Lecture 25 announcements

- HW13 will be posted; due after Thanksgiving break.
- My office hours tomorrow will be 12:00 - 1:00 PM

Lecture 25 outline

- Semi-supervised learning - concluding remarks
- Start Unsupervised learning (USL)
  - Introduction
  - Mixture models, MLE, and Expectation Maximization
  - Example

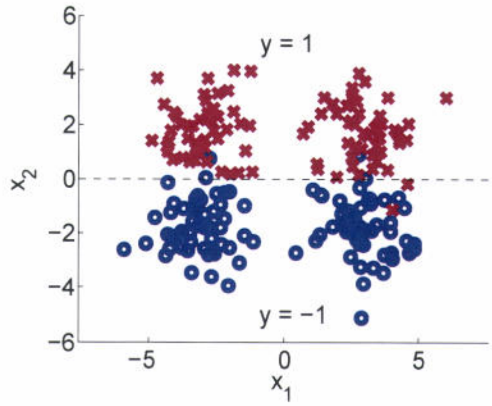


Figure 3.2: Two classes in four clusters (each a 2-dimensional Gaussian distribution).

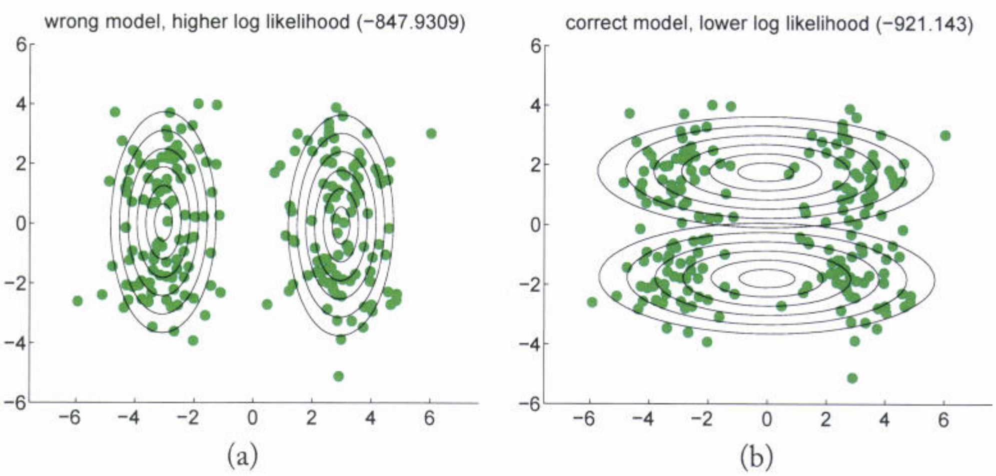


Figure 3.3: (a) Good fit under the wrong model assumption. The decision boundary is vertical, thus producing mass misclassification. (b) Worse fit under the wrong model assumption. However, the decision boundary is correct.

decision boundary would be approximately the line $y = -x$, which would result in only about 25% error.

There are a number of ways to alleviate the danger of using the wrong model. One obvious way is to refine the model to fit the task better, which requires domain knowledge. In the above example, one might model each class itself as a GMM with two components, instead of a single Gaussian.

Another way is to de-emphasize the unlabeled data, in case the model correctness is uncertain. Specifically, we scale the contribution from unlabeled data in the semi-supervised log likeli-

SEMI-SUPERVISED LEARNING – CONCLUDING REMARKS

[FIG. 3.2 & 3.3 OF SSL TEXT]

EM w/ MIXTURE DENS. FOR SSL

- WORKS WELL WHEN MODEL IS ~ CORRECT.
- OTHERWISE MIGHT NOT WORK WELL.

OTHER TOPICS IN SSL (OPTIONAL – N.R.F.)

- CLUSTER-THEN-LABEL METHODS. (END OF Ch. 3)
- CO-TRAINING (Ch. 4)
  (2 VIEWS OF DATA).
- GRAPH BASED METHODS
- SVM BASED TECHNIQUES
- BOUNDS ON E_{out} [INTRO. TO Ch 8].
- MODELING HOW HUMANS LEARN WITH SSL.

UNSUPERVISED LEARNING (USL)

SAMPLES (DATA POINTS) IN $\mathcal{D}$ HAVE NO LABELS.

