Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN, modified by Leonardo Bobadilla and some parts from
http://www.cs.tau.ac.il/~apartzin/MachineLearning/
and
www.cs.princeton.edu/courses/archive/fall01/cs302/notes/11.../EM.ppt
© The MIT Press, 2010 alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2m
Outline

Previous class
Ch 6: Dimensionality reduction
This class:
Ch 7: Clustering
CHAPTER 7:

Clustering
Clustering: Motivation

- Optical Character Recognition
  - Two ways to write 7 (w/o horizontal bar)
  - Can’t assume single distribution
  - Mixture of unknown number of templates

- Compared to classification
  - Number of classes is known
  - Each training sample has a label of a class
  - Supervised Learning
Example: Color quantization

- Image: each pixels represented by 24 bit color
- Colors come from different distribution (e.g. sky, grass)
- Don’t have labeling for each pixels if it’s sky or grass
- Want to use only 256 colors in palette to represent image as close as possible to original
- Quantize uniformly: assign single color to each $2^{24}/256$ interval
- Waste values for rarely occurring intervals
Quantization

- Sample (pixels): \( X = \{x^t\}_{t=1}^N \)
- \( k \) reference vectors (palette): \( m_j, j = 1, \ldots, k \)
- Select reference vector for each pixel: 
  \[
  \|x^t - m_i\| = \min_j \|x^t - m_j\|
  \]
- Reference vectors: codebook vectors or code words
- Compress image: 
  \[
  E\left(\{m_i\}_{i=1}^k X\right) = \sum_t \sum_i b_i^t \|x^t - m_i\|
  \]
- Reconstruction error: 
  \[
  b_i^t = \begin{cases} 
  1 & \text{if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\
  0 & \text{otherwise}
  \end{cases}
  \]
Encoding/Decoding

$$
\begin{align*}
\min_j & - x_j - m_i \\
& 0
\end{align*}
$$

Based on E ALPAYDIN 2004 Introduction to Machine Learning © The MIT Press (V1.1)
K-means clustering

- Minimize reconstruction error
  \[ E\left( \{ m_i \}_{i=1}^k \| X \right) = \sum_t \sum_i b_i^t x^t - m_i \]

- Take derivatives and set to zero
  \[ m_i = \frac{\sum_t b_i^t x^t}{\sum_t b_i^t} \]

- Reference vectors is the mean of all instances it represents
K-Means clustering

- Iterative procedure for finding reference vectors
- Start with random reference vectors
- Estimate labels b
- Re-compute reference vectors as means
- Continue till converge
k-means Clustering

Initialize $m_i, i = 1, \ldots, k$, for example, to $k$ random $x^t$

Repeat

For all $x^t \in \mathcal{X}$

$$b_i^t \leftarrow \begin{cases} 
1 & \text{if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\
0 & \text{otherwise}
\end{cases}$$

For all $m_i, i = 1, \ldots, k$

$$m_i \leftarrow \frac{\sum_t b_i^t x^t}{\sum_t b_i^t}$$

Until $m_i$ converge
Figure 7.2  Evolution of $k$-means. Crosses indicate center positions. Data points are marked depending on the closest center.
Expectation Maximization: Learning from Data

We want to learn a model with a set of parameter values $\Phi$
We are given a set of data $X$.
An approach: $\arg\max_\Phi \Pr(X|\Phi)$
This is the maximum likelihood model (ML).
Super Simple Example

Coin I and Coin II. (biased.)
Pick a coin at random (uniform).
Flip it 4 times.
Repeat.

What are the parameters of the model?
Data

<table>
<thead>
<tr>
<th>Coin I</th>
<th>Coin II</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHHT</td>
<td>TTTTH</td>
</tr>
<tr>
<td>HTHH</td>
<td>THTTT</td>
</tr>
<tr>
<td>HTTH</td>
<td>TTHHT</td>
</tr>
<tr>
<td>THHH</td>
<td>HTHTH</td>
</tr>
<tr>
<td>HHHH</td>
<td>HTTTT</td>
</tr>
</tbody>
</table>
Probability of \( X \) Given \( \Phi \)

\( p \): Probability of \( H \) from Coin I
\( q \): Probability of \( H \) from Coin II

Let’s say \( h \) heads and \( t \) tails for Coin I. \( h' \) and \( t' \) for Coin II.

\[
\Pr(X|\Phi) = p^h (1-p)^t q^{h'} (1-q)^{t'}
\]

How maximize this quantity?
Maximizing $p$

Use maximum likelihood.

