Lecture 12 vol. 2

EM

Information Systems
(Machine Learning)

Andrey Filchenkov

13.12.2018

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM

- The presentation is prepared with materials of the K.V. Vorontsov's course "Machine Leaning".

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM

The two problems of probabilistic classification

First problem: **probability density recovering**

Given: $T^\ell = \{(x_i, y_i)\}_{i=1}^{\ell}$.

Problem: find empirical estimates $\widehat{\Pr}(y)$ и $\hat{p}(x|y)$, $y \in Y$.

Second problem: **mean risk minimization**

Given:

- prior probabilities $\Pr(y)$,
- likelihood $p(x|y), y \in Y$.

Problem: find classifier a which minimizes $R(a)$.

Which of these two problems is already solved and what is the answer?

Distributions mixture recovery

Generative distributions mixture model:

$$p(x) = \sum_{j=1}^{k} w_j p_j(x),$$

where $w_j > 0, \sum_{j=1}^{k} w_j = 1; p_j(x) = \varphi\left(x; \theta_j\right)$ is likelihood function of jth mixture component, w_j is its prior probability, k is the number of mixture components.

Two problems:

1) with given sample $X^m \sim p(x)$, number k and function φ estimate parameter vector $\Theta = (w_1, \dots, w_k, \theta_1, \dots, \theta_k)$.
2) find k.

Solving problems

We know how to solve such problems:

by maximizing logarithm of likelihood

$$L(\Theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j\left(x_i; \theta_j\right) \to \max_{\Theta}.$$

What is a problem then?

Solving problems

We know how to solve such problems:

by maximizing logarithm of likelihood

$$L(\Theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i; \theta_j) \rightarrow \max_{\Theta}.$$

It is unclear what to do with logarithm of sum, therefore we cannot find the analytical solution.

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM

EM algorithm idea

Main idea: add hidden variables, such that:

1) they can be expressed with Θ;
2) they can help to split the sum.

$$p(X, H|\Theta) = \prod_{i=1}^{k} p(X|H, \Theta)p(H|\Theta)$$

EM algorithm scheme

EM-algorithm is reiteration of the two steps:

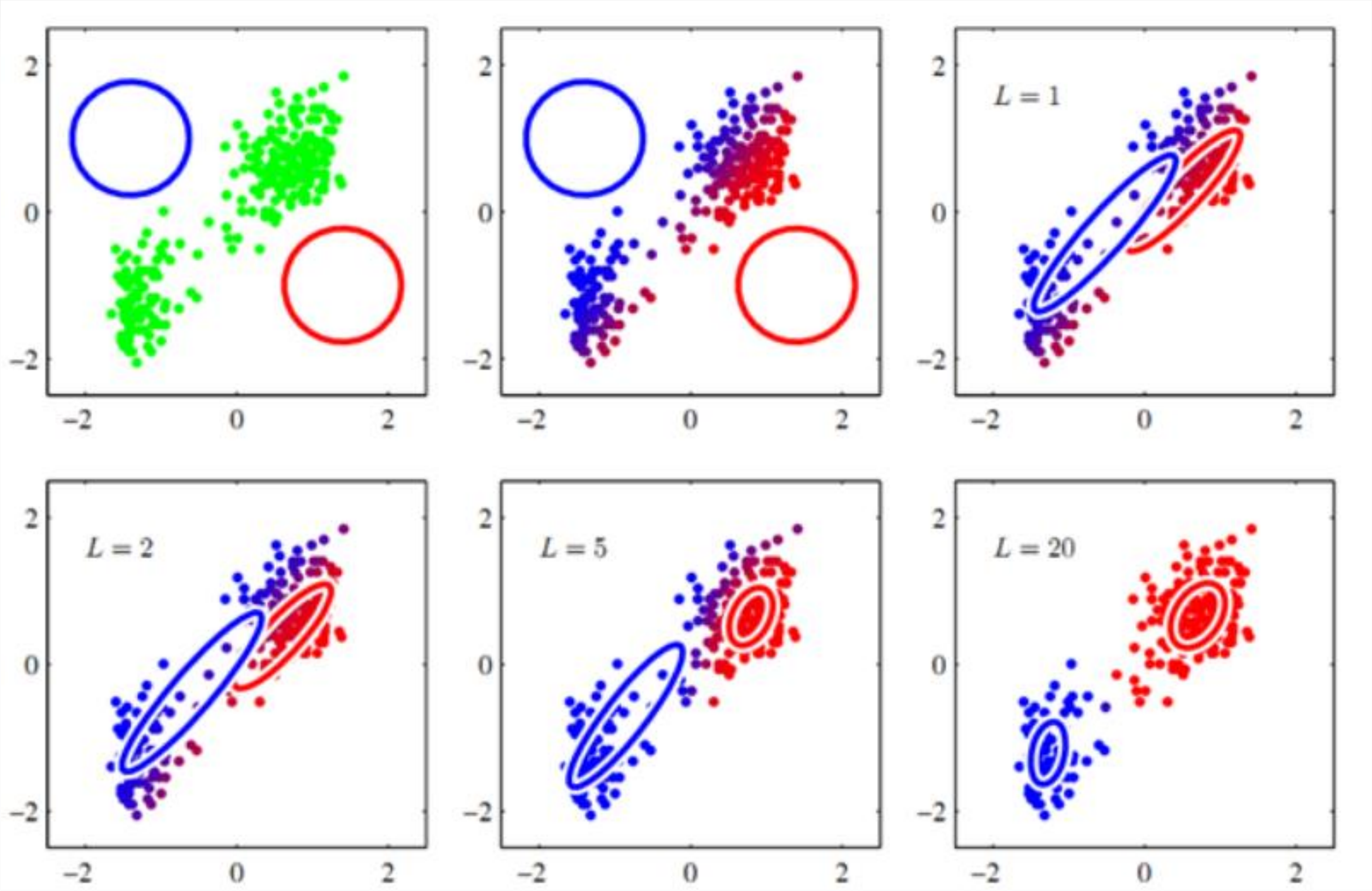
$H \leftarrow \text{E-STEP}(\Theta)$ (expectation):

