Unsupervised Learning

Learning the parametric forms

Unsupervised Learning

- Samples are not labeled
  - Labeling can be very expensive
  - Data mining & pattern discovery
  - Adapting to time varying behaviors
  - Insight into the problem domains

Learning Parametric Forms

- In the supervised learning, we can do one of the three things
  - Parametric Estimation
    - Assume a particular parametric form
  - Nonparametric estimation
    - No particular parametric form
  - Discriminant function
    - Decision boundary
- We can do similar things for unsupervised learning

Learning Density Function

- The simplest way – just collect samples and put them in the *d*-dimensional (*d*: # of features) collection bins, without regard to where they come from
- The same old technique applies here
- A description of the *mixture*, not the *components*

Learning Parametric Forms

- Samples are not labeled but follow a particular distribution
  - For simplicity, we will assume Gaussian
- Might not know
  - How many Gaussian
  - What are the priors
  - What are the means
  - What are the variances

Difficulty

- With labeled samples
  - Separate learning into *c* identical problems – that of learning the mean and variance of each class
- With unlabeled samples
  - Separate learning is not possible, a sample may come from one of the *c* classes (we might not even know how many!)
  - Dealing with *mixture densities*
- The problem is harder and may not even have a solution

Mixture Density

$$p(x \mid \theta) = \sum_{j=1}^{c} p(x \mid w_j, \theta_j) P(w_j)$$

$$\theta = (\theta_1, \theta_2, \cdots, \theta_c)$$

- May not know
  - # of classes
  - Class priors
  - Form
  - Mean
  - Variance
  - …
- E.g., will assume we know
  - # of classes
  - Class priors
  - Form (Gaussian)
  - Variance
- Do not know
  - mean

Identifiably

- Can we actually do anything about this?
- $p(x|\theta)$ is *identifiable* if there exists an x such that $p(x|\theta) \mathrel{!}= p(x|\theta')$ if $\theta \mathrel{!}= \theta'$
  - *I.e.*, you can at least make observations that show different behaviors

Example

- A discrete, binary distribution where $x=0,1$ from a mixture distribution
- Samples can be used to estimate
  - $p(x|\theta)=1$ & $p(x|\theta)=0$
- But all we can say is about $\theta_1+\theta_2$

$$p(x\,|\,\theta)=\frac{1}{2}\theta_1^x(1-\theta_1)^{1-x}+\frac{1}{2}\theta_2^x(1-\theta_2)^{1-x}$$

$$=\begin{cases}\frac{1}{2}(\theta_1+\theta_2) & x=1\\ 1-\frac{1}{2}(\theta_1+\theta_2) & x=0\end{cases}$$

- Two coins
- With θ_1 and θ_2 probability of head
- Randomly choose one to perform the experiment (equal chance for two)
- Register the outcome

Caveats

- Any discrete probability where the number of states of nature are less than free variables (*more variables than constraints*) is completely unidentifiable
  - E.g., three coins and two outcomes (head and tail) is completely unidentifiable
- In general, parametric estimation is interesting mainly from a theoretical point of view (i.e., if you are a mathematician ☺)
- Our discussion here necessarily will be very brief and limited

Maximum-Likelihood Estimates

- We will illustrate this using examples
  - For mixture of Gaussians

case	μ_i	Σ_i	$P(\varpi_i)$	c
1	?	*yes*	*yes*	*yes*
2	?	?	?	*yes*
3	?	?	?	?

General Formula

- Be warned: this is not pretty

$$p(D|\theta)=\prod_{k=1}^{n} p(x_k|\theta)$$

$$l=\sum_{k=1}^{n}\ln p(x_k|\theta)$$

$$\nabla_{\theta_i} l=\sum_{k=1}^{n}\frac{1}{p(x_k|\theta)}\nabla_{\theta_i} p(x_k|\theta)$$

$$=\sum_{k=1}^{n}\frac{1}{p(x_k|\theta)}\nabla_{\theta_i}\sum_{j=1}^{c} p(x_k|\varpi_j,\theta_j)P(\varpi_j)$$

