Announcements

- Homework 2 is due tomorrow
- Homework 3 will be posted

Today's Lecture

- Bayesian inference (regression)
- Summary of estimation techniques

— These 2 pages are from Lecture 4, pp. 7–8. —

参考2

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

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

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

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

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

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

BAYESIAN INFERENCE (part 2)

$\underline{w} = \underline{w}^{(t)}$. [not always true in Murphy 7.6].

GIVEN OUR MODEL, WE CAN COMPUTE THE LIKELIHOOD:

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

USE BAYES THEOREM:

(1)
$$\boxed{p(\underline{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \underline{\theta})\, p(\underline{\theta})}{p(\mathcal{D})}}$$

IN WHICH: $p(\mathcal{D}) = \int p(\mathcal{D} \mid \underline{\theta})\, p(\underline{\theta})\, d\underline{\theta}$

(OR SUM IF $\underline{\theta}$ IS DISCRETE)

IN Ch.3 READING, $\underline{\theta} = h =$ HYPOTHESIS.

IN REGRESSION, $\underline{\theta} = \underline{w}$ (AND MAYBE σ^2).

EXAMPLES OF (1):

- NUMBERS GAME (HW2, M Ch.3)

$$h \in \{h_2, h_4\}$$

$$p(h \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid h)\, p(h)}{\underbrace{p(\mathcal{D})}}$$

LIKELIHOOD (use strong sampling assumpt.)

PRIOR.

$$p(\mathcal{D}) = \sum_h p(\mathcal{D} \mid h)\, p(h)$$

- REGRESSION (APT. RENT EX.)

  $\underline{\theta}$ = COEFFICIENTS (WEIGHTS) ON INPUTS (LIVING AREA, #ROOMS, ETC.)

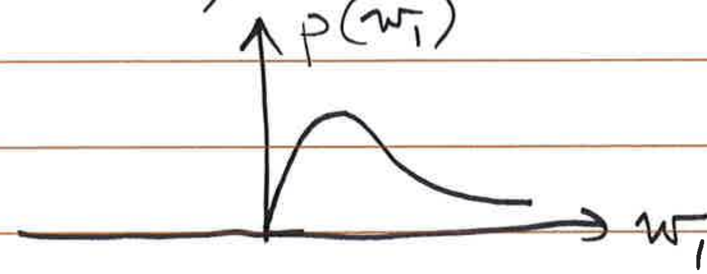
$|\theta_i|$ = IMPORTANCE ~~OF~~ OF FEATURE x_i. <u>IF</u> ALL ~~[illegible]~~ x_j'S ARE NORMALIZED TO SAME RANGE, (e.g., STANDARIZED TO $\mu_j = 0$, $\sigma_j^2 = 1$), $j = 1, \dots, D$.

$p(\underline{\theta} \mid \mathcal{D})$ = DENS. OF w_j COEFF. OF x_j

$$= \left[\prod_{i=1}^{N} \underbrace{p(y_i \mid \underline{x}_i, \underline{\theta})} \right] p(\underline{\theta}) / K$$

USE OUR MODEL, E.G.: $N(y_i \mid \underline{w}^T x_i, \sigma^2)$.

$p(\underline{\theta})$ = PRIOR ON $\underline{\theta}$, E.G. MAYBE:



THEN GET POSTERIOR PREDICTIVE

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

FROM: $p(y) = \int p(y \mid \underline{\theta})\, p(\underline{\theta})\, d\underline{\theta}$