$h/(t+h) = p$
Missing Data

HHHT
TTTH
THTT
TTHT
THHH

HTTH
HTHH
HTTT
HHHH
HTHT
Oh Boy, Now What!

If we knew the labels (which flips from which coin), we could find ML values for p and q.
What could we use to label? p and q!
Computing Labels

\[ p = \frac{3}{4}, \quad q = \frac{3}{10} \]

\[
\Pr(\text{Coin I} \mid \text{HHTH})
= \Pr(\text{HHTH} \mid \text{Coin I}) \, \Pr(\text{Coin I}) / c
= (\frac{3}{4})^3 \left(\frac{1}{4}\right) \left(\frac{1}{2}\right) / c = \frac{.052734375}{c}
\]

\[
\Pr(\text{Coin II} \mid \text{HHTH})
= \Pr(\text{HHTH} \mid \text{Coin II}) \, \Pr(\text{Coin II}) / c
= (\frac{3}{10})^3 \left(\frac{7}{10}\right) \left(\frac{1}{2}\right) / c = \frac{.00945}{c}
\]
### Expected Labels

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>II</td>
<td></td>
<td>I</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td>HHHH</td>
<td>.85</td>
<td>.15</td>
<td>HTTH</td>
<td>.44</td>
<td>.56</td>
</tr>
<tr>
<td>TTTT</td>
<td>.10</td>
<td>.90</td>
<td>HTHH</td>
<td>.85</td>
<td>.15</td>
</tr>
<tr>
<td>THTT</td>
<td>.10</td>
<td>.90</td>
<td>HTTT</td>
<td>.10</td>
<td>.90</td>
</tr>
<tr>
<td>TTHT</td>
<td>.10</td>
<td>.90</td>
<td>HHHH</td>
<td>.98</td>
<td>.02</td>
</tr>
<tr>
<td>THHH</td>
<td>.85</td>
<td>.15</td>
<td>HTHT</td>
<td>.44</td>
<td>.56</td>
</tr>
</tbody>
</table>
Wait, I Have an Idea

Pick some mode $\Phi_0$

Expectation
- Compute expected labels via $\Phi_i$

Maximization
- Compute ML model $\Phi_{i+1}$

Repeat
Could This Work?

Expectation-Maximization (EM)
Pr(X| \Phi_i) will not decrease.
Sound familiar? Type of search.
Mixture Densities

\[ p(x) = \sum_{i=1}^{k} p(x|G_i) P(G_i) \]

• where \( G_i \) the components/groups/clusters, 
  \( P(\ G_i) \) mixture proportions (priors), 
  \( p(\ x \mid G_i) \) component densities

• Gaussian mixture where \( p(x|G_i) \sim N(\mu_i, \Sigma_i) \) parameters \( \Phi = \{P(\ G_i), \mu_i, \Sigma_i\}^{k_i=1} \)

unlabeled sample \( X=\{x^t\}_t \) (unsupervised learning)
Example
Expectation Maximization (EM): Motivation

• Date came from several distributions
• Assume each distribution is known up to parameters
• If we would know for each data instance from what distribution it came, could use parametric estimation
• Introduce unobservable (latent) variables which indicate source distribution
• Run iterative process
  - Estimate latent variables from data and current estimation of distribution parameters
  - Use current values of latent variables to refine parameter estimation
EM

- Log-Likelihood
  \[ L(\Phi \mid X) = \log \prod_t \prod \rho(x_t \mid \Phi) \]
  \[ = \sum_t \log \sum_{i=1}^k \rho(x_t \mid G_i) P(G_i) \]

- Assume hidden variables \( Z \), which when known, make optimization much simpler.
- Complete likelihood, \( L_c(\Phi \mid X, Z) \), in terms of \( X \) and \( Z \).
- Incomplete likelihood, \( L(\Phi \mid X) \), in terms of \( X \).
Latent Variables

