Multi-label Linear Discriminant Analysis

Wenke Sun

Outline

- Classification
- Multi-class vs. Multi label
- Linear Discriminant Analysis (LDA)
- Difficulties of LDA in Multi-label Classification
- Multi-label Linear Discriminant Analysis (MLDA)
- Connections to Related Works
- Experimental Results

Classification vs. Clustering

- Classification: We have a set of predefined classes and try to know which class a new object belongs to.

- Clustering: We don't have any predefined classes and you just tried to group a set of objects and find whether there are any relationships between the objects.
- In the context of machine learning, classification is **supervised learning** and clustering is **unsupervised learning**

Multiclass classification vs. Multilabel classification

- **Multiclass classification** means a classification task with more than two classes.

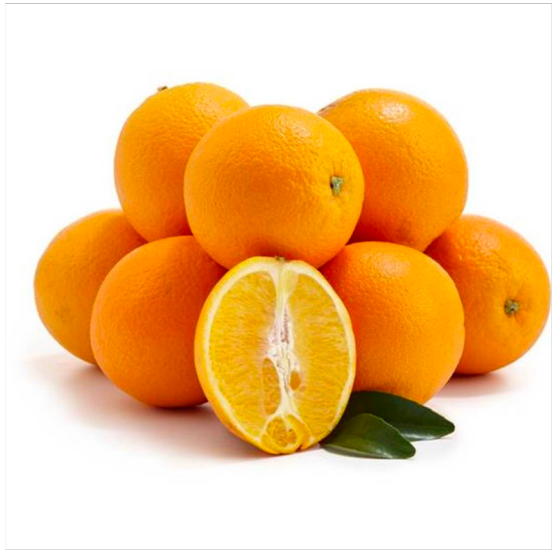
- **Multilabel classification** assigns to each sample a set of target labels. This can be thought as predicting properties of a data-point that are not mutually exclusive, such as topics that are relevant for a document. A text might be about any of religion, politics, finance or education at the same time or none of these.

Multiclass classification

- Classes:
  - A: Banana
  - B: Apple
  - C: Pears
  - D: Orange
- This is also called single-label classification.