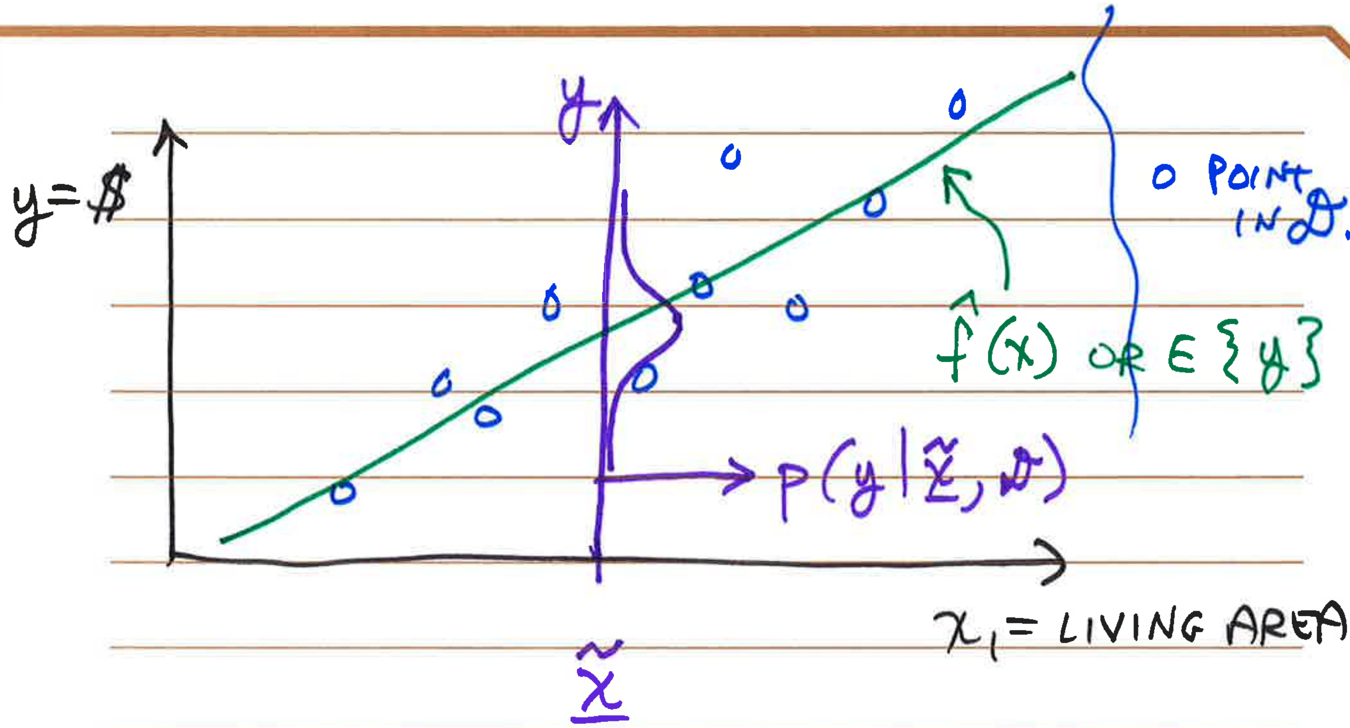
WE GET:

(2)
$$\boxed{\begin{aligned} p(y \mid \underline{x}, \mathcal{D}) &= \int p(y \mid \underline{x}, \underline{\theta}, \mathcal{D})\, p(\underline{\theta} \mid \underline{x}, \mathcal{D})\, d\underline{\theta} \\ \text{USUALLY: } &= \int \underbrace{p(y \mid \underline{x}, \underline{\theta})}_{\text{OUR MODEL}}\, p(\underline{\theta} \mid \mathcal{D})\, d\underline{\theta} \end{aligned}}$$

OUR MODEL
(REGR,
DISCRIMINATIVE)

$p(\underline{w} / \mathcal{D})$
– GET FROM (1).

OR CAN BE OBTAINED FROM
$p(\underline{x} / y, \underline{\theta})$ (CLASS'N, GENERATIVE)



BAYESIAN REGRESSION

1. MODEL IS $p(y|\underline{x}, \underline{\theta}) = p(y|\underline{x}, \underline{w}, \underline{\theta}')$

any other unknowns.

EX: $= p(y|\underline{w}^T\underline{x})$ (LINEAR CASE)

$= p(y|\underline{w}^T\underline{\phi}(x))$

2. PARAMETER POSTERIOR

FROM EQ. (1):

$$p(\underline{\theta} \mid \mathcal{D}) = p(\underline{w} \mid \underline{y}, \underline{\underline{X}})$$

likelihood

prior

$$= \frac{p(\underline{y} \mid \underline{w}, \underline{\underline{X}})\, p(\underline{w} \mid \underline{\underline{X}})}{K}$$

$$K = \int p(\underline{y} \mid \underline{w}, \underline{\underline{X}})\, p(\underline{w} \mid \underline{\underline{X}})\, d\underline{w}$$

3. POSTERIOR PREDICTIVE

FROM EQ. (2):

$$p(y \mid \underline{x}, \mathcal{D}) = \int p(y \mid \underline{x}, \underline{w}, \cancel{\mathcal{D}})\, p(\underline{w} \mid \cancel{\underline{x}}, \mathcal{D})\, d\underline{w}$$

$$= \int p(y \mid \underline{x}, \underline{w})\, p(\underline{w} \mid \mathcal{D})\, d\underline{w}.$$

$p(\mathcal{D} \mid \underline{w})$

no data.