finding the most probable values of the hidden variables

$\Theta \leftarrow \text{M-STEP}(H, \Theta)$ (maximization):

finding the most probable parameters given the hidden variables values

Example (Gaussians)



E-STEP

$$p(x_i, \theta_j) = p(x) \Pr(\theta_j | x) = w_j p_j(x)$$

Hidden variables $H = \left(h_{ij}\right)_{m \times k}$,

where $h_{ij} = \Pr(\theta_j | x_i)$ are degrees of how likely x_i belongs to the jth component:

$$h_{ij} = \frac{w_j p_j(x_i)}{p(x_i)} = \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)},$$

$$\sum_{j=1}^{k} h_{ij} = 1.$$

M-STEP

Theorem

If hidden variables are known, then the problem of minimizing $Q(\Theta)$ can be reduced to k independent subproblems

$$\theta_j = \operatorname{argmax}_\theta \sum_{i=1}^{m} h_{ij} \ln \varphi\,(x_i, \theta),$$

and optimal weights are equal to

$$w_i = \frac{1}{m}\sum_{i=1}^{m} h_{ij}.$$

We will maximize θ_j.

Expectation minimization

Input: $X^m, k, \Theta^{[0]}$

1. Repeat

2. **E-step**: for all $i = 1, \dots, m; j = 1, \dots, k$

$$h_{ij} = \frac{w_j \varphi(x_i; \theta_j)}{\sum_{s=1}^{k} w_s \varphi(x_i; \theta_j)};$$

3. **M-step**: for all $j = 1, \dots, k$

$$\theta_j = \operatorname{argmax}_\theta \sum_{i=1}^{m} h_{ij} \ln \varphi(x_i, \theta); w_j = \frac{1}{m} \sum_{i=1}^{m} h_{ij};$$

4. Until a **stopping criterion** is satisfied

Return $\Theta = \left(\theta_j^{[L]}, w_j^{[L]}\right)_{j=1}^{k}$.

Algorithm discussion

Advantages:

- Converges in many situations
- Can be easily turned to be insensitive to noise
- Most flexible approach

Questions:

1. When to stop?
2. How to accelerate convergence?
3. How to choose an initial approximation?
4. How to choose k?

Some answers (1/2)

1. When to stop?

Until the result do not stabilize. It is recommended to do it with respect to g:

$$\max_{i,j}\left|h_{ij} - h_{ij}^{[0]}\right| > \delta_1$$

$$\max_i \sum_j |h_{ij}^{[t]} - h_{ij}^{[t-1]}| > \delta_2$$

$$\dots$$

2. How to accelerate convergence?

Accelerate M-step.

Some answers (2/2)

3. How to choose an initial approximation?
- Uniformly.
- Choose from distant point neighborhoods.

…

4. How to choose k?
- Iteratively check for each k.
- Check for some values of k and recover the plot.

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM

Improvements of EM

- **Changing number of components**

try to add or to delete components

- **Generalized EM-algorithm (GEM)**

do not try to find a good solution of M-step

- **Stochastic EM-algorithm (SEM)**

try to find the maximum of unweighted likelihood on M-step

- **Hierarchical EM-algorithm (HEM)**

try to split “bad” components