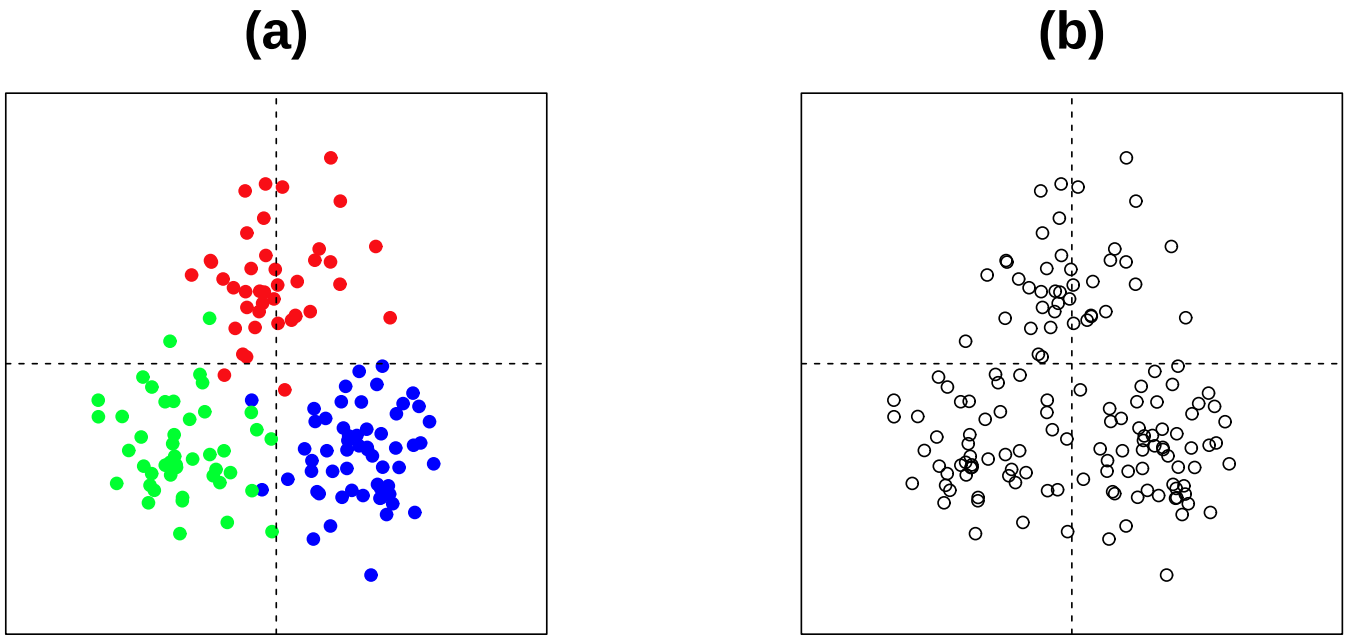
Lecture 3 – Part 1
Unsupervised Learning

(January 23, 2015)
with revisions after lecture

Mu Zhu
University of Waterloo



(a) data from three groups; (b) group label unobserved

Clustering

- unsupervised classification
- $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^d$
- given $y_i = k$, $\boldsymbol{x}_i \sim p_k(\boldsymbol{x})$, for $k = 1, 2, ..., K$
- but class label $y_i \in \{1, 2, ..., K\}$ unobserved
- often, assumes $p_k(\cdot) = p(\cdot; \boldsymbol{\theta}_k)$ — same family, different parameters

Latent Variables

- suppose K is known
- introduce latent variables

$$z_{ik} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \text{ belongs to class } k; \\ 0, & \text{otherwise} \end{cases}$$

with $\pi_k \equiv \mathbb{P}(z_{ik} = 1)$, where $\sum \pi_k = 1$

Likelihood

- parameters

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k, \pi_k : \quad k = 1, 2, ..., K\}$$

