Lecture 4 **EE 660** Aug 30, 2018

Announcements

- Homework 2 was posted.

Today's Lecture

- Ridge regression (finish)
- Notation comment
- Bayesian inference

NOTE: A COUPLE LINES HAVE BEEN ADDED TO PG. 8 FOR CLARIFICATION.

Ridge Regression, Part 2

Subs. (2) & (3) $\rightarrow$ (1):

$$\hat{\underline{w}} = \hat{\underline{\theta}} = \underset{\underline{w}}{\arg\max} \left\{ \overbrace{\sum_{i=1}^{N} \ln N(y_i \mid w_0 + \underline{w}^T \underline{x}_i, \sigma^2)}^{1^{st}\text{ term}} + \underbrace{\sum_{j=1}^{D} \ln N(w_j \mid 0, \tau^2)}_{2^{nd}\text{ term}} \right\}$$

$$\hat{\underline{w}} = \underset{\underline{w}}{\arg\max} \left\{ -\tilde{J}_R(\underline{w}, \mathcal{D}) \right\}$$

$\Rightarrow \tilde{J}_R$ IS OBJECTIVE FCN. (IN UNSIMPLIFIED FORM)

1^{st} term of $\tilde{J}_R(\underline{w}, \mathcal{D})$:

$$-\sum_{i=1}^{N} \ln N(y_i \mid w_0 + \underline{w}^T \underline{x}_i, \sigma^2)$$

$$(4) \quad = \cancel{-\sum_{i=1}^{N} (\text{const. of } \underline{w})} + \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - (w_0 + \underline{w}^T \underline{x}_i)\right)^2$$

2^{nd} term of $\tilde{J}_R$:

$$-\sum_{j=1}^{D} \ln N(w_j \mid 0, \tau^2)$$

$$= -\sum_j \ln \left[\frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{ -\frac{w_j^2}{2\tau^2} \right\} \right]$$

$$= \cancel{+ D \cdot \tfrac{1}{2} \ln(2\pi\tau^2)} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}$$

$$(5) \quad = \frac{1}{2\tau^2} \left\| \underline{w} \right\|_2^2$$

$2\sigma^2 \left[(4) + (5) \right] \Rightarrow$

CORRECT VERSION OF EQ. (7.32)

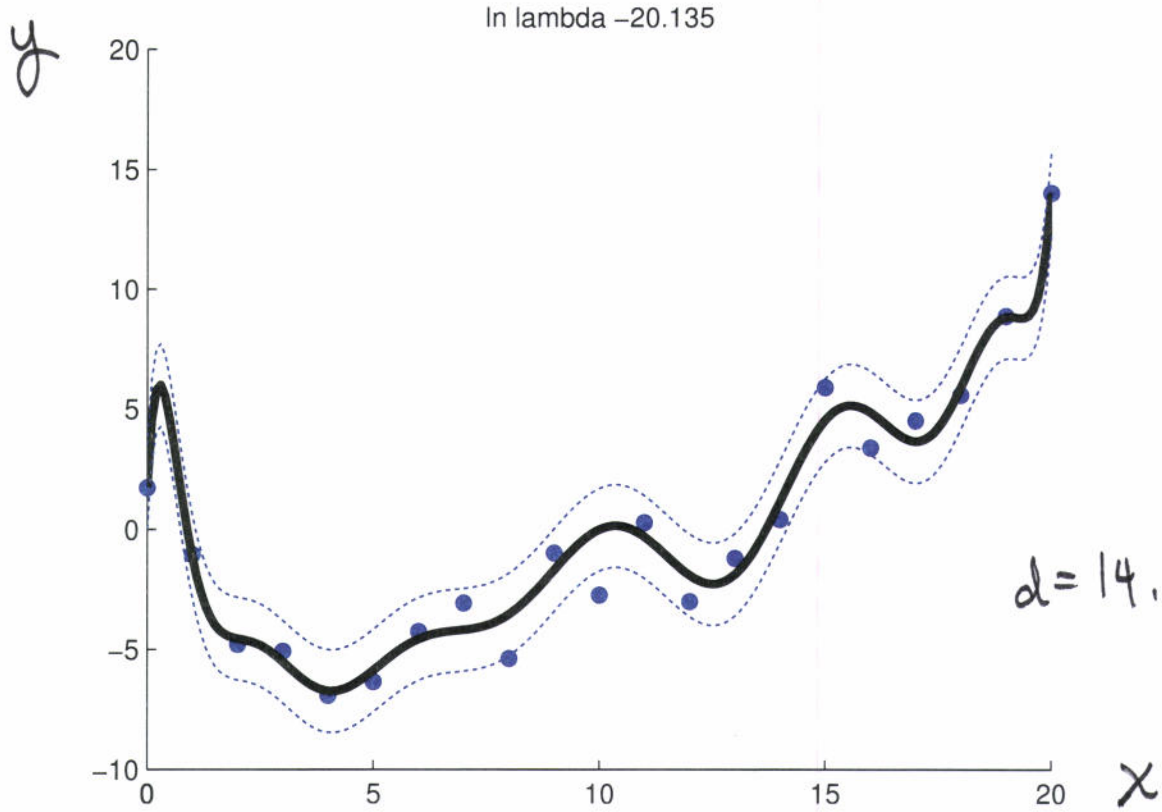
$$\boxed{\tilde{J}_R(w, \mathcal{D}) = \underbrace{\sum_{i=1}^{N} \Big[\underset{\substack{\uparrow \\ \text{TRAINING SAMPLE OUTPUT.}}}{y_i} - \underbrace{(w_0 + \underline{w}^T \underline{x}_i)}_{\hat{f}(x_i)} \Big]^2}_{J_{MLE}} + \underbrace{\underbrace{\lambda \left\| \underline{w} \right\|_2^2}_{\text{REGULARIZER}}}_{\substack{\text{new} \\ \text{(prior term)}}} \qquad \lambda = \frac{\sigma^2}{\tau^2}}$$

