Lecture 24 **EE 660** Nov 13, 2018

Lecture 24 announcements

- HW 12 - due Thursday

Lecture 24 outline

- Semi-supervised learning (part 2)
  - Mixture models for SSL
  - Maximum likelihood estimate (MLE)
  - Expectation maximization (EM)

(*) CORRECTIONS ON p. 4, 10.

MIXTURE MODELS FOR SSL

WE WANT TO FIND $p(y \mid \underline{x})$

LET'S MODEL EACH CLASS AS A SPECIFIED DENSITY WITH UNKNOWN PARAMETERS:

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

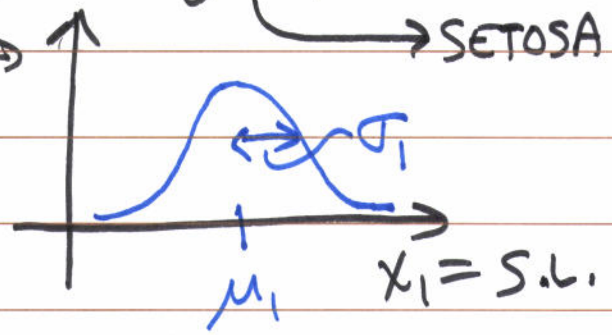
$$= \underbrace{p(\underline{x} \mid y, \underline{\theta})}_{\text{CLASS-CONDITIONAL DENSITY, CONDITIONED ON } \underline{\theta}} \pi_y \qquad (1)$$

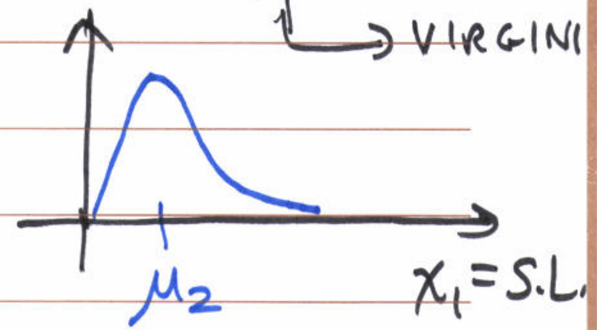
$(\pi_y = p(y)$ $=$ PRIOR ON $y)$.

OUR MODEL:

LABELED DATA:

$p(\underline{x} \mid y=1, \underline{\theta}) \longrightarrow$ SETOSA $\qquad p(\underline{x} \mid y=2, \underline{\theta}) \longrightarrow$ VIRGINICA



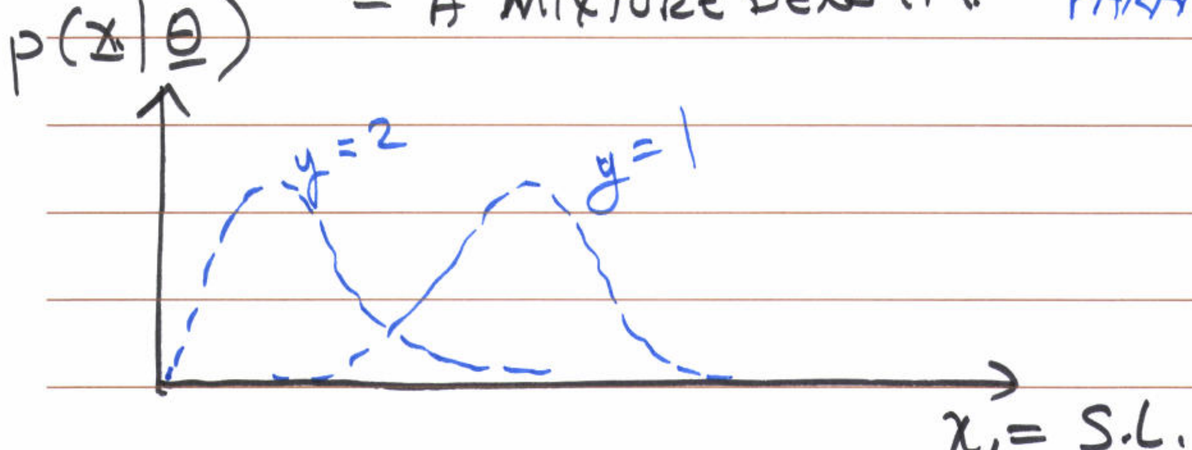


$$\underline{\theta} = \begin{bmatrix} \mu_1 \\ \sigma_1 \\ \mu_2 \\ \vdots \end{bmatrix}$$