$$\because \theta_i,\theta_j \text{ independent}, i\neq j$$

$$=\sum_{k=1}^{n}\frac{1}{p(x_k|\theta)}\nabla_{\theta_i} p(x_k|\varpi_i,\theta_i)P(\varpi_i)$$

$$\because P(\varpi_i|x_k,\theta)=\frac{p(x_k|\varpi_i,\theta_i)P(\varpi_i)}{p(x_k|\theta)}$$

$$=\sum_{k=1}^{n}P(\varpi_i|x_k,\theta)\frac{1}{p(x_k|\varpi_i,\theta_i)P(\varpi_i)}\nabla_{\theta_i} p(x_k|w_i,\theta_i)P(\varpi_i)$$

$$=\sum_{k=1}^{n}P(\varpi_i|x_k,\theta)\nabla_{\theta_i}\ln p(x_k|\varpi_i,\theta_i)P(\varpi_i)$$

As applied to Gaussian Case I

$$\ln p(x_k \mid \varpi_i, \mu_i) = -\ln[(2\pi)^{d/2} |\Sigma_i|^{1/2}] - \frac{1}{2}(x_k - \mu_i)^t \Sigma_i^{-1}(x_k - \mu_i)$$

$$\nabla_{u_i} \ln p(x_k \mid \varpi_i, \mu_i) P(\varpi_i) = \Sigma_i^{-1}(x_k - \mu_i)$$

$$\sum_{k=1}^{n} P(\varpi_i \mid x_k, \hat{\mu}) \Sigma_i^{-1}(x_k - \hat{\mu}_i) = 0$$

$$\hat{\mu}_i = \boxed{\frac{\sum_{k=1}^{n} P(\varpi_i \mid x_k, \hat{\mu}) x_k}{\sum_{k=1}^{n} P(\varpi_i \mid x_k, \hat{\mu})}} \qquad cf \quad \hat{u}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} x_k \text{ in supervised training}$$

- A weighted average of samples
- Weight ~ how likely is the sample in class *i*

As applied to Gaussian (cont.)

- However, how do you determine $P(\varpi_i \mid x_k, \hat{\mu})$
- If you are still awake, then that is a posterior probability and the good way to estimate it is to use Bayes rule to convert it into a prior + conditional

$$P(\varpi_i \mid x_k, \hat{\mu}) = \frac{p(x_k \mid \varpi_i, \hat{\mu}) P(\varpi_i)}{\sum_{j=1}^{c} p(x_k \mid \varpi_j, \hat{\mu}) P(\varpi_j)}$$

- Even with Gaussian assumption, this expression is hard to evaluate (no closed form solution)

As applied to Gaussian (cont.)

- Here, advanced optimization technique such as EM is used (or gradient descent is used)

$$\hat{\mu}_i^{(m+1)} = \frac{\sum_{k=1}^{n} P(\varpi_i \mid x_k, \hat{\mu}^{(m)}) x_k}{\sum_{k=1}^{n} P(\varpi_i \mid x_k, \hat{\mu}^{(m)})} \quad \text{with known } \hat{\mu}^{(0)}$$

A concrete example

$$p(x\,|\,\mu_1,\mu_2)=\frac{1}{3}\frac{1}{\sqrt{2\pi}}e^{-(x-u_1)^2}+\frac{2}{3}\frac{1}{\sqrt{2\pi}}e^{-(x-u_2)^2}$$

$$p(D\,|\,\mu_1,\mu_2)=\prod_{k=1}^{n}[\frac{1}{3}\frac{1}{\sqrt{2\pi}}e^{-(x_k-u_1)^2}+\frac{2}{3}\frac{1}{\sqrt{2\pi}}e^{-(x_k-u_2)^2}]$$

$$l=\sum_{k=1}^{n}\log[\frac{1}{3}\frac{1}{\sqrt{2\pi}}e^{-(x_k-u_1)^2}+\frac{2}{3}\frac{1}{\sqrt{2\pi}}e^{-(x_k-u_2)^2}]$$

- The landscape is fairly complicated even in this simple case
- Solution is not unique, depending on the search start point

