Classification of High Dimensional Data
By Two-way Mixture Models

Jia Li

Statistics Department
The Pennsylvania State University
Outline

• Goals

• Two-way mixture model approach
  – Background: mixture discriminant analysis
  – Model assumptions and motivations
  – Dimension reduction implied by the two-way mixture model
  – Estimation algorithm

• Examples
  – Document topic classification (Discrete)
    * A mixture of Poisson distributions
  – Disease-type classification using microarray gene expression data (Continuous)
    * A mixture of normal distributions

• Conclusions and future work
Goals

- Achieve high accuracy for the classification of high dimensional data.
  - Document data:
    * Dimension: $p > 3400$.
    * Training sample size: $n \approx 2500$.
    * Number of classes: $K = 5$.
    * The feature vectors are sparse.
  - Gene expression data:
    * Dimension: $p > 4000$.
    * Training sample size: $n < 100$.
    * Number of classes: $K = 4$.

- Attribute (variable, feature) clustering may be desired.
  - Document data: which words play similar roles and do not need to be distinguished for identifying a set of topics?
  - Gene expression data: which genes function similarly?
Mixture Discriminant Analysis

- Proposed as an extension of linear discriminant analysis.
- The mixture of normals is used to obtain a density estimation for each class.
- Denote the feature vector by $X$ and the class label by $Y$.
- For class $k = 1, 2, \ldots, K$, the within-class density is:

$$f_k(x) = \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma)$$
• A 2-classes example. Class 1 is a mixture of 3 normals and class 2 a mixture of 2 normals. The variances for all the normals are 3.0.
The overall model is:

\[ P(X = x, Y = k) = a_k f_k(x) \]

\[ = a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma) \]

where \( a_k \) is the prior probability of class \( k \).

Equivalent formulation:

\[ P(X = x, Y = k) = \sum_{m=1}^{M} \pi_m \phi(x | \mu_m, \Sigma) q_m(k) \]

where \( q_m \) is a pmf for the class label \( Y \) within a mixture component.

Here we have \( q_m(k) = 1.0 \) if mixture component \( m \) “belongs to” class \( k \) and zero otherwise.

The ML estimation of \( a_k \) is the proportion of training samples in class \( k \).

EM algorithm is used to estimate \( \pi_{kr}, \mu_k, \) and \( \Sigma \).

Bayes classification rule:

\[ \hat{y} = \arg \max_k a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma) \]
Assumptions for the Two-way Mixture

- For each mixture component, the variables are independent.
  - As a class may contain multiple mixture components, the variables are NOT independent in general given the class.
  - To approximate the density within each class, the restriction on each component can be compensated by having more components.
  - Convenient for extending to a two-way mixture model.
  - Efficient for treating missing data.

- Suppose $X$ is $p$-dimensional, $x = (x_1, x_2, \ldots, x_p)^T$. The mixture model is:

$$P(X = x, Y = k) = \sum_{m=1}^{M} \pi_{mqm(k)} \prod_{j=1}^{p} \phi(x_j | \theta_{m,j})$$

We need to estimate parameter $\theta_{m,j}$ for each dimension $j$ in each mixture component $m$. 
• When the dimension is very high (sometimes \( p \gg n \)), we may need an even more parsimonious model for each mixture component.

• Clustering structure on the variables:
  
  – Assume that the \( p \) variables belong to \( L \) clusters. Two variables \( j_1, j_2 \), in the same cluster have \( \theta_{m,j_1} = \theta_{m,j_2}, m = 1, 2, \ldots, M \).
  
  – Denote the cluster identity of variable \( j \) by \( c(j) \in \{1, \ldots, L\} \).
  
  – For a fixed mixture component \( m \), only need to estimate \( L \theta's \).
  
  – The \( \theta_{m,j} \)'s are shrunk to \( L \theta_{m,c(j)} \)'s.

\[
P(X = x, Y = k) = \sum_{m=1}^{M} a_m q_m(k) \prod_{j=1}^{p} \phi(x_j | \theta_{m,c(j)})
\]

• This way of regularizing the model leads to variable clustering.
• Properties of variable clusters:
  – Variables in the same cluster have the same distributions within each class.
  – For each cluster of variables, only a small number of statistics are sufficient for predicting the class label.
Dimension Reduction

• Within each mixture component, variables in the same cluster are i.i.d. random variables.

• For i.i.d. random variables sampled from an exponential family, the dimension of the sufficient statistic for the parameter $\theta$ is fixed w.r.t. the sample size.