- observed data

$$\boldsymbol{X} = \{\boldsymbol{x}_i : \quad i = 1, 2, ..., n\}$$

- unobserved "data"

$$\boldsymbol{Z} = \{z_{ik} : \quad i = 1, 2, ..., n; \quad k = 1, 2, ..., K\}$$

- likelihood

$$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) \propto \prod_i \prod_k \left[p(\boldsymbol{x}_i; \boldsymbol{\theta}_k)\right]^{z_{ik}} \times \pi_k^{z_{ik}}$$

EM Algorithm

- log-likelihood (aside from constant)

$$\ell(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_i \sum_k z_{ik} \log\left[p(\boldsymbol{x}_i; \boldsymbol{\theta}_k)\right] + \sum_i \sum_k z_{ik} \log(\pi_k)$$

- would like to maximize $\ell(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z})$ over $\boldsymbol{\Theta}$ but $\boldsymbol{Z}$ missing

- idea:
  - given current estimate $\widehat{\boldsymbol{\Theta}}$, replace $\boldsymbol{Z}$ with $\widetilde{\boldsymbol{Z}} = \mathbb{E}(\boldsymbol{Z}|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}})$
  - given $\widetilde{\boldsymbol{Z}}$, re-estimate $\boldsymbol{\Theta}$ by maximizing over $\ell(\boldsymbol{\Theta}; \boldsymbol{X}, \widetilde{\boldsymbol{Z}})$
  - iterate

EM Algorithm

- E-Step: compute $\mathbb{E}(z_{ik}|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}})$

- M-Step: maximize

$$Q(\boldsymbol{\Theta}; \boldsymbol{X}) = \sum_i \sum_k \mathbb{E}(z_{ik}|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}}) \log\left[p(\boldsymbol{x}_i; \boldsymbol{\theta}_k)\right] + \sum_i \sum_k \mathbb{E}(z_{ik}|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}}) \log(\pi_k)$$

over $\boldsymbol{\Theta}$

M-Step

- write

$$w_{ik} \equiv \mathbb{E}(z_{ik}|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}})$$

- for the "$\boldsymbol{\theta}_k$ part", just **weighted** maximum likelihood

$$\max_{\boldsymbol{\theta}_k} \quad \sum_i w_{ik} \log\left[p(\boldsymbol{x}_i; \boldsymbol{\theta}_k)\right] \quad \forall \quad k \tag{1}$$

- for the "π_k part", usual multinomial likelihood with w_{ik}

$$\max_{\pi_1,\ldots,\pi_K} \quad \sum_i \sum_k w_{ik} \log(\pi_k) \quad \text{s.t.} \quad \sum_k \pi_k = 1 \tag{2}$$

M-Step

Exercise

(a) Solve (1) for the case of $p_k \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.

(b) Solve (2).

(c) What would happen if we modelled each p_k as $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ instead? (Hard)

(d) What can we do in order to strike a compromise between (a) and (c)? [See, e.g., Fraley & Raftery (2002; *JASA*).]

E-Step

Finally,

$$\begin{aligned} w_{ik} &\equiv \mathbb{E}(z_{ik}|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}}) \\ &= \mathbb{P}(z_{ik}=1|\boldsymbol{x}_i, \widehat{\boldsymbol{\Theta}}) \\ &= \frac{\widehat{\pi}_k p(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_k)}{\widehat{\pi}_1 p(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_1) + ... + \widehat{\pi}_K p(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_K)}. \end{aligned}$$

Remark The denominator above shows the EM algorithm has just allowed us to fit a mixture distribution,

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}; \boldsymbol{\theta}_k),$$

to the sample $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. Question: How to cluster each $\boldsymbol{x}_i$?