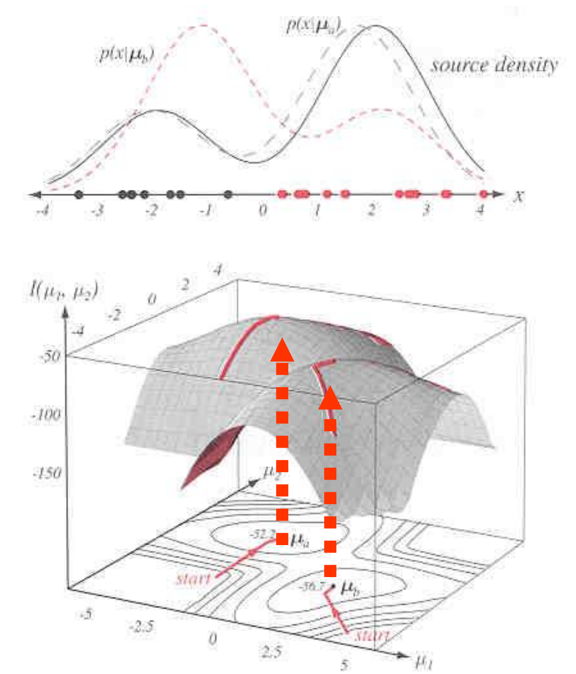


FIGURE 10.1. (Above) The source mixture density used to generate sample data, and two maximum-likelihood estimates based on the data in the table. (Bottom) Log-likelihood of a mixture model consisting of two univariate Gaussians as a function of their means, for the data in the table. Trajectories for the iterative maximum-likelihood estimation of the means of a two-Gaussian mixture model based on the data are shown as red lines. Two local optima (with log-likelihoods −52.2 and −56.7) correspond to the two density estimates shown above.

Case II

- Impossible to solve with so many parameters, theoretically
- Can make likelihood estimator arbitrarily large (e.g., by have *u* to be one of the samples and σ as zero)

$$p(x \mid \mu,\sigma) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-u)^2}{\sigma^2}} + \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

$$p(x_1(=u) \mid \mu,\sigma) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2}$$

$$p(x_i(\neq u) \mid \mu,\sigma) \geq \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2}$$

$$p(x_1, x_2, \cdots, x_n \mid \mu,\sigma) \geq (\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-x_1^2}) \frac{1}{(2\sqrt{2\pi})^{n-1}} e^{-\sum_{i=2}^{n} x_i^2}$$

What does it mean?

- To maximize probability
  - Make something very unlikely to happen
    - E.g. One of the Gaussian has very narrow spread
  - Try to fit the data to make that unlikely thing to happen
    - E.g., make a data point to coincide with the Gaussian mean
  - Then, all others notwithstanding, because of this highly unlikely event, the particular model will win

Case II (cont.)

- Pathological solutions aside, in general
  - Prior $$\hat{P}(w_i) = \frac{1}{n}\sum_{k=1}^{n}\hat{P}(w_i \mid x_k, \theta)$$
    - To be the posterior class likelihood of given samples and estimated parameters
  - Mean
    - To be the weighted average of all samples, weighted by likelihood of samples in that particular class
  - Variance
    - To be the weighted average of sample variances, weighted by likelihood of samples in that particular class

