Guided Tour of Machine Learning in Finance

Week 3: Unsupervised Learning

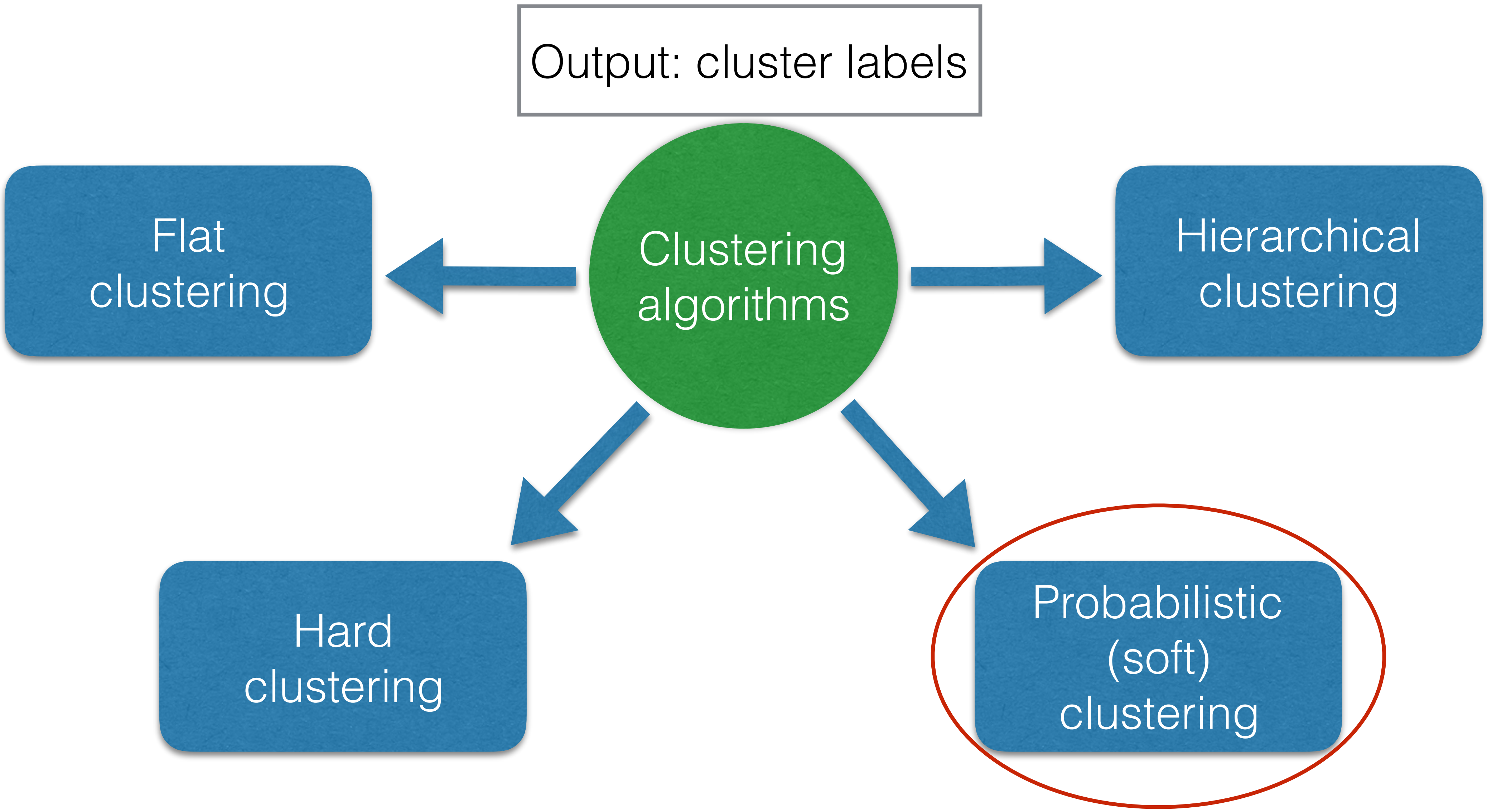
Probabilistic clustering

Igor Halperin

NYU Tandon School of Engineering, 2017

Types of clustering

Now consider Soft (Probabilistic) clustering methods



Mixtures of Gaussians

Mixture models are used to describe complex distributions.
A mixture of Gaussians is a popular common choice:

$$p(y \mid \Theta) = \sum_{k-1}^{K} \pi_k \, p(y \mid \theta_k)$$

Here each component is a Gaussian with $\theta_k = (\mu_k, \Sigma_k)$ for its mean and variance.
The component weight is $0 \leq \pi_k \leq 1$ with

$$\sum_{k-1}^{K} \pi_k = 1$$

s_i

Mixtures of Gaussians

Mixture models are used to describe complex distributions.
A mixture of Gaussians is a popular common choice:

$$p(y \mid \Theta) = \sum_{k-1}^{K} \pi_k \, p(y \mid \theta_k)$$

Here each component is a Gaussian with $\theta_k = (\mu_k, \Sigma_k)$ for its mean and variance.
The component weight is $0 \le \pi_k \le 1$ with

$$\sum_{k-1}^{K} \pi_k = 1$$

s_i

Weights can be described by a latent (hidden) variable s such that $s = k$ if the data point was generated by component k. Then we can write it as

$$p(y \mid \Theta) = \sum_{k-1}^{K} P(s = k \mid \pi) \, p(y \mid s = k, \theta)$$