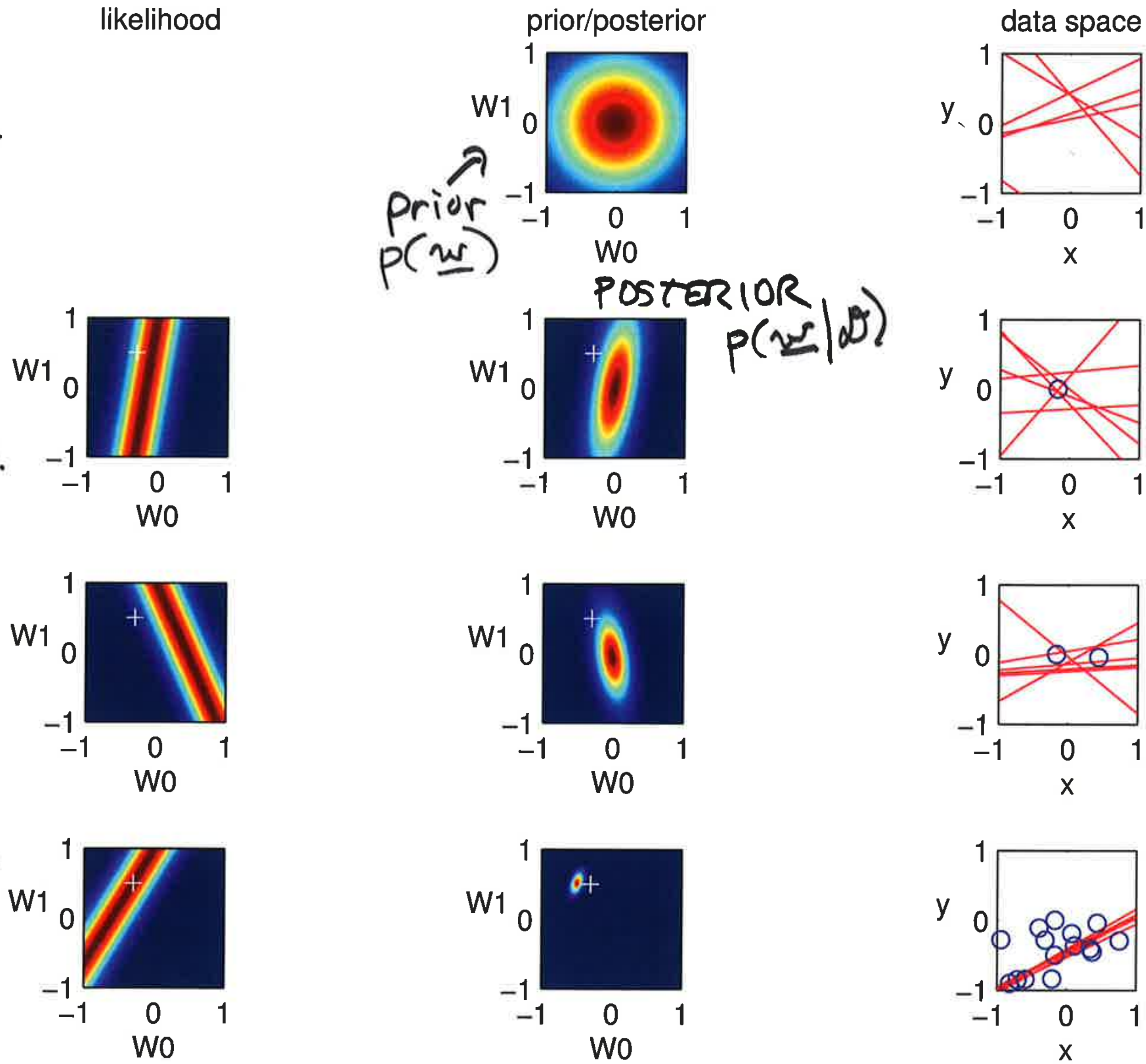
prior $p(\underline{w})$

POSTERIOR $p(\underline{w} \mid \mathcal{D})$

N=1 DATA POINT.

N=2

N=20



$\hat{f}(x) = w_0 + w_1 x$

Murphy Fig. 7.11.

Linear regr. w/ scalar input x_1.

$$\underline{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$y(x_1, \underline{w}) = w_0 + w_1 x_1 + \epsilon$ ← noiseterm.

SUMMARY OF ESTIMATION TECHNIQUES

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

GAUSSIAN CASE $\Longrightarrow$ $J_{MLE}(\underline{w}, \mathcal{D}) = MSE$

MAP: $\hat{\underline{\theta}}_{MAP} = \underset{\underline{\theta}}{\operatorname{argmax}} \left\{ \ln p(\mathcal{D} | \underline{\theta}) + \ln p(\underline{\theta}) \right\}$

GAUSSIAN CASE $\Longrightarrow$ $J_{MAP}(\underline{w}, \mathcal{D}) = N \cdot (MSE) + \lambda \|\underline{w}\|_2^2$

BAYESIAN: $p(\underline{\theta} | \mathcal{D}) \propto p(\mathcal{D} | \underline{\theta}) \, p(\underline{\theta})$

$$p(y | \underline{x}, \mathcal{D}) = \int p(y | \underline{x}, \underline{\theta}) \, p(\underline{\theta} | \mathcal{D}) \, d\underline{\theta}$$

POSTERIOR PREDICTIVE