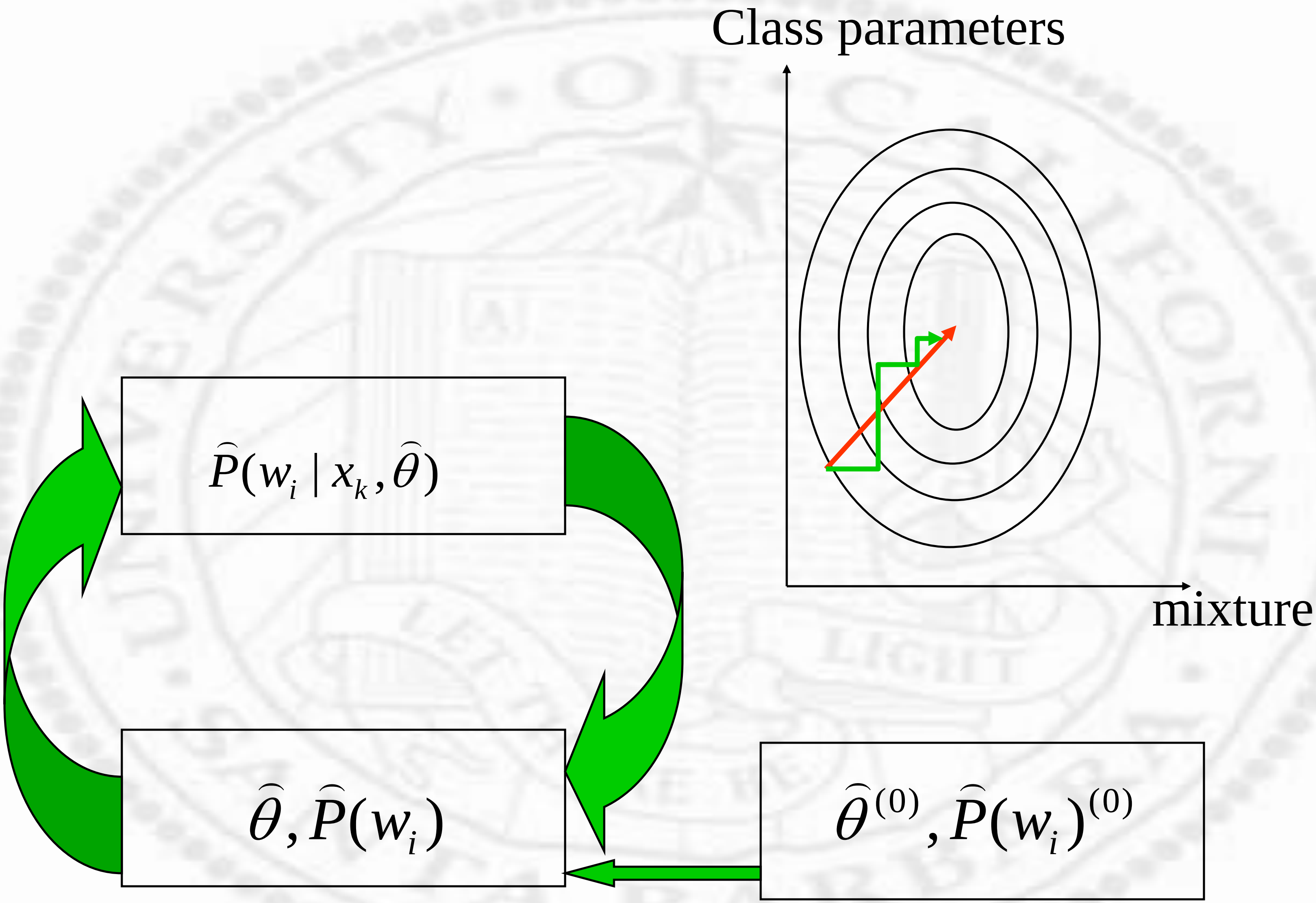
EM (Expectation & Maximization)

- An iterative algorithm (with an initial guess)
  - E stage: given the *parameters*, finding the right *mixture* *(where does each sample come from?)*
  - M stage: given the *mixtures*, finding the right *parameters* *(what is the traits of each class?)*
  - Can be considered a gradient descent technique which guarantees convergence to a local minimum
  - Global minimum requires good initial guess (or many different starting points)

Intuition

1) How do we know the traits of each class?
   - Estimate that from samples (as always)
   - Need to know which samples to use
2) How do we know where each sample come from?
   - If we know the class parameters (traits)
   - The sample comes from the class with a probability proportional to how likely is a class to generate that sample (i.e., Bayes rule)

Intution (cont)



EM Example

- The equation looks terribly complex and very confusing
- But the concept is quite easy to understand in English
- Assume that there *n* mixtures with (u_i, Σ_i, P_i) (mean, variance, and prior)
- There are k samples ($\mathbf{x}_k$) which are drawn from these *n* mixtures, but don't know where they are from