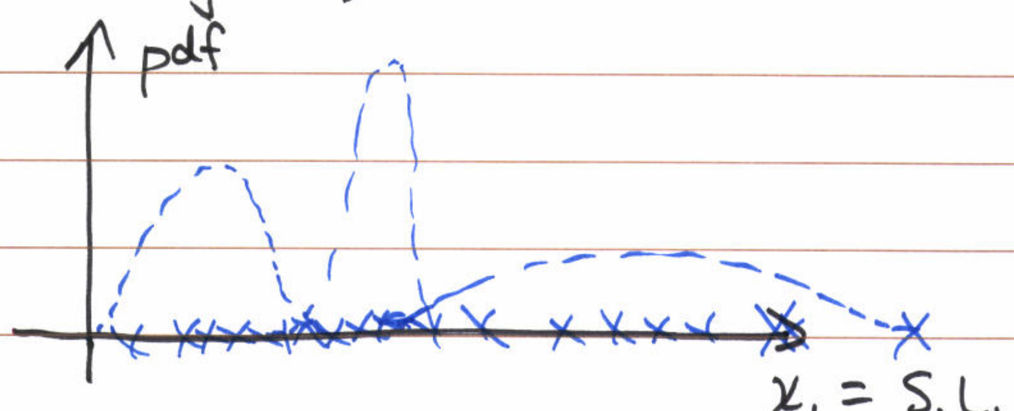
EXAMPLES OF ITS USEFULLNESS:

- CLUSTERING
  e.g.: RECOMMENDER SYSTEMS (MOVIES, NEWS)
- DIMENSIONALITY REDUCTION (e.g., PCA)
- FILLING IN / DEALING WITH MISSING DATA.
- LEARNING ABOUT CHARACTERISTICS IN THE DATA.

MIXTURE DENSITIES FOR USL / CLUSTERING

(MIXTURE MODELS)

MODEL EACH CLUSTER AS A pdf ~~with~~ WITH UNKNOWN PARAMETERS (TYPICAL FOR CONTINUOUS FEATURES $x_j \in \mathbb{R}$).



THUS:

$$p(\underline{x}|\underline{\theta}) = \sum_{k=1}^{K} p(\underline{x}|z=k, \underline{\theta}_k)\, \underbrace{p(z=k|\underline{\theta})}_{\pi_k}$$

z = CLUSTER INDEX

$\underline{\theta}_k$ = PARAMETERS FOR CLUSTER k.

= A MIXTURE MODEL

$$p(\underline{x}|\underline{\theta}) = \sum_{k=1}^{K} \pi_k\, \underbrace{p(\underline{x}|z=k, \underline{\theta}_k)}$$

OUR MODEL FOR CLUSTER k.

AND $\sum_{k=1}^{K} \pi_k = 1.$

MIXING PARAMETER (OR WEIGHT) FOR CLUSTER k.

FIND MLE OF $\underline{\theta}$?

LIKELIHOOD: $p(\mathcal{D}|\underline{\theta}) = \prod_{i=1}^{N} \underbrace{p(x_i|\underline{\theta})}_{\text{MIXTURE.}}$

→ GENERALLY NOT SOLVABLE ALGEBRAICALLY.

→ TYPICALLY USE AN ITERATIVE OPTIMIZATION.

→ USE EM (AGAIN).

EM FOR CLUSTERING USING MIXTURE MODELS

SAME BASIC ALGORITHM AS EM FOR SSL.

LET $\mathcal{D} = \{\underline{x}_i\}_{i=1}^{N}$

$\mathcal{H} = \{z_i\}_{i=1}^{N}$, z_i IS CLUSTER LABEL FOR DATA PT. $\underline{x}_i$

$z_i \in \{1, 2, \cdots, K\}$.

ALGORITHM

1. INITIALIZE $t=0$ AND $\underline{\theta}^{(0)}$
2. ITERATE (INDEX t):

2.1 E STEP: COMPUTE $p(\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)})$

2.2. M STEP: FIND:

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

2.3 $t \leftarrow t+1$

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

3. OUTPUT $\hat{\underline{\theta}} = \underline{\theta}^{(t)}$

EQUATIONS FOR E STEP

$$p(\mathcal{H} \mid \mathcal{D}, \underline{\theta}) = \prod_{i=1}^{N} \underbrace{p(z_i \mid \underline{x}_i, \underline{\theta})}$$

$$p(z=k \mid \underline{x}, \underline{\theta}_k) = \frac{\overbrace{p(\underline{x} \mid z=k, \underline{\theta}_k)}^{\text{our model}} \overbrace{p(z=k \mid \underline{\theta}_k)}^{\pi_k}}{\sum_{k'=1}^{K} p(\underline{x} \mid z=k', \underline{\theta}_{k'})\, p(z=k' \mid \underline{\theta}_{k'})}$$

$\rightarrow$ SOFT LABEL $\gamma_{ik}^{(t)} = p(z_i = k \mid \underline{x}_i, \underline{\theta}_k^{(t)})$