C

A



D

B

Multilabel classification

- Classes:
- A: Banana
- B: Apple
- C: Pears
- D: Orange

B,D,A

A,B,C

A,D

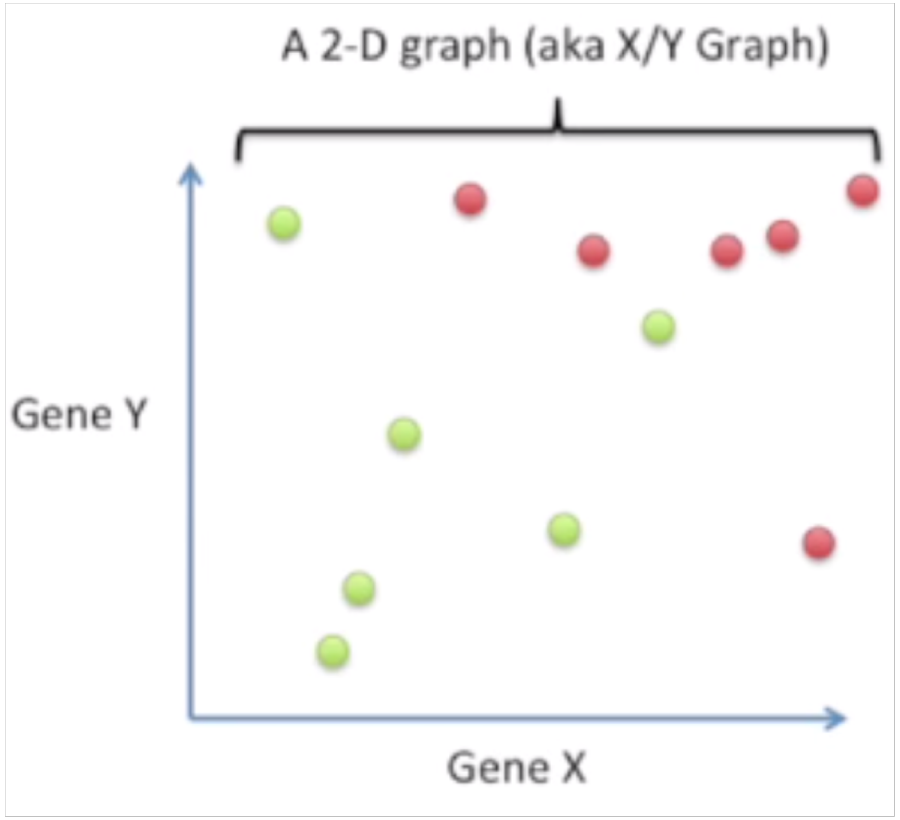
B,C

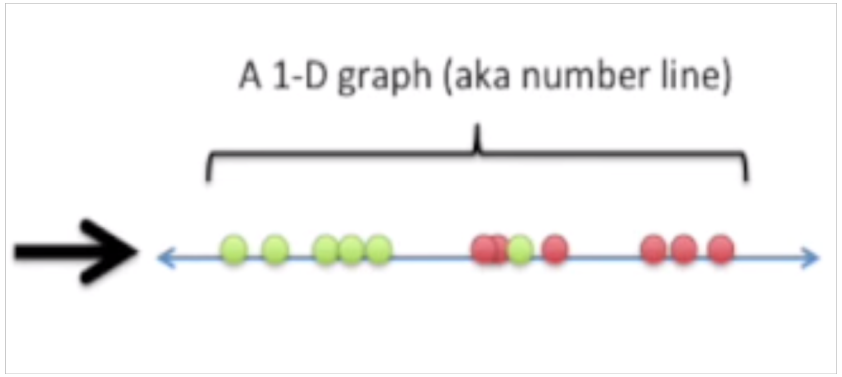
Linear Discriminant Analysis (LDA)

- Find a linear combination of features that characterizes or separates two or more classes of objects or events.
- Just like PCA, LDA also does dimension reduction.
- LDA focuses on maximizing the seperatibility among know categories

Linear Discriminant Analysis (LDA)

A sample two dimension examples.