- Unknown
- Can’t compute complete likelihood $L_c(\Phi | X, Z)$
- Can compute its expected value

$$E\text{-step:} Q(\Phi | \Phi^l) = E\left[ L_c(\Phi | X, Z) | X, \Phi^l \right]$$
E- and M-steps

- Iterate the two steps:
  1. E-step: Estimate \( z \) given \( X \) and current \( \Phi \)
  2. M-step: Find new \( \Phi' \) given \( z, X, \) and old \( \Phi \).

\[
\text{E-step: } Q(\Phi|\Phi^l) = E \left[ L_C(\Phi|X,Z)|X,\Phi^l \right]
\]

\[
\text{M-step: } \Phi^{l+1} = \arg \max \Phi Q(\Phi|\Phi^l)
\]
Example:

- Data came from mix of Gaussians
- Maximize likelihood assuming we know latent “indicator variable”
  
  $m_{i+1}^{l+1} = \frac{\sum_t h_i^t x_t}{\sum_t h_i^t}$
  
  $s_{i+1}^{l+1} = \frac{\sum_t h_i^t (x_t - m_{i+1}^{l+1}) (x_t - m_{i+1}^{l+1})^T}{\sum_t h_i^t}$

- E-step: expected value of indicator variables

$h_i^t = \frac{\pi_i |S_i|^{-1/2} \exp[-(1/2) (x_t - m_i)^T S_i^{-1} (x_t - m_i)]}{\sum_j \pi_j |S_j|^{-1/2} \exp[-(1/2) (x_t - m_j)^T S_j^{-1} (x_t - m_j)]}$

Based on E ALPAYDIN 2004 Introduction to Machine Learning © The MIT Press (V1.1)
\[ P(G_1|x) = h_1 = 0.5 \]
EM for Gaussian mixtures

- Assume all groups/clusters are Gaussians
- Multivariate Uncorrelated
- Same Variance
- Harden indicators
  - EM: expected values are between 0 and 1
  - K-means: 0 or 1
- Same as k-means
Dimensionality Reduction vs. Clustering

- Dimensionality reduction methods find correlations between features and group features
  - Age and Income are correlated

- Clustering methods find similarities between instances and group instances
  - Customer A and B are from the same cluster
Clustering: Usage for supervised learning

- Describe data in terms of cluster
  - Represent all data in cluster by cluster mean
  - Range of attributes

- Map data into new space (preprocessing)
  - $d$-dimension original space
  - $k$-number of clusters
  - Use indicator variables as data representations
  - $k$ might be larger than $d$
Mixture of Mixtures

• In classification, the input comes from a mixture of classes (supervised).
• If each class is also a mixture, e.g., of Gaussians, (unsupervised), we have a mixture of mixtures:

\[
p(x | C_i) = \sum_{j=1}^{k_i} p(x | G_{ij}) P(G_{ij})
\]

\[
p(x) = \sum_{i=1}^{K} p(x | C_i) P(C_i)
\]
Hierarchical Clustering

• Probabilistic view
  – Fit mixture model to data
  – Find codewords minimizing reconstruction error

• Hierarchical clustering
  – Group similar items together
  – No specific model/distribution
  – Items in groups is more similar to each other than instances in different groups
Hierarchical Clustering

Minkowski ($L_p$) (Euclidean for $p = 2$)

$$d_m(x^r, x^s) = \left[ \sum_{j=1}^{d} (x_j^r - x_j^s)^p \right]^{1/p}$$

City-block distance

$$d_{cb}(x^r, x^s) = \sum_{j=1}^{d} |x_j^r - x_j^s|$$
Agglomerative Clustering

- Start with clusters each having single point
- At each step merge similar clusters
- Measure of similarity
  - Minimal distance (single link)
    - Distance between closest points in 2 groups
  - Maximal distance (complete link)
    - Distance between most distant points in 2 groups
  - Average distance
    - Distance between group centers
Example: Single-Link Clustering

Dendrogram
Choosing k

- Defined by the application, e.g., image quantization
- Plotting data in two dimensions using PCA
- Incremental (leader-cluster) algorithm: Add one at a time until “elbow” (reconstruction error/log likelihood/intergroup distances)