EQNS. FOR M STEP

$$p(\mathcal{D}, \mathcal{H} \mid \underline{\theta}) = \prod_{i=1}^{N} p(\underline{x}_i, z_i \mid \underline{\theta})$$

$$= \prod_{i=1}^{N} \underbrace{p(\underline{x}_i \mid z_i, \underline{\theta})}_{\text{our model}} \underbrace{p(z_i \mid \underline{\theta})}_{\pi_k}$$

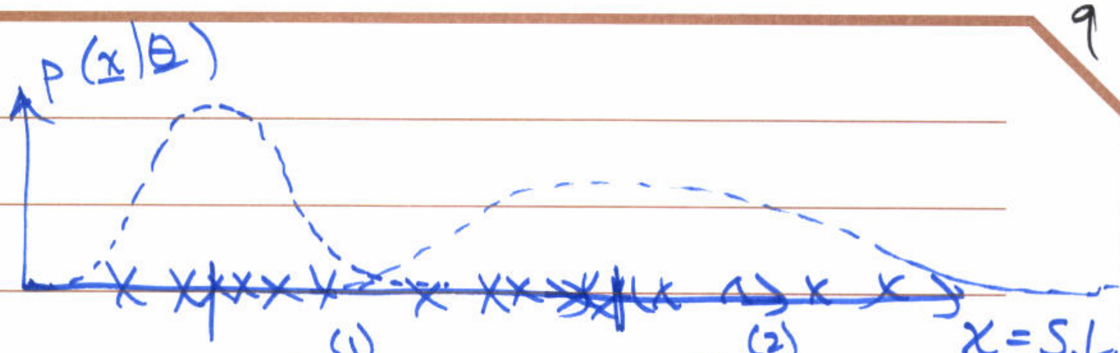
$$\rightarrow \underline{\theta}^{(t+1)} = \underset{\underline{\theta}}{\operatorname{argmax}} \left\{ \sum_{\mathcal{H}} p(\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)}) \cdot \ln p(\mathcal{D}, \mathcal{H} \mid \underline{\theta}) \right\}$$

COMMENTS

1. ALGORITHM CHARACTERISTICS – SAME AS FOR SSL EM.
2. CHOICE OF $\underline{\theta}^{(0)}$ – IF NO PRIOR KNOWLEDGE, CAN RUN ALG. A NUMBER OF TIMES WITH DIFFERENT (~RANDOM) CHOICES OF $\underline{\theta}^{(0)}$. COMPARE RESULTS USING $p(\mathcal{D} \mid \underline{\theta})$.
3. DOES THERE EXIST A UNIQUE BEST SOLUTION? IF MIXTURE $p(\underline{x} \mid \underline{\theta})$ IS <u>IDENTIFIABLE</u>, THEN ~~YES~~ IN LIMIT $N \to \infty$, YES.

DEF: A DENSITY $p(\underline{x} \mid \underline{\theta})$ IS <u>IDENTIFIABLE</u> IF $p(\underline{x} \mid \underline{\theta}_1) = p(\underline{x} \mid \underline{\theta}_2) \ \forall \underline{x}$ $\Rightarrow \underline{\theta}_1 = \underline{\theta}_2$ (UP TO A PERMUTATION OF MIXTURE COMPONENT INDICES).*

IN PRACTICE, IF $x_j \in \mathbb{R}$, AND HAVE A SUFFICIENT NUMBER OF DATA POINTS, IDENTIFIABILITY IS USUALLY SATISFIED.



1st SOL'N: $\mu_1 = \mu_0^{(1)}$, $\sigma_1^2 = \sigma_0^{(1)^2}$ $\qquad$ $\mu_2 = \mu_0^{(2)}$, $\sigma_2^2 = \sigma_0^{(2)^2}$

2nd SOL'N: $\mu_2 = \mu_0^{(1)}$, $\sigma_2^2 = \sigma_0^{(1)^2}$ $\qquad$ $\mu_1 = \mu_0^{(2)}$, $\sigma_1^2 = \sigma_0^{(2)^2}$

STILL IDENTIFIABLE.

EX: GAUSSIAN MIXTURE MODEL (GMM) USING EM, $K=2$

$$p(\underline{x} \mid z=k, \underline{\theta}) = N(\underline{x} \mid \underbrace{\underline{\mu}_k, \underline{\underline{\Sigma}}_k}_{\underline{\theta}_k})$$

(a) $\underline{\underline{\Sigma}}^{(0)}_{z=1} = \underline{\underline{\Sigma}}^{(0)}_{z=2} = \underline{\underline{I}}$

$\underline{\mu}_k^{(0)}$ = shown.

– [FIG. 9.8 OF BISHOP (p. 437) — EM TO CLUSTER OLD FAITHFUL DATA.]