Thursday, 11/12/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 24

p. 1

Lecture 24	EE 660	Nov 12, 2020

Announcements

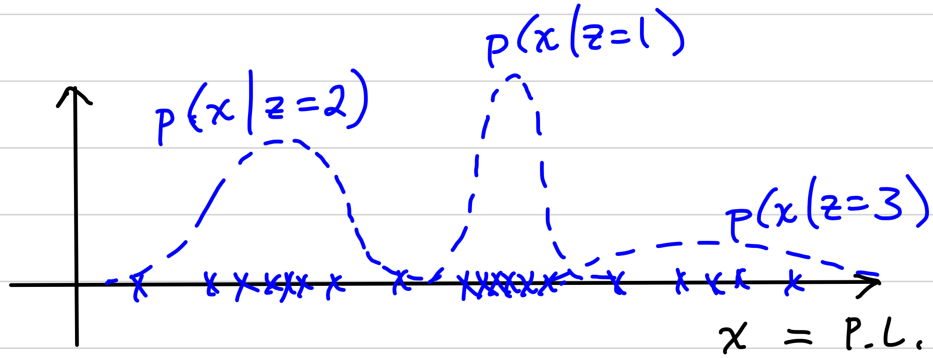
- Homework 9 will be posted

Today's topics

- Unsupervised learning (USL) (part 2)
  - Expectation Maximization (EM)
    - Mixture models in USL
    - EM algorithm and equations
    - Example
  - Similarity and dissimilarity measures for clustering

Mixture Models for USL

- Model each cluster as a pdf with unknown parameters
  - usually for continuous features $x_j \in \mathbb{R}$



- $p(\underline{x}|\underline{\theta}) = \underbrace{\sum_{k=1}^{K}}_{\text{mixture}} \underbrace{p(\underline{x}|z=k,\underline{\theta})}_{\text{use } \underline{\theta}_k = \text{ parameters for cluster } k} \underbrace{p(z=k|\underline{\theta})}_{\pi_k}$, z = cluster index (label), K = total # of clusters

(1) $p(\underline{x}|\underline{\theta}) = \sum_{k=1}^{K} \pi_k \, \underbrace{p(\underline{x}|z=k,\underline{\theta}_k)}_{\text{our model for cluster } k}$, with $\sum_{k=1}^{K} \pi_k = 1$

π_k: prior or mixing paramter (weight) for cluster k.

- Goal: find MLE of $\underline{\theta}$, or $\underline{\theta}_k$ for $k=1,2,\cdots,K$
  - Likelihood: $p(\mathcal{D}|\underline{\theta}) = \prod_{i=1}^{N} p(\underline{x}_i|\underline{\theta})$
  - Generally not solvable analytically $\Rightarrow$ Use EM

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

- Algorithm (EM for estimating $\underline{\theta}$ and $\mathcal{H}$)

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

  2. Iterate (index t: $0, 1, 2, \cdots, T-1$)

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

     2.2 M step: Find $\underline{\theta}^{(t+1)} = \underset{\underline{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{\mathcal{H} \mid \mathcal{D}, \underline{\theta}^{(t)}} \{ \ln p(\mathcal{D}, \mathcal{H} \mid \underline{\theta}) \}$

     2.3 Halt when $p(\mathcal{D} \mid \underline{\theta}^{(t+1)})$ converges

  3. Output $\hat{\underline{\theta}} = \underline{\theta}^{(T)}$

Equations for E step

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

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

→ soft label (responsibility) $\gamma_{ik}^{(t)} = p(z_i=k|\underline{x}_i,\underline{\theta}_k^{(t)})$

k: cluster index

i: data pt. index

Equations for M step

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

for $z_i=k$: $\underbrace{p(\underline{x}_i|z_i=k,\underline{\theta}_k)}_{\text{our model}}\,\pi_k$

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

Comments

1. Algorithm characteristics – same as for SSL EM.

2. Choice of $\underline{\theta}^{(0)}$ if no prior knowledge?
   - Use a simpler clustering algorithm (e.g., Kmeans), then use its resulting clusters to calculate $\underline{\theta}^{(0)}$ for EM.