UNLABELED DATA:

$$(2) \quad p(\underline{x} \mid \underline{\theta}) = \sum_{y=1}^{C} \underbrace{p(\underline{x} \mid y, \underline{\theta})}_{\text{COMPONENT DENS.}} \; \underbrace{\overbrace{p(y \mid \cancel{\underline{\theta}})}^{\pi_y}}_{\text{MIXING PARAMETER}}$$

= A MIXTURE DENSITY.



HOW USE BOTH MODELS (FOR $\mathcal{D}_L$ AND $\mathcal{D}_U$) TOGETHER?

FIND $\underline{\theta}$ FROM DATA USING MLE

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

$$p(\mathcal{D}|\underline{\theta}) = \prod_{i=1}^{l} p(\underline{x}_i, y_i|\underline{\theta}) \cdot \prod_{i=l+1}^{u+l} p(\underline{x}_i|\underline{\theta})$$

$$\ln p(\mathcal{D}|\underline{\theta}) = \underbrace{\sum_{i=1}^{l} \ln p(\underline{x}_i, y_i|\underline{\theta})}_{\mathcal{D}_L \text{ use (1)}} + \underbrace{\sum_{i=l+1}^{l+u} \ln p(\underline{x}_i|\underline{\theta})}_{\mathcal{D}_U \text{ use (2)}}$$

$$= \sum_{i=1}^{l} \ln p(\underline{x}_i|\overset{y_i,}{\underline{\theta}}) + \ln \pi_{y_i}$$

$$+ \sum_{i=l+1}^{u} \ln\left[\sum_{y=1}^{C} p(\underline{x}_i|y, \underline{\theta})\, \pi_{y}\right] \quad (*)$$

LET $\mathcal{D} \stackrel{\Delta}{=} \{\mathcal{D}_L, \mathcal{D}_U\}$ $\quad \overbrace{i = l+1, \cdots, u+l,}$

TREAT UNKNOWN LABELS y_i, ~~the $\mathcal{D}_U$~~ AS "HIDDEN VARIABLES", DENOTED $\mathcal{H}$.

→ USE EXPECTATION MAXIMIZATION (EM).

TO EST. $\mathcal{H}$ AND $\underline{\theta}$.

EM ALGORITHM (GENERAL FORMULATION)

[FOLLOWS SSL TEXT; ALSO MURPHY 11.4]

COMMONLY USED FOR:

- WORKING WITH MISSING DATA.
- FINDING MLE IN DIFFICULT SITUATIONS.
- ESTIMATING QUANTITIES IN MIXTURE MODELS.

LET $\mathcal{D}$ BE ALL THE DATA:

$$\{(\underline{x}_i, y_i),\ i=1,\cdots,l;\ \underline{x}_h,\ h=l+1,\cdots,l+u\}$$

LET $\mathcal{H}$ BE THE HIDDEN LABELS

$$\{y_h,\ h=l+1,\cdots,l+u\}$$

EM ALGORITHM (t = ITERATION INDEX)

INITIALIZE $t=0$ AND $\underline{\theta}^{(0)}$

E STEP:

COMPUTE BEST EST. OF $\mathcal{H}$ AS:

$$p(\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)})$$

(EST. HIDDEN LABELS y_h)

M STEP:

EST. PARAMETERS $\underline{\theta}^{(t+1)}$ BY:

$$\underline{\theta}^{(t+1)} = \underset{\underline{\theta}}{\operatorname{argmax}} \; \mathbb{E}_{\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)}} \left\{ \ln p(\mathcal{D}, \mathcal{H} \mid \underline{\theta}) \right\}$$

(USE y_h EST.'s TO COMPUTE NEW MLE OF $\underline{\theta}$).

$t \leftarrow t+1$

HALT WHEN $p(\mathcal{D} \mid \underline{\theta}^{(t)})$ CONVERGES.

