Lecture 13

EM + PCA

Machine Learning

Andrey Filchenkov

15.05.2018

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM
- Dimensionality reduction
- PCA
- EM PCA

- 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
- Dimensionality reduction
- PCA
- EM PCA

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 $w_j > 0, \sum_{j=1}^{k} w_j = 1; p_j(x) = \varphi(x; \theta_j)$ 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

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

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
- Dimensionality reduction
- PCA
- EM PCA

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:

$H \leftarrow$E-STEP(Θ) (expectation):

finding the most probable values of the hidden variables

$\Theta \leftarrow$M-STEP(H, Θ) (maximization):

finding the most probable parameters given the hidden variables values

E-STEP

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

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

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 $i = 1, \dots, m; j = 1, \dots, k$

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

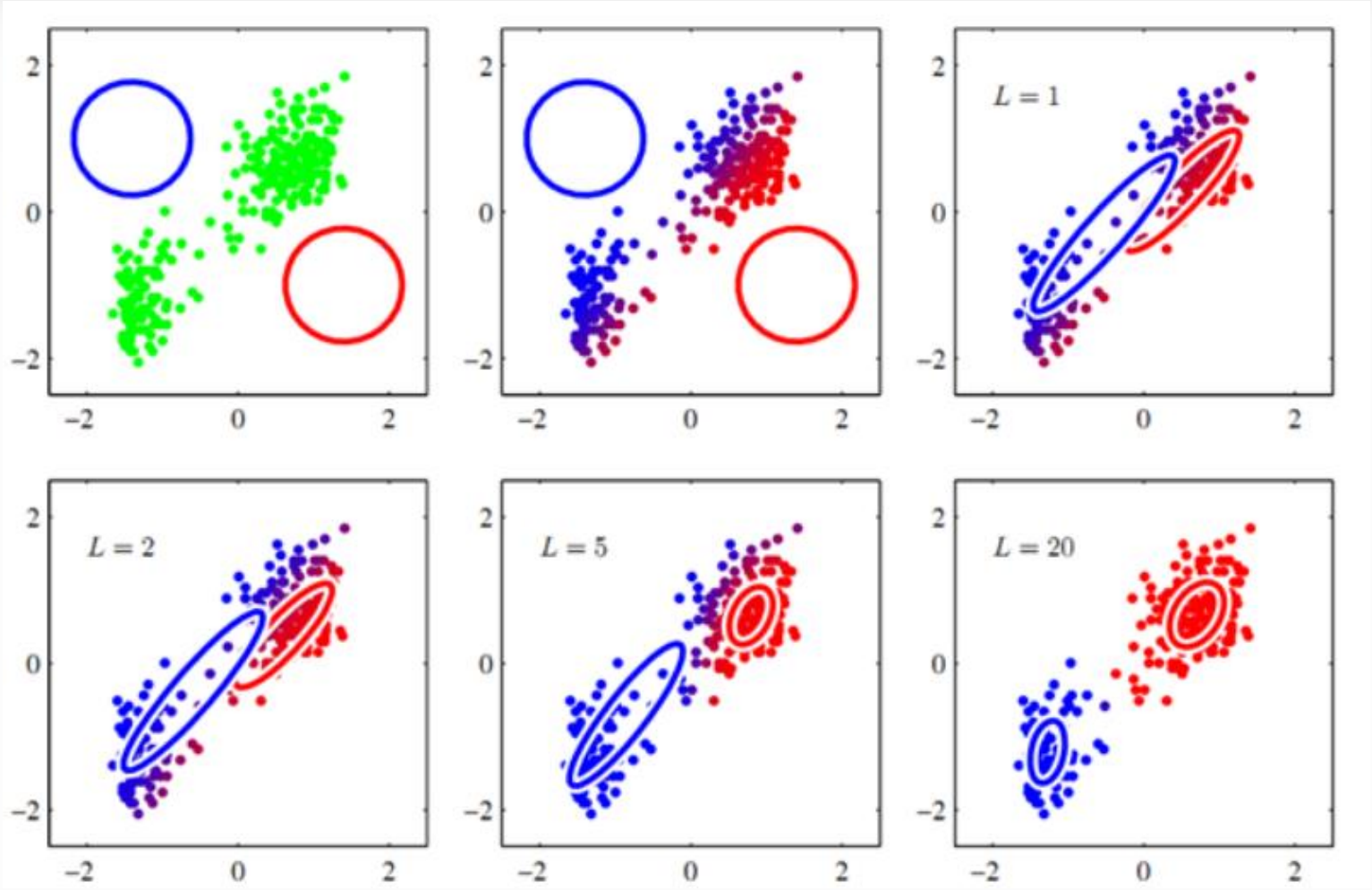
3. **M-step**: for all $j = 1, \dots, k$