EM Example: E Step

models

samples

	1	2	…	n
1				
2				
…				
k				

$$\hat{P}(w_i \mid x_k, \hat{\theta}) = \frac{p(x_k \mid w_i, \hat{\theta}_i)\hat{P}(w_i)}{\sum_{j=1}^{c} p(x_k \mid w_j, \hat{\theta}_j)\hat{P}(w_j)}$$

- E-Step:
  - Given the parameters
    - Prior $P(w_i)$ and θ's
  - Estimate the mixture
    - $P(w_i|X_k,\theta)$

EM Example: M Step

- Give the mixture, update the prior and other parameters

- Prior $$\widehat{P}(w_i) = \frac{1}{n}\sum_{k=1}^{n}\widehat{P}(w_i \mid x_k, \widehat{\theta})$$

models

samples

	1	2	…	n
1				
2				
…				
k				

EM Example: M Step

- Prior

$$\hat{P}(w_i) = \frac{1}{n}\sum_{k=1}^{n}\hat{P}(w_i \mid x_k, \hat{\theta})$$

- Mean

$$\hat{u}_i = \frac{\sum_{k=1}^{n}\hat{P}(w_i \mid x_k, \hat{\theta})x_k}{\sum_{k=1}^{n}\hat{P}(w_i \mid x_k, \hat{\theta})}$$

- Variance

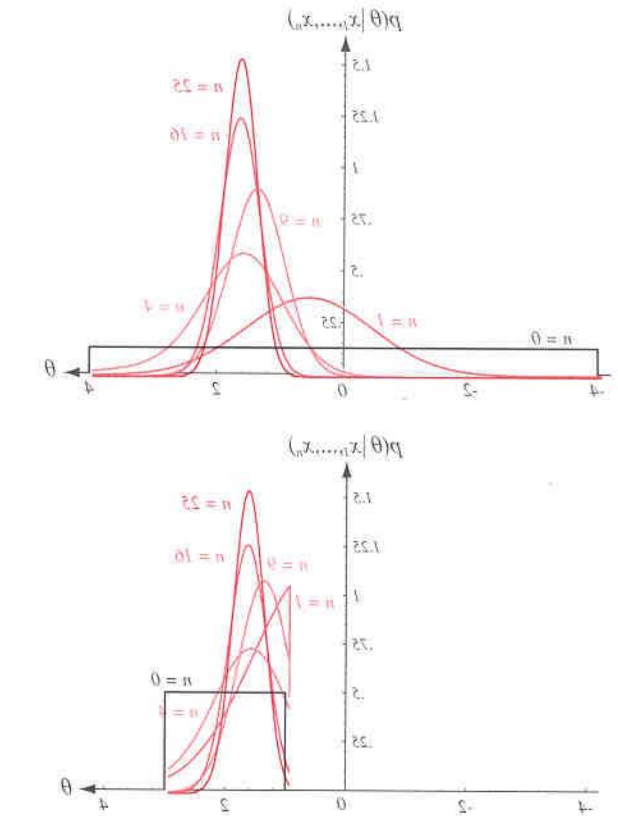
$$\hat{\Sigma}_i = \frac{\sum_{k=1}^{n}\hat{P}(w_i \mid x_k, \hat{\theta})(x_k - \hat{u}_i)(x_k - \hat{u}_i)^t}{\sum_{k=1}^{n}\hat{P}(w_i \mid x_k, \hat{\theta})}$$

- where

$$\hat{P}(w_i \mid x_k, \hat{\theta}) = \frac{p(x_k \mid w_i, \hat{\theta}_i)\hat{P}(w_i)}{\sum_{j=1}^{c} p(x_k \mid w_j, \hat{\theta}_j)\hat{P}(w_j)}$$

Bayesian Learning

- Similar thing can happen as in supervised learning (a sharpening of belief)
- However, the math is much more complicated to say the least



In unsupervised Bayesian learning of the parameter θ, the density becomes more peaked as the number of samples increases. The top figure uses a wide uniform prior p(θ) = 1/8, for −4 ≤ θ ≤ 4 while the bottom figure uses a narrower one, p(θ) = 1/2, for 1 ≤ θ ≤ 3. Despite these different prior distributions, after all 25 samples have been used, the posterior densities are virtually identical in the two cases—the information in the samples overwhelms the prior information.