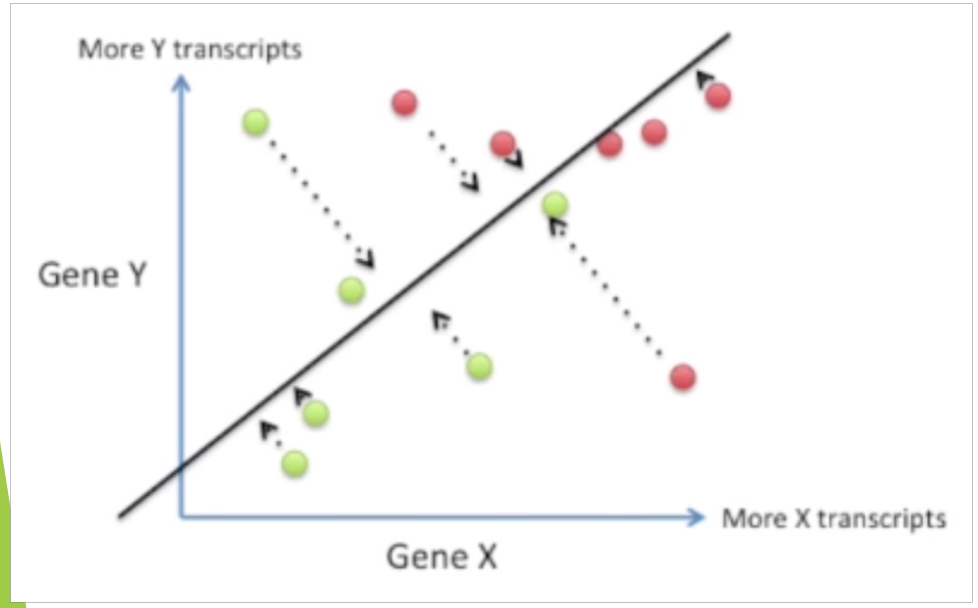


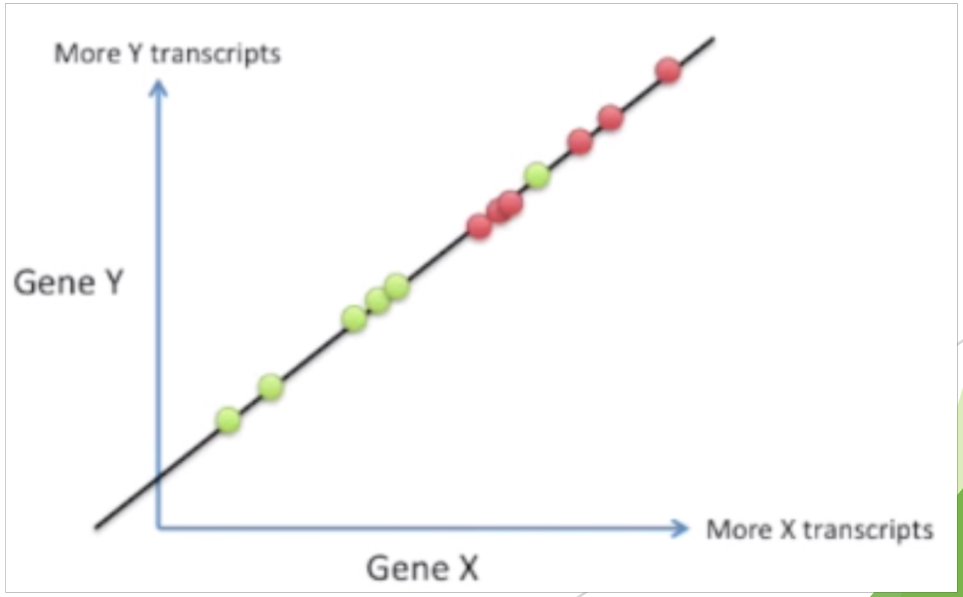


Green dots: A drug works.
Red dots: A drug doesn't work.

Linear Discriminant Analysis (LDA)

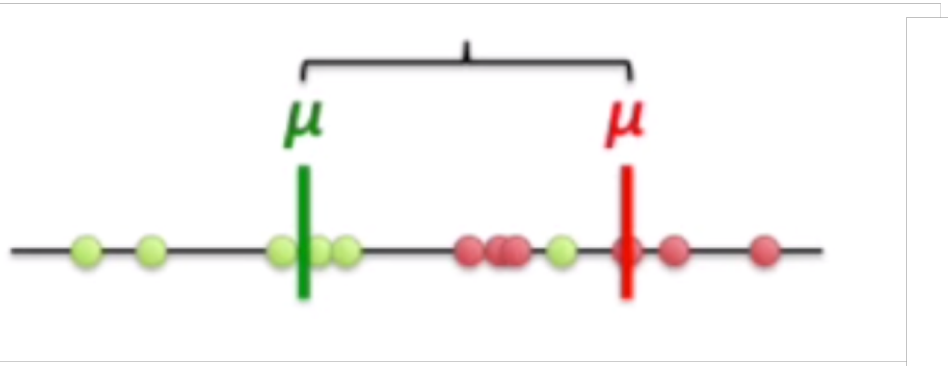
LDA finds the linear combinations of the features and the resulting combination is used as a linear classifier which maximizes the separations of the two categories



