Tuesday, 11/10/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 23

Announcements

- Homework 9 will be posted later this week
- End-of-semester quiz
  - Tues., 11/24/2020, 5:30-7:00 PM
  - Will be short-answer questions, taken online
  - Will be 60-90 minutes in length
  - Materials allowed to be announced

Today's topics

- Semi-supervised learning (part 3)
  - Expectation Maximization (EM)
- Unsupervised learning (USL) (part 1)
  - Introduction
  - Mixture models and EM → deferred

EM Algorithm

Initialize $t = 0$ and $\underline{\theta}^{(0)}$

E step

Compute best est.
of $\mathcal{H}$ as $p(\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)})$

M step

Estimate parameters $\underline{\theta}^{(t+1)}$ by:

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

$t = t + 1$

Halt when $p(\mathcal{D} \mid \underline{\theta}^{(t)})$ converges

EM Properties

1. Can be shown that $p(\mathcal{D} \mid \underline{\theta})$ increases at every iteration.
2. Converges to local optimum.
3. Result depends on starting point $\underline{\theta}^{(0)}$.

   Common choice is:

$$\underline{\theta}^{(0)} = \underline{\hat{\theta}}_{MLE} \text{ based on } \mathcal{D}_L.$$

How to use it: For E step

Let i index the data pts in $\mathcal{D}_L$; h index the data pts. in $\mathcal{D}_U$.

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

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

(if know $\underline{\theta}$, don't need other $\underline{x}_n$, $\underline{\underline{X}}_L$, or $\underline{Y}_L$ to predict y_n).

(4) $$p(y_h = c \mid \underline{x}_h, \underline{\theta}) = \frac{p(\underline{x}_h \mid y_h = c, \underline{\theta})\, p(y_h = c \mid \cancel{\underline{\theta}})}{\sum_{y_u=1}^{C} \underbrace{p(y_u = c \mid \cancel{\underline{\theta}})}_{\pi_{y_u}}\, p(\underline{x}_h \mid y_u, \underline{\theta})}$$

π_y

Let $\gamma_{hc} \triangleq p(y_h = c \mid \underline{x}_h, \underline{\theta})$ = responsibility of $y_h = c$ for data pt. $\underline{x}_h$.

data pt. index ↑ ↑ class ass'n.

= "soft label" for $\underline{x}_h$.

$p(\mathcal{H} \mid \mathcal{D}, \underline{\theta})$ gives soft labels of all unlabeled data pts.

M step

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

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

$$p(\mathcal{D},\mathcal{H}|\underline{\theta}) = \underbrace{p(\mathcal{H}|\mathcal{D},\underline{\theta})}_{\substack{\text{given above} - (3),(4) \\ \text{from E step}}} \underbrace{p(\mathcal{D}|\underline{\theta})}_{\substack{\text{likelihood} \\ \text{of all known data}}}$$

$$p(\mathcal{D}|\underline{\theta}) = \prod_{i=1}^{\ell} \underbrace{p(\underline{x}_i, y_i|\underline{\theta})} \prod_{h=\ell+1}^{\ell+u} \underbrace{p(\underline{x}_h|\underline{\theta})}$$

from (2): $p(\underline{x}_h|\underline{\theta}) = \sum_{y=1}^{C} p(\underline{x}_h|y,\underline{\theta})\,\pi_y$

(a mixture density)

$$p(\underline{x}_i, y_i|\underline{\theta}) = \underbrace{p(\underline{x}_i|y_i,\underline{\theta})}_{\substack{\text{model w/ labeled data} \\ \text{(class-conditional pdf)}}} \underbrace{p(y_i|\underline{\theta})}_{\text{class prior.}}$$

Comments: (EM w/ mixture dens. for SSL)

1. Works well when model is ~correct.
2. Othewise might not work well.

[Fig. 3.2, 3.3 in Zhu SSL textbook]

Other topics in SSL (N.R.F.)