Estimation of model now reduces to estimation of parameters $\theta_k = (\mu_k, \Sigma_k)$, as well as inference of the hidden variable s

Mixtures of Gaussians

Mixture models are used to describe complex distributions.
A mixture of Gaussians is a popular common choice:

$$p(y \mid \Theta) = \sum_{k-1}^{K} \pi_k \, p(y \mid \theta_k)$$

Here each component is a Gaussian with $\theta_k = (\mu_k, \Sigma_k)$ for its mean and variance.
The component weight is $0 \le \pi_k \le 1$ with

$$\sum_{k-1}^{K} \pi_k = 1$$

s_i

Weights can be described by a latent (hidden) variable s such that $s = k$ if the data point was generated by component k. Then we can write it as

$$p(y \mid \Theta) = \sum_{k-1}^{K} P(s = k \mid \pi) \, p(y \mid s = k, \theta)$$

Estimation of model now reduces to estimation of parameters $\theta_k = (\mu_k, \Sigma_k)$, as well as inference of the hidden variable s

This can be done using the **EM algorithm**!

The EM algorithm

The EM algorithm = Expectation Maximization.
Let $p(y,x\,|\,\theta)$ be the joint probability of observed data y and latent variables x
We have for the likelihood of data

$$L(\theta) = \log p(y\,|\,\theta) = \log \int p(x,y\,|\,\theta)\,dx = \log \int q(x) \frac{p(x,y\,|\,\theta)}{q(x)} dx$$

$$\geq \int q(x) \log \frac{p(x,y\,|\,\theta)}{q(x)} dx \triangleq F(q,\theta)$$

s_i

Here $q(x)$ is some arbitrary density over latent variables, and the last inequality is due to concavity of the logarithm (Jensen's inequality)

The EM algorithm: iterate between maximization of this low bound on the log-likelihood as a function of q, and as a function of θ

The EM algorithm

The EM algorithm = Expectation Maximization.

E-step: Optimize wrt $q(x)$ while keeping θ fixed:

$$q_k(x) = \arg\max_{q(x)} \int q(x) \frac{p(x,y \mid \theta_{k-1})}{q(x)} dx = \arg\max_{q(x)} \left[\log p(y \mid \theta_{k-1}) + \int q(x) \frac{p(x \mid y, \theta_{k-1})}{q(x)} dx \right]$$

$$\Rightarrow q_k(x) = p(x \mid y, \theta_{k-1})$$

The EM algorithm

The EM algorithm = Expectation Maximization.
E-step: Optimize wrt $q(x)$ while keeping θ fixed:

$$q_k(x) = \arg\max_{q(x)} \int q(x) \frac{p(x,y \mid \theta_{k-1})}{q(x)} dx = \arg\max_{q(x)} \left[\log p(y \mid \theta_{k-1}) + \int q(x) \frac{p(x \mid y, \theta_{k-1})}{q(x)} dx \right]$$

$$\Rightarrow q_k(x) = p(x \mid y, \theta_{k-1})$$

M-step: Optimize wrt θ, while holding the distribution $q(x)$ fixed:

$$\theta_k = \arg\max_{\theta} \int q_k(x) \frac{p(x,y \mid \theta)}{q(x)} dx$$

$$\Rightarrow \theta_k = \arg\max_{\theta} \int q(x) \log p(x,y \mid \theta) dx$$

The EM algorithm

The EM algorithm = Expectation Maximization.
E-step: Optimize wrt $q(x)$ while keeping θ fixed:

$$q_k(x) = \arg\max_{q(x)} \int q(x) \frac{p(x,y|\theta_{k-1})}{q(x)} dx = \arg\max_{q(x)} \left[\log p(y|\theta_{k-1}) + \int q(x) \frac{p(x|y,\theta_{k-1})}{q(x)} dx \right]$$

$$\Rightarrow q_k(x) = p(x|y,\theta_{k-1})$$

M-step: Optimize wrt θ, while holding the distribution $q(x)$ fixed:

$$\theta_k = \arg\max_{\theta} \int q_k(x) \frac{p(x,y|\theta)}{q(x)} dx$$

$$\Rightarrow \theta_k = \arg\max_{\theta} \int q(x) \log p(x,y|\theta) dx$$

- The EM algorithm is guaranteed to increase the likelihood or keep it constant at each iteration, and to find a local maximum of the log-likelihood.
- The EM algorithm can be applied to any model with hidden variables: Gaussian mixtures, Factor models, probabilistic PCA etc.
- For Gaussian mixtures, $q(x)$ reduces to a discrete distribution $P(s=k|\pi)$ over the mixture components.

Control question

Select all correct answers

1. Probabilistic Clustering focuses on data points that have reasonably high probability to be present in the data, and ignores the rest.
2. The Expectation Maximization algorithm expects that all data in a dataset is unlabelled, and maximizes this expectation by removing all labels, even when they are present in the data.
3. A Hard (non-probabilistic) Clustering can be obtained as a deterministic limit of Probabilistic Clustering.
4. The EM algorithm iterates between two steps: the E-step eliminates improbable points, and the M-step maximizes the probability to see the rest of points.
5. The EM algorithm iterates between two steps: the E-step estimates the probability distribution on hidden variables with model parameters fixed, and the M-step maximizes the low bound on the log-likelihood by adjusting model parameters while keeping the distribution over hidden variables fixed.

Correct answers: 3, 5