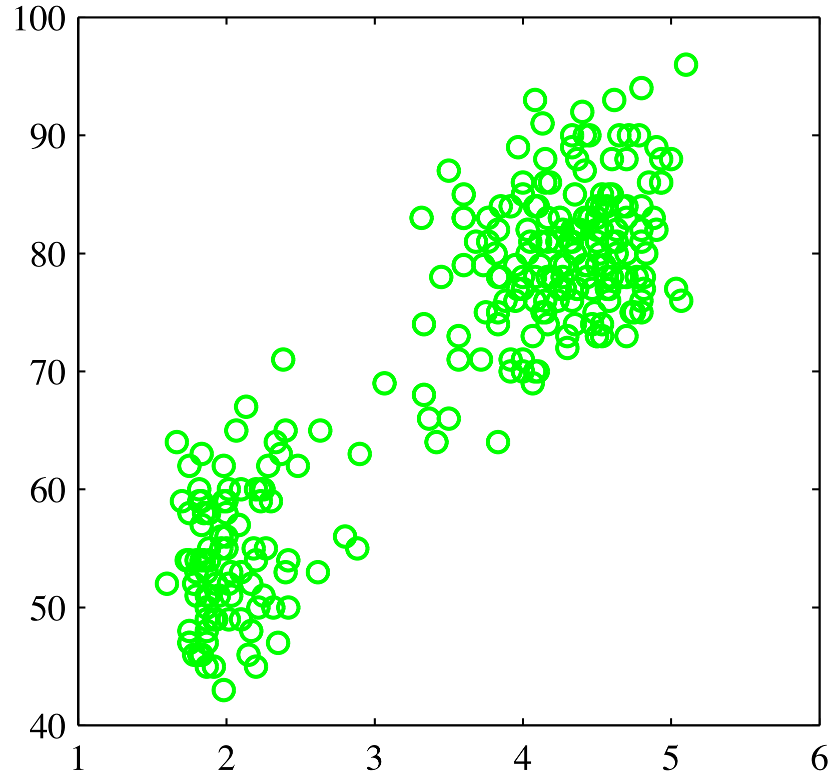
   or
   - Try many different (random) $\underline{\theta}^{(0)}$; compare the results using $p(\mathcal{D} \mid \underline{\theta})$.

3. Quality of clustering result depends on how appropriate the model is for the data.

Example

"Old Faithful" geyser.

Old Faithful Data



Time to next eruption (minutes) vs. duration of eruption (minutes)

From Bishop Fig. A.5

Model

Gaussian mixture model (GMM) using EM, $K=2$.

$$p(\underline{x} \mid \underline{\theta}) = \sum_{k=1}^{2} p(\underline{x} \mid z=k, \underline{\theta}_k)\, p(z=k \mid \underline{\theta}_k) = \sum_{k=1}^{2} N(\underline{x} \mid \underbrace{\underline{\mu}_k, \underline{\underline{\Sigma}}_k}_{\underline{\theta}_k})\, \pi_k$$

Algorithm steps

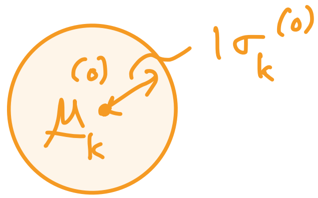
(a) $\underline{\theta}^{(0)}$: $\underline{\underline{\Sigma}}^{(0)}_{z=1} = \underline{\underline{\Sigma}}^{(0)}_{z=2} = \underline{\underline{I}}$. $\underline{\mu}^{(0)}_k$ is shown.

(b) First E step result: $\gamma^{(0)}_{ik} = p(z_i = k \mid \underline{x}_i, \underline{\theta}^{(0)}_k)$

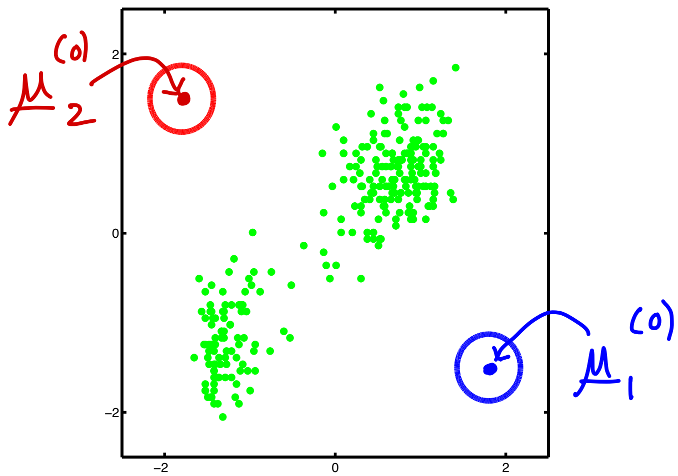
blue ink $\propto \gamma_{i,k=1}$; red ink $\propto \gamma_{i,k=2}$

(c) After first M step: $\underline{\theta}^{(1)}$ gives new $\underline{\mu}^{(1)}_k$, $\underline{\underline{\Sigma}}^{(1)}_k$

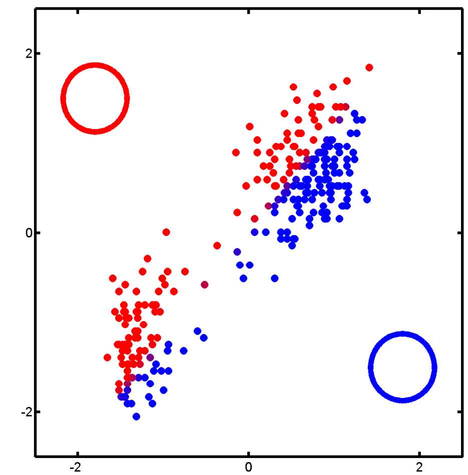
new Gaussians $p(\underline{x} \mid z=k, \underline{\theta}_k)$ are shown.