EM PROPERTIES

1. CAN BE SHOWN THAT $p(\mathcal{D}|\underline{\theta})$ INCREASES AT EVERY ITERATION.
2. CONVERGES TO A LOCAL OPTIMUM.

   Is $-\ln p(\mathcal{D}|\underline{\theta})$ A CONVEX FCN. OF $\underline{\theta}$?

   $\rightarrow$ GENERALLY, NO.

3. RESULT DEPENDS ON STARTING POINT $\underline{\theta}^{(0)}$.

COMMON CHOICE: $\underline{\theta}^{(0)} = \underline{\hat{\theta}}_{MLE}$ BASED ON $\mathcal{D}_L$.

HOW TO USE IT:

FOR E STEP

LET i INDEX THE DATA PTS. IN $\mathcal{D}_L$;
h " " " " IN $\mathcal{D}_U$.

$$p(\mathcal{H} \mid \mathcal{D}, \underline{\theta}) = \prod_{h=l+1}^{l+u} p\left(y_h \mid \underline{\underline{X}}_U, \underline{\underline{Y}}_L, \underline{\underline{X}}_L, \underline{\theta}\right)$$

$$(3) \qquad = \prod_h p\left(y_h \mid \underline{x}_h, \underline{\theta}\right)$$

(KNOW $\underline{\theta}$ ⇒ DON'T NEED OTHER DATA.)

$$(4) \quad p\left(y_h = c \mid \underline{x}_h, \underline{\theta}\right) = \frac{p\left(\underline{x}_h \mid y_h = c, \underline{\theta}\right) p\left(y_h \mid \cancel{\underline{\theta}}\right)}{\sum_{y_h=1}^{C} p\left(y_h \mid \cancel{\underline{\theta}}\right) p\left(\underline{x}_h \mid y_h, \underline{\theta}\right)}$$

LET $\gamma_{hc} \triangleq p\left(y_h = c \mid \underline{x}_h, \underline{\theta}\right)$

DATA PT. INDEX (UNLABELED)

CLASS ASSIGNMENT

γ_{nc} = RESPONSIBILITY OF $y_n = c$ LABEL FOR DATA PT. $\underline{x}_n$.
= "SOFT LABEL" FOR DATA PT. $\underline{x}_n$.

$p(y_n = c \mid \underline{x}_n, \underline{\theta})$.

FOR M STEP

$$\max_{\underline{\theta}} \mathbb{E}_{\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)}} \left\{ \ln p(\mathcal{D}, \mathcal{H} \mid \underline{\theta}) \right\}$$

$$= \max_{\underline{\theta}} \left\{ \sum_{\mathcal{H}} p(\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)}) \ln p(\mathcal{D}, \mathcal{H} \mid \underline{\theta}) \right\}$$

$$p(\mathcal{D}, \mathcal{H} \mid \underline{\theta}) = \underbrace{p(\mathcal{H} \mid \mathcal{D}, \underline{\theta})}_{\text{GIVEN ABOVE. EQ. (3),(4).}} \underbrace{p(\mathcal{D} \mid \underline{\theta})}_{\text{LIKELIHOOD OF ALL KNOWN (GIVEN) DATA.}}$$

$\cancel{p(\mathcal{D} \mid \underline{\theta}}$

$$p(\mathcal{D} \mid \underline{\theta}) = \left[\prod_{i=1}^{l} \underbrace{p(\underline{x}_i, y_i \mid \underline{\theta})} \right] \prod_{h=l+1}^{l+u} \underbrace{p(\underline{x}_h \mid \underline{\theta})}$$

$$p(\underline{x}_i, y_i \mid \underline{\theta}) = \underbrace{p(\underline{x}_i \mid y_i, \underline{\theta})}_{\text{MODEL FOR LABELED DATA.}} \; \underbrace{p(y_i \mid \cancel{\underline{\theta}})}_{\pi_{y_i}}$$

MIXTURE DENSITY (EQ. (2)):

$$p(\underline{x}_h \mid \underline{\theta}) = \sum_{y_h=1}^{C} p(\underline{x}_h \mid y_h, \underline{\theta}) \, \pi_{y_h} .$$ *