K-Means as a Special Case

- assume $p_k(\boldsymbol{x}) \sim \mathrm{N}(\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})$

- E-Step: in addition to computing the "soft" parameter w_{ik}, also make a "hard" decision

$$\widehat{z}_{ik} = \begin{cases} 1, & \text{if } w_{ik} = \max\limits_{k'} w_{ik'} \\ 0, & \text{otherwise} \end{cases}$$

- M-Step: do the same with $\widehat{z}_{ik}$ in place of w_{ik}

Exercise The use of $\widehat{z}_{ik}$ in place of w_{ik} "happens automatically" if we let $\sigma^2 \to 0$. (Why is this algorithm called K-means?)

Variation for Text Data

- each $\boldsymbol{x}_i = (x_{i1}, ..., x_{id})^{\mathrm{T}}$ a text document
- x_{ij} = count of word j in document i (bag of words)
- want to cluster $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into K groups
- can use the same framework, except choose

$$p_k(\boldsymbol{x}_i) \sim \text{multinomial}(m_i; \boldsymbol{\theta}_k),$$

  where $\boldsymbol{\theta}_k = (\theta_{k1}, ..., \theta_{kd})^{\mathrm{T}}$, and m_i = total number of words in document i (treated as fixed constants)

- many sophisticated extensions, e.g., latent Dirichlet allocation (Blei, Ng & Jordon, 2003; *JMLR*)

Latent Dirichlet Allocation

- instead of modelling each document ($\boldsymbol{x}_i$) to be a mixture

$$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{X}) = \prod_i \left[\sum_k \pi_k p(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \right],$$

  model each word ($x_{it} \in \{1, 2, ..., d\}$) to be a mixture, and allow each document to have different mixing proportions

$$\mathcal{L} = \prod_i \prod_t \left[\sum_k \pi_{ik} p(x_{it}; \boldsymbol{\theta}_k) \right], \quad p(x_{it}; \boldsymbol{\theta}_k) = \prod_{j=1}^{d} \theta_{kj}^{I(x_{it}=j)}$$

- instead of EM, put (Dirichlet) prior distributions on

$$\underline{\text{both}} \quad (\pi_{i1}, ..., \pi_{iK})^{\mathrm{T}} \quad \underline{\text{and}} \quad (\theta_{k1}, ..., \theta_{kd})^{\mathrm{T}}$$

  and use a Bayesian approach

Empirical Bayes

- data:

$$x_1, \quad x_2, \quad ..., \quad x_n$$

- model:

$$x_i \sim f(\cdot|\theta_i), \quad x_1, ..., x_n \quad \text{all independent}$$

- source of information for each θ_i:

$$x_i \quad \underline{\text{alone}}$$

Empirical Bayes

- now, put (common) prior on each θ_i:

$$\theta_i \quad \sim \quad \pi(\cdot|\psi),$$

  where ψ is the parameter for the prior

- then, posterior of θ_i given data is

$$p(\theta_i|x_i,\psi) = \frac{f(x_i|\theta_i)\pi(\theta_i|\psi)}{\int f(x_i|\theta_i)\pi(\theta_i|\psi)d\theta_i}$$

- source of information for each θ_i:

$$\underline{\text{both}} \quad x_i \quad \underline{\text{and}} \quad \psi$$

- next, consider ψ unknown

Empirical Bayes

- in Bayesian framework, $f(x_i|\theta_i)$ is a conditional distribution (of x_i given θ_i)
- can integrate out θ_i to get the marginal distribution of x_i,

$$m(x_i|\psi) = \int f(x_i|\theta_i)\pi(\theta_i|\psi)d\theta_i$$

- $x_1, ..., x_n \overset{iid}{\sim} m(\cdot|\psi)$ all contain some information about ψ
- can find an estimate $\widehat{\psi}$, using all x_i, and use the "plug-in" posterior, $p(\theta_i|x_i, \widehat{\psi})$
- source of information for each θ_i [based on $p(\theta_i|x_i, \widehat{\psi})$]:

all $x_1, ..., x_n$ [not just x_i; borrow strength]

Ex: James-Stein Estimator

Set-up Suppose

$$\begin{aligned} x_i &\sim \mathrm{N}(\theta_i, \sigma^2) \quad \text{all independent;} \\ \theta_i &\sim \mathrm{N}(\mu, \tau^2), \end{aligned}$$

where σ^2 (but not τ^2) is assumed to be known.