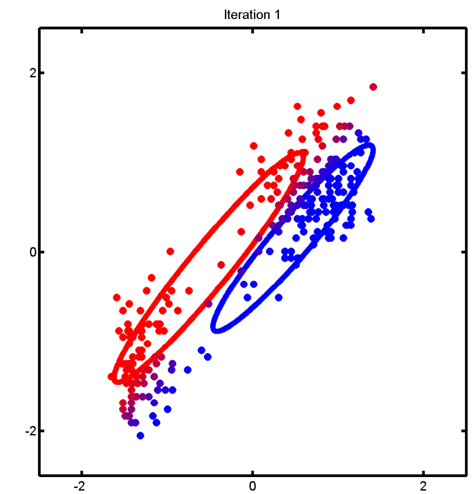
EM Example: Old Faithful Data



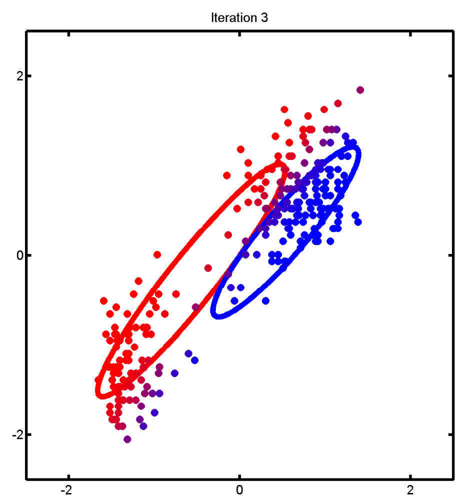
(a) Initial condition



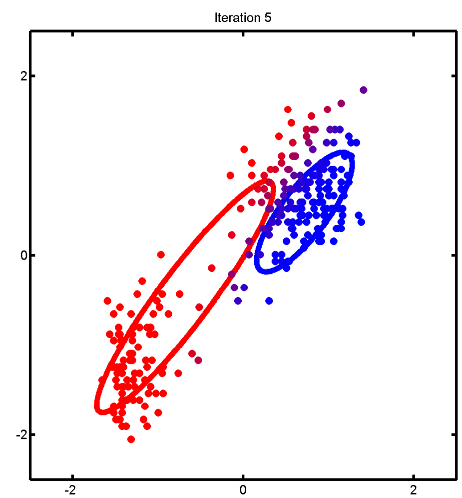
(b) After first E step



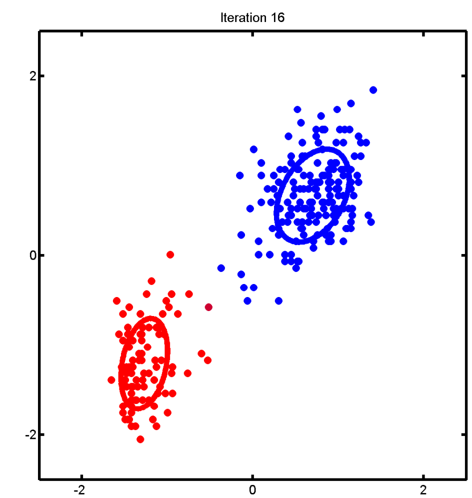
(c) After first M step



(d) After 3 iterations



(e) After 5 iterations



(f) After 16 iterations

From Murphy Fig. 11.11

Similarity and Dissimilarity Measures

Need measures for similarity or dissimilarity between:

→ • 2 points $\underline{x}_i, \underline{x}_j$.
• A point and a cluster
• 2 clusters

Need measure for quality of a partitioning [difficult problem]

Let $\Delta(\underline{x}_i, \underline{x}_{i'}) = d_{ii'}$ denote a dissimilarity function

Can use a distance function, e.g.:

$$\Delta(\underline{x}_i, \underline{x}_{i'}) = d_{ii'} = \sum_{j=1}^{D} \Delta_j(x_{ij}, x_{i'j})$$

$\Delta_j(x_{ij}, x_{i'j})$ can be: $(x_{ij} - x_{i'j})^2$ (Eucl. dist.)^2

$|x_{ij} - x_{i'j}|$ (ℓ_1 norm
or city block dist.
or Manhattan dist.)

For <u>nominal</u> features (symbolic, categorical, or labels):

$$\Delta(x_{ij}, x_{i'j}) = \#\text{ of features that are different}$$

$$= \sum_{j=1}^{D} \mathbb{I}(x_{ij} \neq x_{i'j})$$

$$= \text{Hamming distance.}$$

Can let $\Delta(\underline{x}_i, \underline{x}_{i'})$ = this or some other $d_{ii'}$, s.t. $d_{ii'} \geq 0 \ \forall i, i'$
and $d_{ii} = 0 \ \forall i$.

or $f(d_{ii'})$, f = any monotonically increasing fcn.