUSSIAN)
3. WILL PREFER w_j VALUES CLOSE TO 0.

BUT NOT NECESSARILY THE BEST CHOICE.