Exercise The posterior of θ_i is normal with mean

$$\left[\frac{\tau^2}{\sigma^2+\tau^2}\right] x_i + \left[\frac{\sigma^2}{\sigma^2+\tau^2}\right] \mu = \mu + \left[1 - \frac{\sigma^2}{\sigma^2+\tau^2}\right] (x_i - \mu),$$

which can be used as a (Bayesian) point estimate of θ_i (if μ, τ^2 are known).

Ex: James-Stein Estimator

Empirical estimate of μ and $1/(\sigma^2+\tau^2)$ Marginally,

$$x_i|\mu \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2+\tau^2),$$

so we can estimate μ by

$$\widehat{\mu} = \bar{x} = \frac{x_1 + ... + x_n}{n}.$$

Exercise Show that, if $x_1, ..., x_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2+\tau^2)$, then

$$\mathbb{E}\left[\frac{n-3}{\sum(x_i-\bar{x})^2}\right] = \frac{1}{\sigma^2+\tau^2}.$$

Ex: James-Stein Estimator

The J-S estimator

$$\widehat{\theta}_i^{JS} = \bar{x} + \left[1 - \frac{(n-3)\sigma^2}{\sum(x_i - \bar{x})^2}\right](x_i - \bar{x})$$

has been shown to have smaller mean-squared error than $\widehat{\theta}_i = x_i$ for $n > 3$.

Hierarchical Bayes

- alternatively, put prior on unknown ψ (another layer), e.g.,

$$\begin{aligned} x_i &\sim f(\cdot|\theta_i); \\ \theta_i &\sim \pi(\cdot|\psi); \\ \psi &\sim \xi(\cdot) \end{aligned}$$

- here, we have suppressed any parameters for ξ — eventually, *something* must be treated as fixed

- instead of point estimate $\widehat{\psi}$, rely on posterior of ψ

- posterior of ψ depends on all x_i ... still borrow strength

Some Challenges

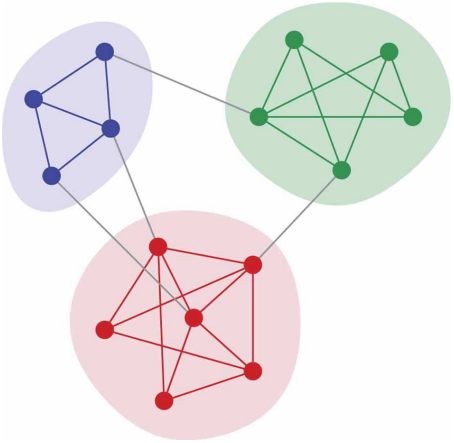
- for EM
  - many local solutions in practice
- in reality, K unknown
  - use a different prior for $(\pi_1, ..., \pi_K)^{\mathrm{T}}$
  - e.g., Chinese restaurant process, etc
- for Bayesian approach
  - effective computational techniques
  - e.g., collapsed Gibbs sampler, variational inference, etc

Summary

- key ideas:
  - unsupervised learning; clustering
  - latent variables; mixture models
  - empirical Bayes; hierarchical Bayes; borrow strength
- specific methods:
  - EM algorithm
  - K-means
  - bag of words; latent Dirichlet allocation
  - James-Stein estimator

Next ...

- lecture @ 2 pm by Professor E. Kolaczyk on network data
- a short, 10-minute break
- current research with collaborators on clustering nodes in a transactional network (e.g., an NBA basketball game)



Community detection in networks. [Image source: Newman (2012), *Nature Physics* **8**, pp. 25–31.]