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

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

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

Example (Gaussians)



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:

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

$$\max_{i} \sum_{j} |h_{ij}^{[t]} - h_{ij}^{[t-1]}| > \delta_2$$

…

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
- Dimensionality reduction
- PCA
- EM PCA

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

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM
- Dimensionality reduction
- PCA
- EM PCA

Dimensionality reduction problem

Objects in given dataset is described with features $\mathcal{F} = (f_1, \dots, f_n)$.

Problem: find a new set of features $\mathcal{G} = (g_1, \dots, g_k)$: $k < n$ (often $k \ll n$) with the minimal information loss.

- Processing algorithm speedup
- Noise filtering and multicollinearity
- Getting more understandable and visualize data

Two main approaches

Feature selection are set of approaches where $\mathcal{G} \subset \mathcal{F}$. They

- work fast;
- cannot find more complex features which describes data very well.

Feature extraction are all the other method (including such that $k > n$).

- work longer in general;
- complex features can by synthesized.

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM
- Dimensionality reduction
- PCA
- EM PCA

Case of one dimension

Given a set of points in R^n. We want to describe it with a single feature.

It is unsupervised regression problem.

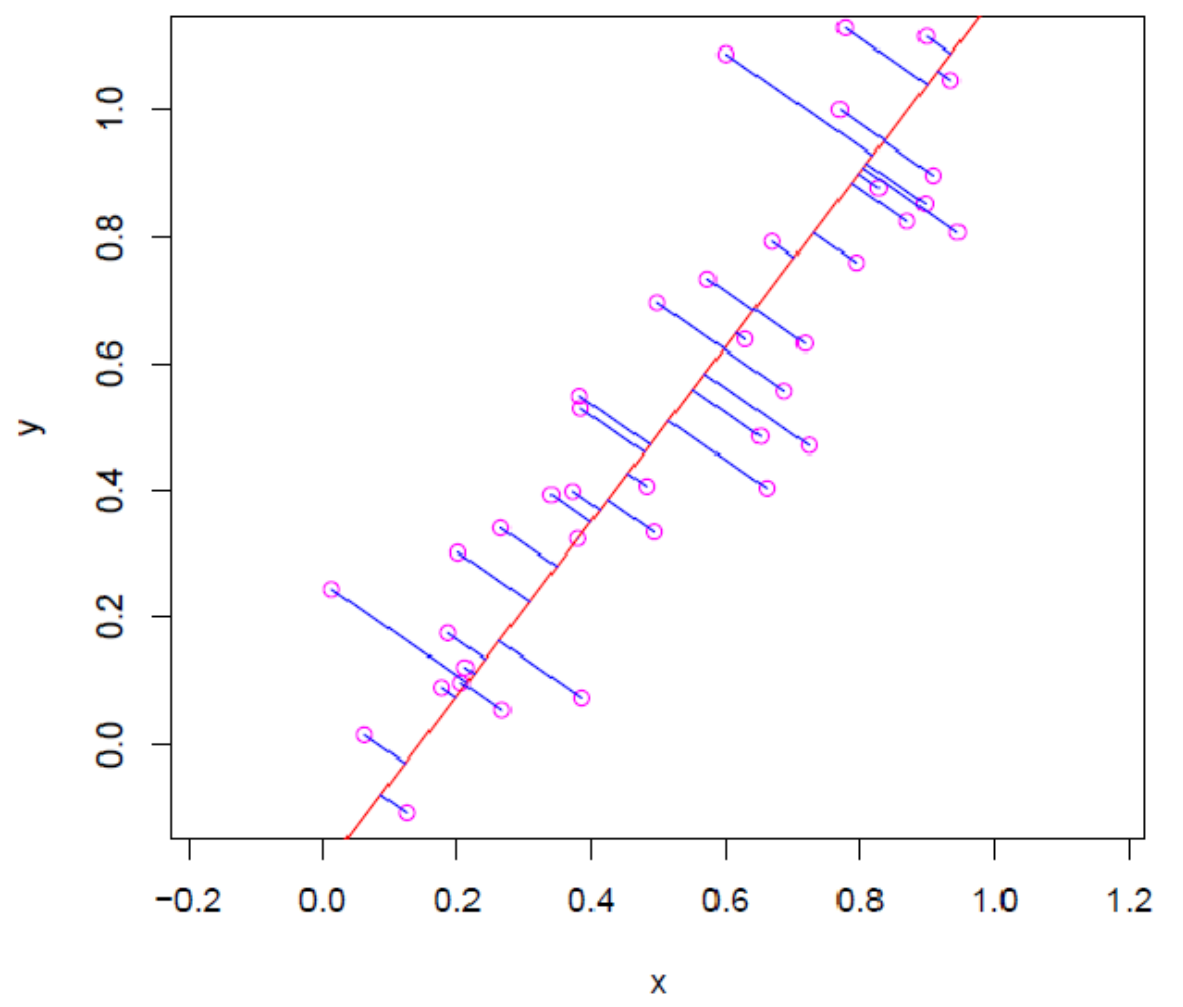
Main idea: find a line, such as

1) the distance to it is minimal;

2) dispersion of projection is maximal.

These requirements are equal.

Illustration



General case

Approximate data by linear manifolds of lower dimensions:

- minimization of the distance
- maximization of project dispersion
- maximization of distances between projections
- correlation between projection axes is zero (new!)

Many variants how solution can be expressed

Mathematical view

Problem statement:

$$\left\|GU^{\top} - F\right\|^2 \to \min_{G,U},$$

where $F = \begin{pmatrix} f_1(x_1) & \dots & f_1(x_\ell) \\ \dots & \dots & \dots \\ f_n(x_1) & \dots & f_n(x_\ell) \end{pmatrix}$, $G = \begin{pmatrix} g_1(x_1) & \dots & g_1(x_\ell) \\ \dots & \dots & \dots \\ g_k(x_1) & \dots & g_k(x_\ell) \end{pmatrix}$,