RIDGE REGRESSION OBJECTIVE FCN.

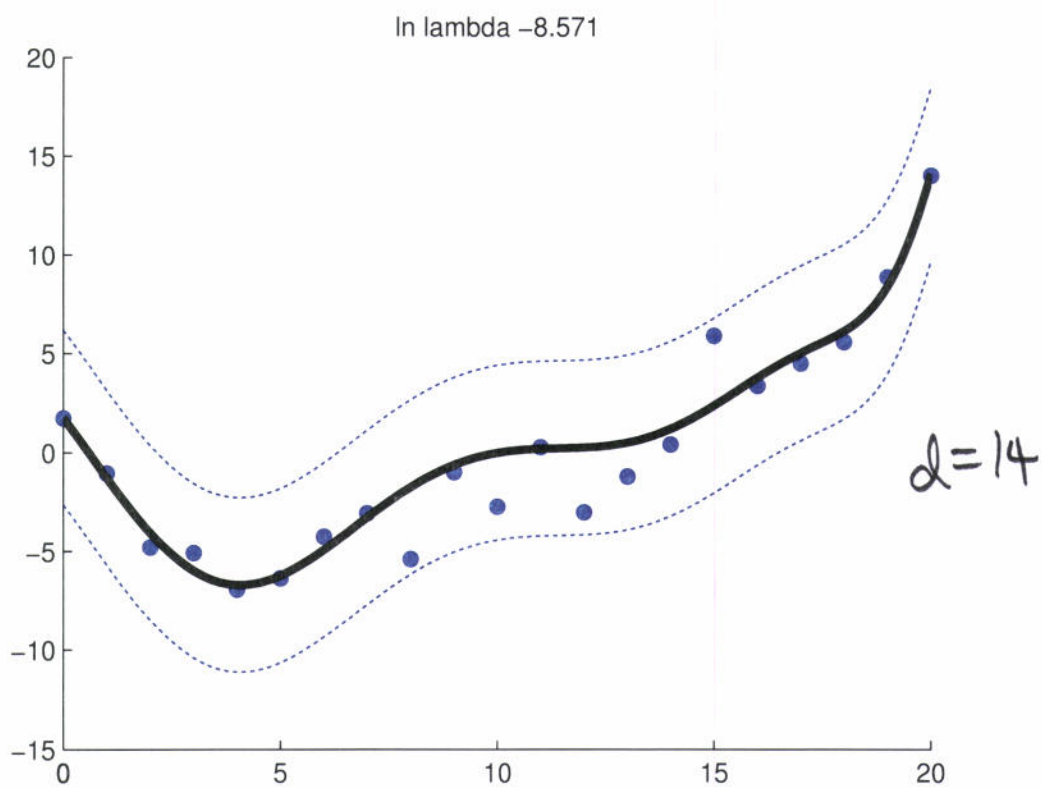
ASSUMED: NORMAL DENSITIES $p(y \mid \underline{x})$, $p(\underline{w})$

& INDEP. OF w_j.

POINTS IN $\mathcal{D}$ ARE i.d.