• Assume the exponential family to be:

$$p_\theta(x_j) = \exp \left( \sum_{s=1}^{S} \eta_s(\theta) T_s(x_j) - B(\theta) \right) h(x_j)$$

**Proposition:** For $X_j$'s in cluster $l$, $l = 1, \ldots, L$, define

$$\bar{T}_{l,s}(x) = \sum_{j : c(j) = l} T_s(x_j) \quad s = 1, 2, \ldots, S.$$ 

Given $\bar{T}_{l,s}(x)$, $l = 1, \ldots, L$, $s = 1, \ldots, S$, the class of a sample $Y$ is conditionally independent of $X_1, X_2, \ldots, X_p$. 
• **Dimension reduction:** “sufficient information” for predicting $Y$ is of dimension $LS$. We often have $LS \ll p$.

• Examples:

  – Mixtures of Poisson: $S = 1$.

    $$\tilde{T}_{l,1}(x) = \sum_{j:c(j)=l} x_j$$

  – Mixtures of normal: $S = 2$.

    $$\tilde{T}_{l,1}(x) = \sum_{j:c(j)=l} x_j$$
    $$\tilde{T}_{l,2}(x) = \sum_{j:c(j)=l} x_j^2$$

    Equivalently:

    **Sample mean:**
    $$\bar{T}_{l,1}(x) = \frac{\sum_{j:c(j)=l} x_j}{\sum_j I(c(j) = l)}$$

    **Sample variance:**
    $$\tilde{T}_{l,2}(x) = \frac{\sum_{j:c(j)=l} (x_j - \bar{T}_{l,1}(x))^2}{\sum_j I(c(j) = l)}$$
Model Fitting

- We need to estimate the following:
  - Mixture component prior probabilities \( a_m, m = 1, ..., M \).
  - Parameters of the Poisson distributions: \( \theta_{m,l}, m = 1, ..., M, l = 1, ..., L \).
  - The variable clustering function \( c(j), j = 1, ..., p, c(j) \in \{1, ..., L\} \).

- Criterion: Maximum likelihood estimation.

- Algorithm: EM.
  - E-step: compute the posterior probability of each sample coming from each mixture component.
  - M-step:
    * Update the parameters \( a_m, \theta_{m,l} \).
    * Update the variable clustering function \( c(j) \) by optimizing \( c(j) \) individually for each \( j, j = 1, ..., p \) with all the other parameters fixed.

- Computational perspective:
  - E-step: a “soft” clustering of samples into mixture components, “row-wise” clustering.
Document Topic Classification

- Classify documents into different topics.
  
  1. *comp.graphics*
  2. *rec.sport.baseball*
  3. *sci.med*
  4. *sci.space*
  5. *talk.politics.guns*

- Classification is based on word counts.
  

- Each document is represented by a vector of word counts. Every dimension corresponds to a particular word.

- Each class contains about 1000 documents. Roughly half of them are randomly selected as training data, and the others testing.

- Pre-processing: for each document class, the 1000 words with the largest total counts in the training data are used as variables.

- The dimension of the word vector is *p = 3455*, \( p > n \).
Mixture of Poisson Distribution

- The Poisson distribution is uni-modal.

\[ P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}. \]

- Example mixtures of Poisson distributions:

- Mixture of multivariate independent Poisson distributions with variable clustering:

\[ P(X = x, Y = k) = \sum_{m=1}^{M} a_m q_m(k') \prod_{j=1}^{p} \frac{\lambda_{m,c(j)}^{x_j}}{x_j!} \cdot e^{-\lambda_{m,c(j)}} \]
Results

- Classification error rates achieved without variable clustering. \#components per class = 1 \sim 20.

\begin{center}
\begin{figure}
\end{figure}
\end{center}

- \(L = 30 \sim 3455\), \#components per class = 6.
• Confusion table for $M = 30$, without word clustering, $p = 3455$. Classification error rate: 11.22%.

<table>
<thead>
<tr>
<th></th>
<th>graphics</th>
<th>baseball</th>
<th>sci.med</th>
<th>sci.space</th>
<th>politics.guns</th>
</tr>
</thead>
<tbody>
<tr>
<td>graphics</td>
<td>463</td>
<td>5</td>
<td>9</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>baseball</td>
<td>3</td>
<td>459</td>
<td>4</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>sci.med</td>
<td>22</td>
<td>12</td>
<td>435</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>sci.space</td>
<td>27</td>
<td>14</td>
<td>28</td>
<td>409</td>
<td>18</td>
</tr>
<tr>
<td>politics.guns</td>
<td>11</td>
<td>27</td>
<td>17</td>
<td>17</td>
<td>434</td>
</tr>
</tbody>
</table>

• For $M = 30$, $L = 168$. Classification error rate: 12.51%.