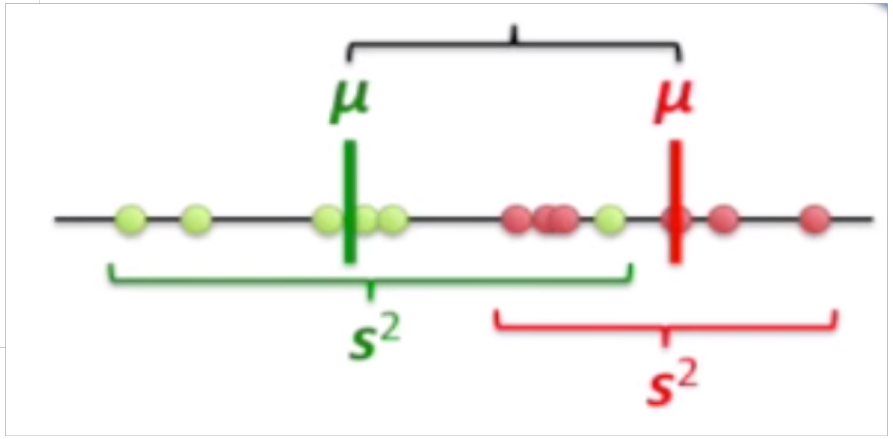


Linear Discriminant Analysis (LDA)

- The linear combination is created based on two criteria.



Maximum the distance between means.



Minimize the variation (LDA calls ”scatter” within each category.

“Arg max” linear transformation G

$$\frac{(\mu - \mu)^2}{s^2 + s^2}$$

Scatter between-class, within-class and total-class in one dimension.

- Suppose we have n groups of data.
- For group i: Average : μi and Variance: σi^2
- Scatter within-class: $\sigma 1^2$ $\sigma 2^2$ …. σn^2
- Scatter between-class: Variance of all μi's.
- Scatter total-class = Scatter between-class + Scatter within-class

Notation Defined

- We have n samples and K classes.
- We have P features so $\mathbf{x}_i \in R^{\boldsymbol{n}}$
- $\boldsymbol{y}_i \in \{0,1\}^K$, $\boldsymbol{y}_i$(k) = 1 if $\mathbf{x}_i$ belongs to the k-th class and 0 otherwise
- $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$
- $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^T = [\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(K)}]$
- $\boldsymbol{Y}_{(K)} \in \{0,1\}^n$ is the class-wise label indication vector for the k-th class.

The between-class, within-class and total-class scatter matrices of LDA

- LDA seeks a linear transformation G = $R^{p\times r}$

$$S_b = \sum_{k=1}^{K} n_k \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)^T,$$

$$S_w = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \pi_k} \left(\mathbf{x}_i - \mathbf{m}_k\right)\left(\mathbf{x}_i - \mathbf{m}_k\right)^T,$$

$$S_t = \sum_{i=1}^{n} \left(\mathbf{x}_i - \mathbf{m}\right)\left(\mathbf{x}_i - \mathbf{m}\right)^T,$$

Class mean：

$$\mathbf{m}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in \pi_k} \mathbf{x}_i$$

Global mean:

$$\mathbf{m} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

The between-class, within-class and total-class scatter matrices of LDA

- $S_b + S_w = S_t$
- The optimal G is chosen such that the **between-class distance is maximize** and **the within-class distance is minimized** in the low-dimensional projected space
- The standard LDA optimization objective as follows:

