Mixture Models and Order Selection

Lesson 6 : Lab Session

Advanced Machine Learning, CentraleSupelec

Teacher's Assistant: Omar CHEHAB

Professors : Emilie CHOUZENOUX, Frederic PASCAL

General Information

- **Assignment :** alone or in pairs, you will code the algorithms you learnt in 'scikit-learn formalism', and apply them to images and text.

- **Due :** the 5 lab assignments for lessons 3-7 are due a week from when they are given, at aml.centralesupelec.2020@gmail.com

- **Grading** : each assignment is worth 4 points — your 4 best labs out of the 5 will be retained and will count for half of your final grade.

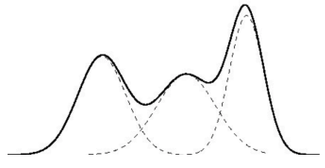
- **Questions :** questions or feedback are welcome after class or by email at l-emir-omar.chehab@inria.fr

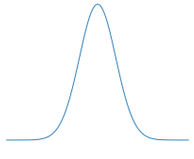
Mixture Model

Builds a rich (e.g. multimodal) distribution
…from (a weighted sum of) simpler ones.

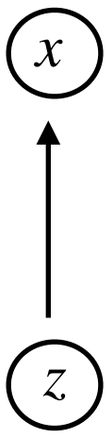
$$p_\theta(x) = \sum_{z=1}^{K} p_\theta(x \mid z = k)\; p_\theta(z = k)$$

complex model — simple model k — its weight





Gaussian Mixture Model



$$X \mid z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

each cluster is gaussian (mean and variance)

$$= \theta$$

$$Z \sim Multinom(\pi_1, \ldots, \pi_K)$$

prior on K clusters

Inference

Maximum-Likelihood

X and Z are observed

data-point and cluster assignment $\underset{\in \mathbb{R}^p \ \in \mathbb{R}}{(x_i, z_i)}_{i=1,\dots,n}$

'easy' $\max_{p_\theta} \log p_\theta(x, z)$

direct

MLE

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} 1_{z_i=k}$$

$$\hat{\mu}_k = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} 1_{z_i=k} \, x_i$$

$$\hat{\sigma}_k^2 = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} 1_{z_i=k} \left(x_i - \hat{\mu}_k\right) \left(x_i - \hat{\mu}_k\right)^T$$

X only is observed (*Z is hidden*)

$\underset{\in \mathbb{R}^p}{(x_i)}_{i=1,\dots,n}$ data-point without cluster assignment

$\max_{p_\theta} \log p_\theta(x)$ *'hard'*

indirect (via E.M.)

E-Step

update (soft) assignment

$$p_{\theta_t}\left(Z = k \mid X = x_i\right) = \frac{\hat{\pi}_k^t \times p_{\theta_t}(x_i \mid z = k)}{\sum_{l=1}^{K} \hat{\pi}_l^t \times p_{\theta_t}(x_i \mid z = l)}$$

like in K-means but 'soft'

M-Step

update cluster

$$\underset{\in \mathbb{R}}{\hat{\pi}_k^{t+1}} = \frac{1}{n} \sum_{i=1}^{n} P_{\theta_t}\left(Z = k \mid X = x_i\right)$$

$$\underset{\in \mathbb{R}^p}{\hat{\mu}_k^{t+1}} = \frac{1}{n\hat{\pi}_k^{t+1}} \sum_{i=1}^{n} P_{\theta_t}\left(Z = k \mid X = x_i\right) x_i$$

$$\underset{\in \mathbb{R}^{p \times p}}{\hat{\Sigma}_k^{t+1}} = \frac{1}{n\hat{\pi}_k^{t+1}} \sum_{i=1}^{n} P_{\theta_t}\left(Z = k \mid X = x_i\right) \left(x_i - \hat{\mu}_k^{t+1}\right) \left(x_i - \hat{\mu}_k^{t+1}\right)^T$$

The idea behind E.M.

marginal log-likelihood $\log p_\theta(x)$

$\geq$ Jensen

lower-bound $\mathscr{L}(p_\theta, q)$

$= H(Q) - H(Q,P)$

$= -\mathbb{E}_{q(z)}[\log q(z)] + \mathbb{E}_{q(z)}[\log p_\theta(z,x)]$

max_{p_θ} ↑ *M-step :* maximize

max_q ↑ *E-step* : tighten the bound } gap

$= KL(q(z)||p(z|x))$

The marginal log-likelihood is difficult to maximize directly
…so do so via a tractable lower-bound!

E-Step $\max_q \mathscr{L}(p_\theta, q)$ or equivalently $\min_q gap(p_\theta, q)$ $\Rightarrow$ $q(z) = p(z|x)$

M-Step $\max_{p_\theta} \mathscr{L}(p_\theta, q) \iff \max_{p_\theta} \log \mathbb{E}_{z\sim q(z)}[p(x,z)]$

<u>stochastic EM:</u>
replace Expectation over q with a sample

<u>N-EM:</u>
replace Expectation over q with N samples

<u>EM:</u>
use Expectation

Model order

I can model my data $\mathscr{D} = (x_i)_{i \in [1,N]}$ with any model(s) from $(\mathscr{M}_1, \ldots, \mathscr{M}_K)$: which is best?

Choose a model:

$$\max_j$$

	model **quality**	capacity to **overfit**	capacity to **underfit**
$AIC(j)$	$=$	$\mathscr{L}(\hat{\theta}_{MLE})$	$- M_j$
$BIC(j)$	$=$	$\mathscr{L}(\hat{\theta}_{MLE})$	$- \frac{1}{2} M_j N$
		rewards **complexity**	penalizes **complexity**

M_j = nb of unknown params for model j

harsher

tradeoff

Reduce a model:
(specific here to Gaussian Mixture Model)

after training, merge two clusters (k, k') by the weighted sum of their statistics:

$$\mu_{new} = \frac{n_k \mu_k + n_{k'} \mu_{k'}}{n_k + n_{k'}} \qquad \Sigma_{new} = \frac{n_k \Sigma_k + n_{k'} \Sigma_{k'}}{n_k + n_{k'}}$$

$$\pi_{new} = \pi_k + \pi_{k'}$$

with n_k points in cluster k

correlation criterion: if the profiles of their assigned points are ‘similar enough’

$$PearsonR\left(p_{\theta^*}(z = k_1 \,|\, x = \ldots), p_{\theta^*}(z = k_1 \,|\, x = \ldots)\right) > 1 - \epsilon$$

distance criterion: or if their centroids are ‘close enough’

$$\|\mu_k - \mu_{k'}\| \leq \epsilon$$

parameter criterion: or if it reduces the intra-class variance

$$d_{new} < d_k + d'_k - \epsilon$$

where d_k is the variance of cluster k

Assignment: plan

1. Gaussian Mixture Model : train using E.M. algorithm (*your own code*)

2. Choose number of clusters : model order *selection* (AIC, BIC)

   or model order *reduction* (3 criteria) (*your own code*)

3. Application : vision (MNIST digits dataset) (*your own code*)