<table>
<thead>
<tr>
<th></th>
<th>graphics</th>
<th>baseball</th>
<th>sci.med</th>
<th>sci.space</th>
<th>politics.guns</th>
</tr>
</thead>
<tbody>
<tr>
<td>graphics</td>
<td>458</td>
<td>1</td>
<td>12</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>baseball</td>
<td>3</td>
<td>446</td>
<td>2</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>sci.med</td>
<td>23</td>
<td>9</td>
<td>408</td>
<td>21</td>
<td>42</td>
</tr>
<tr>
<td>sci.space</td>
<td>24</td>
<td>9</td>
<td>21</td>
<td>404</td>
<td>38</td>
</tr>
<tr>
<td>politics.guns</td>
<td>4</td>
<td>15</td>
<td>18</td>
<td>17</td>
<td>452</td>
</tr>
</tbody>
</table>
• For $M = 30$, $L = 168$, median cluster size is 7. Highly skewed cluster sizes: the largest 10 clusters account for more than half of the 3455 words.

![Graph of word cluster index versus number of words in each cluster](image1)

• The corresponding weighted average of $\lambda_{m,l}$’s for each cluster $l$, $\sum_{m=1}^{M} a_{m} \lambda_{m,l}$, is shown below. The largest few word clusters have very low average counts.

![Graph of word cluster index versus average $\lambda$](image2)
If the $612 + 220 + 180 + 166 + 137 = 1315$ words in the largest five clusters are not used when classifying test samples, the error rate is only slightly increased from $12.15\%$ to $12.99\%$.

Words in all of the clusters with size 5:

- patient, eat, food, treatment, physician
- nasa, space, earth, mission, satellit
- compil, transform, enhanc, misc, lc
- game, team, player, fan, pitcher
- unit, period, journal, march, sale
- wai, switch, describ, directli, docum
- faq, resourc, tool, distribut, hardwar
- approxim, aspect, north, angl, simul
- recogn, wisdom, vm, significantli, breast
- bought, simultan, composit, walter, mag
- statu, ny, dark, eventu, phase
- closer, po, paid, er, huge
- necessarili, steven, ct, encourag, dougla
- replac, chri, slow, nl, adob
Disease Classification by Microarray Data

- The microarray data are provided at the web site: http://llmpp.nih.gov/lymphoma/
- Every sample in the data set contains expression levels of 4026 genes.
- There are 96 samples divided into 9 classes.
- Four classes of 78 samples are chosen for the classification experiment.
  - DLBCL (diffuse large B-cell lymphoma): 42
  - ABB (activated blood B): 16
  - FL (follicular lymphoma): 9
  - CLL (chronic lymphocytic leukemia): 11
- Five-fold cross-validation is used to assess the accuracy of classification.
- Mixture of normal distribution with variable clustering:

\[
P(X = x, Y = k) = \sum_{m=1}^{M} a_m q_m(k) \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi \sigma^2_{m,c(j)}}} \exp \left( \frac{-(x_j - \mu_{m,c(j)})^2}{2\sigma^2_{m,c(j)}} \right)
\]
Results

- Classification error rates achieved without variable clustering. $M = 4 \sim 36$.

- Minimum error rate 10.26% is achieved at $M = 6$.

- Due to the small sample size, classification performance degrades rapidly when $M$ increases.
• Classification error rates achieved with gene clustering. \( L = 10 \sim 100, M = 4, 18, 36. \)

![Graph showing classification error rates](image)

• Gene clustering improves classification.

<table>
<thead>
<tr>
<th>Error rate (%)</th>
<th>( M = 4 )</th>
<th>( M = 6 )</th>
<th>( M = 12 )</th>
<th>( M = 18 )</th>
<th>( M = 36 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No clustering</td>
<td>12.82</td>
<td>10.26</td>
<td>14.10</td>
<td>29.49</td>
<td>43.59</td>
</tr>
<tr>
<td>( L = 50 )</td>
<td>8.97</td>
<td>10.26</td>
<td>7.69</td>
<td>5.13</td>
<td>5.13</td>
</tr>
<tr>
<td>( L = 100 )</td>
<td>8.97</td>
<td>8.97</td>
<td>6.41</td>
<td>7.69</td>
<td>3.85</td>
</tr>
</tbody>
</table>
• Variable clustering allows us to have more mixture components than the sample size.

• The number of parameters in the model is small due to clustering along variables.

• Fix $L = 20$ (20 gene clusters). $M = 4 \sim 144$.

• When $M \geq 36$, the classification error rate remains below 8%.
Conclusions

• A two-way mixture model approach is developed to classify high dimensional data.
  – This model implies dimension reduction.
  – Attributes are clustered in a way to preserve information about the class of a sample.

• Applications of both discrete and continuous models have been studied.

• Future work:
  – Can the two-way mixture approach be extended to achieve dimension reduction under more general settings?