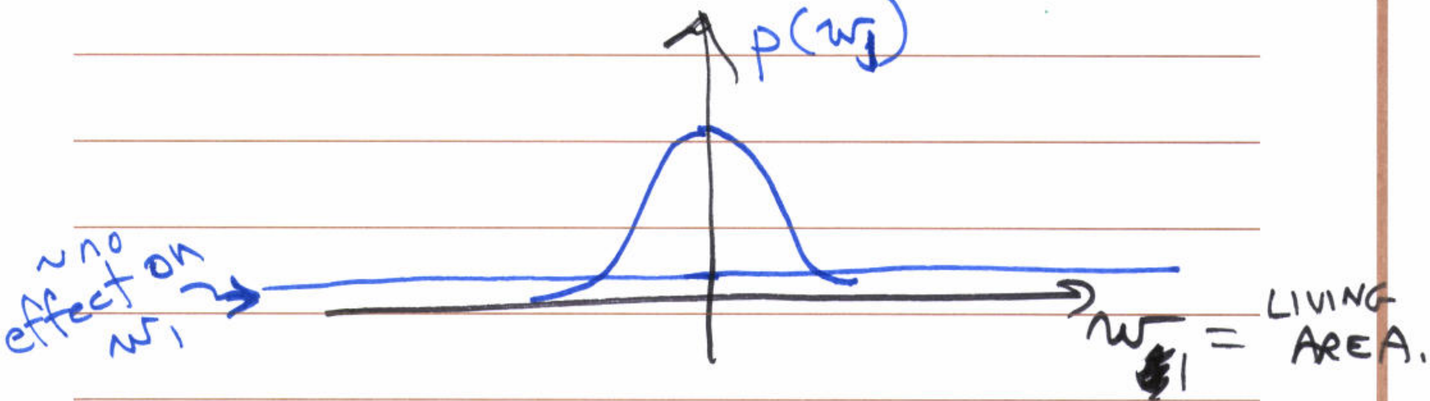


Murphy Fig. 7.7 (a)-(b). $N = 21$ data points, fit using regression function that is polynomial of degree 14, and differing amounts of L2 regularization.



$$\lambda = \frac{\text{VARIANCE OF DATA (FROM MODEL)}}{\text{VARIANCE OF PRIOR.}}$$

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



[USING MORE DATA CAN HAVE A REGULARIZING-TYPE EFFECT — MURPHY FIG. 7.10]

NOTATION COMMENT

$\underline{x}$ = GENERAL INPUT VARIABLE.

y = " OUTPUT " (VALUE OR CLASS)

WHICH ARE DIFFERENT THAN:

VALUE OF DATAPOINTS IN $\mathcal{D}$.

$\underline{\underline{X}}$ = INPUT VALUES OF DATASET $\mathcal{D}$ (TRAINING DATA)

$$= \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}$$

~~X_i = INPUT VALUES OF~~

y_i = OUTPUT VALUE FOR INPUT $\underline{x}_i$ OF $\mathcal{D}$.

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

x_i = $i^{\underline{th}}$ COMPONENT OF FEATURE VECTOR $\underline{x}$.

BAYESIAN INFERENCE

WE WANT TO ESTIMATE $\underline{\theta}$

INSTEAD OF FINDING A POINT ESTIMATE $\hat{\underline{\theta}}$,

LET'S ESTIMATE ~~THE~~ DENSITY:

$$p(\underline{\theta} \mid \mathcal{D}).$$

WE HAVE A MODEL:

(i) $p(y \mid \underline{x}, \underline{\theta})$ [REGRESSION]

(ii) OR $p(\underline{x} \mid \underline{y}, \underline{\theta})$ [CLASSIFICATION].

EE559: $p(\underline{x} \mid S_i) \leftarrow$ ASSUME A MODEL.

(i) DISCRIMINATIVE APPROACH

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

(ii) GENERATIVE APPROACH.

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

NOTE: MODELING $p(\underline{x} \mid y=c, \underline{\theta})$ IN CLASSIFICATION, WE CAN:

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

AND:

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

(b) OR $= p(\underline{x} \mid y=c, \underline{\theta})\, p(y=c)$

FOR (a), WE CAN USE $p(\underline{x} \mid \underline{\theta}) = \sum_{c=1}^{C} p(\underline{x} \mid y=c, \underline{\theta})\, p(y \mid \underline{\theta})$

$$\Longrightarrow p(\underline{x}) = \sum_{c=1}^{C} p(\underline{x} \mid y=c, \underline{\theta})\, p(y).$$