$\operatorname{rank}(U) = \operatorname{rank}(G) = k.$

Main theorem

Theorem:
if $k \leq \text{rank}(F)$, then the minimum is reached when rows of U are eigenvectors of $F^\top F$ which corresponds to the k maximal eigenvalues, and $G = UF$.

Sequences:

1. $U^\top U = \mathbf{E}_r$.
2. $G^\top G = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$.
3. $U\Lambda = F^\top F U; G\Lambda = F^\top F G$.
4. $\left\| GU^\top - F \right\|^2 = \|F\|^2 - \text{tr}\,\Lambda = \sum_{r+1}^{n} \lambda_i$.

Principal components

G presents a linear manifold.
Axes of G are **principal components**.

Iterative search of the principal components:

find c_1 which is a line, minimizing distance to it.

Repeat: find c_i which is orthogonal to $\{c_j\}_{j=1}^{i-1}$ and minimizing distance to it.

Choosing k

The problem is similar to choose of k in EM.

Sort eigenvalues of $F^\top F$ decreasingly: $\lambda_{[1]} \geq \cdots \geq \lambda_{[n]}$.

$$E(k) = \frac{\left\|F - UG^\top\right\|^2}{\|F\|^2} = \frac{\lambda_{[k+1]} + \cdots + \lambda_{[n]}}{\lambda_{[1]} + \cdots + \lambda_{[n]}}$$

$E(k)$ characterize the part of information which is lost after a projection.

We can choose k using $E(k)$.

PCA analysis

- Linear transformation with all its pros and cons
- Works quit long
- Is wildly used for data compression and visualization
- Improvement: non-linear methods (principal curves and principal manifolds).

Lecture plan

- Mixture of distributions
- EM algorithm
- Improvements of EM
- Dimensionality reduction
- EM PCA

PCA in EM manner

- Joint distribution

$$p(F, G|\Theta) = \prod_{i=1}^{m} p(x_i|g_i, \Theta)p(g_i|\Theta)$$

- Θ is $n \times k$ matrix V and a scalar parameter
- We can use EM algorithm to find

$$\arg\max_{H} p(\left\|GU^\top - F\right\|^2|\Theta)$$

Advantages

- EM updates have complexity $O(kmn)$ instead of $O(mn^2)$
- Can process missing parts in F and present parts in Θ
- Can determine k if $p(\Theta)$ established
- Can be extended to mixture model

Mixture model

- Joint distribution

$$p(F, H, G|\Theta) = \prod_{i=1}^{m} p(x_i|g_i, h_i, \Theta)p(g_i|\Theta)p(h_i|\Theta)$$

- Θ consists of matrices $\{V_k\}$ and scalars and hidden variable a priory probabilities