- Cluster-then-label methods (end of Ch. 3)
- Co-training (Ch. 4)
  - Each instance $\underline{x}_i$ has 2 views (feature sets): $\underline{x}^{(1)}, \underline{x}^{(2)}$.
    e.g.: words/letters in a phrase to classify: $\underline{x}^{(1)}$
    context (nearby words): $\underline{x}^{(2)}$
  - Use both for SSL
- Graph-based methods
- SVM based methods
- Bounds on E_{out} [Intro. in Ch. 8]

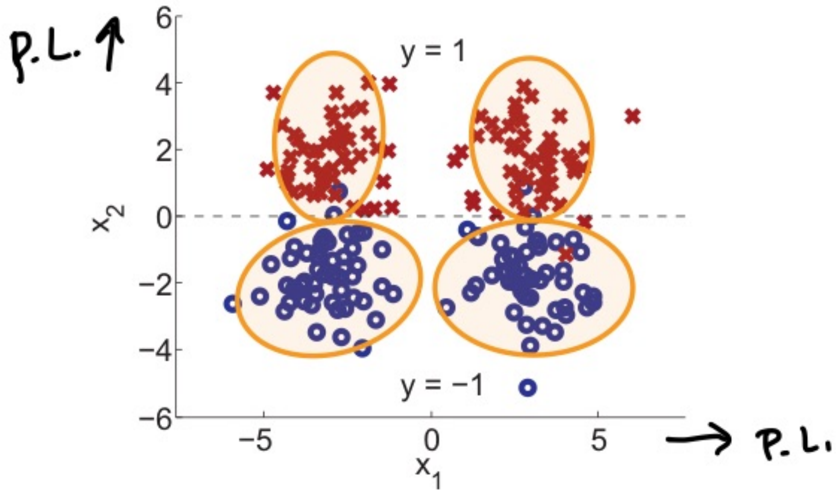

Figure 3.2: Two classes in four clusters (each a 2-dimensional Gaussian distribution).

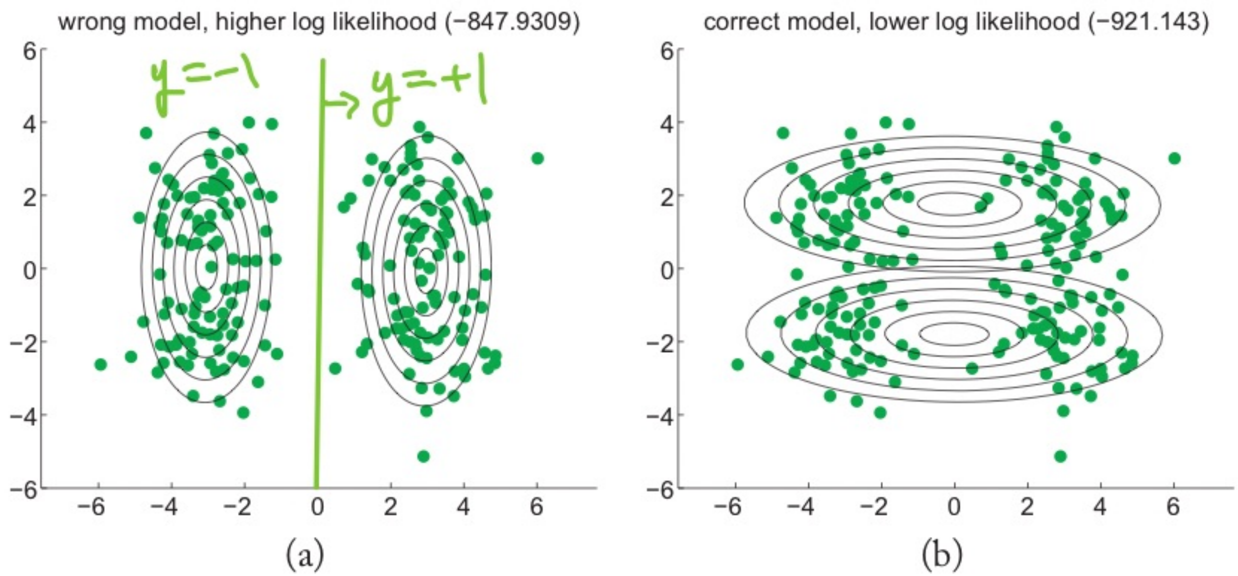

Figure 3.3: (a) Good fit under the wrong model assumption. The decision boundary is vertical, thus producing mass misclassification. (b) Worse fit under the wrong model assumption. However, the decision boundary is correct.

[From Zhu et al., Intro. to Semi-Supervised Learning.]

Unsupervised Learning (USL)

Data pts. in $\mathcal{D}$ have no class labels.

What can USL do for us?

- Learn structure in the data – significance of clusters.
- Feature selection or feature discovery
  - e.g.: cluster data
    - find centroid of each cluster μ_k
    - use distance $d(\underline{x}_i, \mu_k)$ to each μ_k as a feature value.
    - → bag-of-words representation
- Adapt a classifier to changes over time by revising cluster/class boundaries as new data is observed.