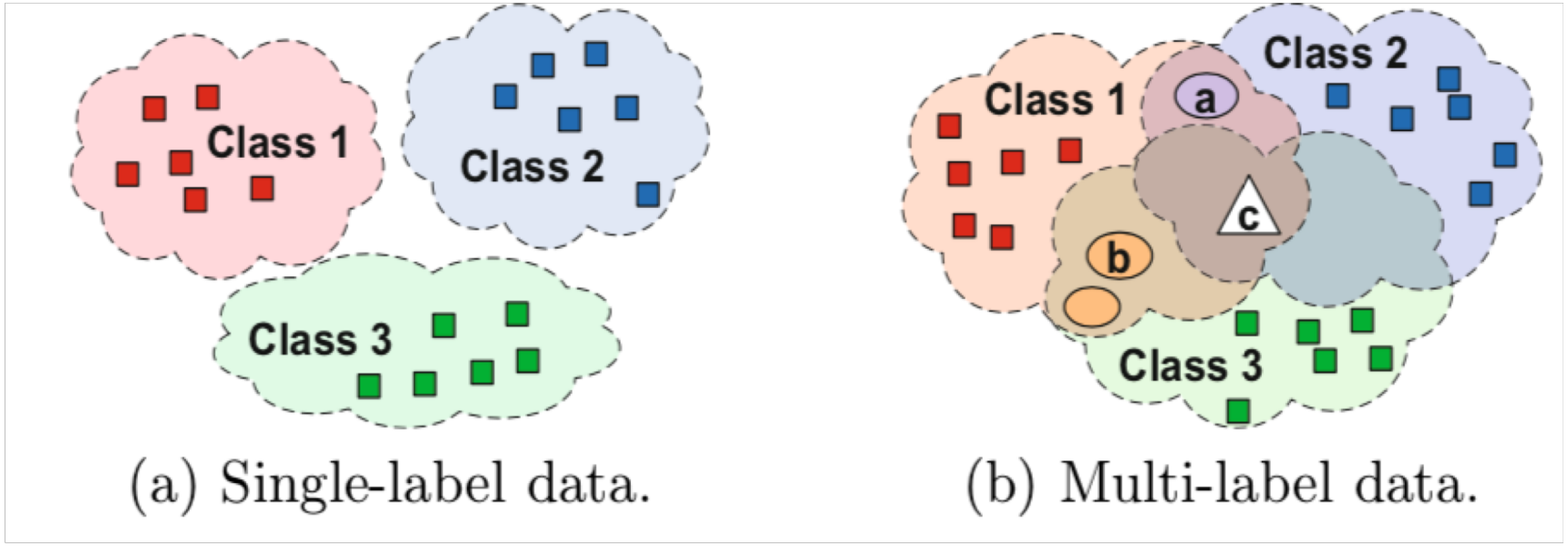
$$\arg\max_{G} J = \mathbf{tr}\left(\frac{G^T S_b G}{G^T S_w G}\right).$$

Difficulty to apply the classical LDA in multi-label classification

- Classes in single-label classification are assumed mutually exclusive, where within-class, between-class and total-class scatter matrices are well defined.

- In multi-label case, there definitions become obscure, since there are overlaps.

Difficulty to apply the classical LDA in multi-label classification



(a) Single-label data. (b) Multi-label data.

Multi-label Linear Discriminant Analysis (MLDA)

- A new way to define the previous within-class, between-class and total-class scatter matrices
- Computing them by class-wise.
- Benefits:
  1. Variances of training data are represented more lucidly
  2. Scatter matrices are easier to be constructed.
  3. Ambiguity can ne avoided and label correlations can be incorporated.

The between-class, within-class and total-class scatter matrices of MLDA

- The class-wise between-class scatter matrix:

$$S_b = \sum_{k=1}^{K} S_b^{(k)}, \; S_b^{(k)} = \left(\sum_{i=1}^{n} Y_{ik} \right) (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T,$$

- The class wise within-class scatter matrix:

$$S_w = \sum_{k=1}^{K} S_w^{(k)}, \; S_w^{(k)} = \sum_{i=1}^{n} Y_{ik} \left(\mathbf{x}_i - \mathbf{m}_k \right) \left(\mathbf{x}_i - \mathbf{m}_k \right)^T$$

The between-class, within-class and total-class scatter matrices of MLDA

- The class-wise total scatter matrix:

$$S_t = \sum_{k=1}^{K} S_t^{(k)}, \quad S_t^{(k)} = \sum_{i=1}^{n} Y_{ik}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$$

$$\mathbf{m}_k = \frac{\sum_{i=1}^{n} Y_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} Y_{ik}}$$

Mean of class k

$$\mathbf{m} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n} Y_{ik}\mathbf{x}_i}{\sum_{k=1}^{K}\sum_{i=1}^{n} Y_{ik}}$$

Multi-label global mean

Two Important Theorems

- This two theorems guarantee the reasonable of the new definition of the scatter matrices.
- **Theorem 1:**

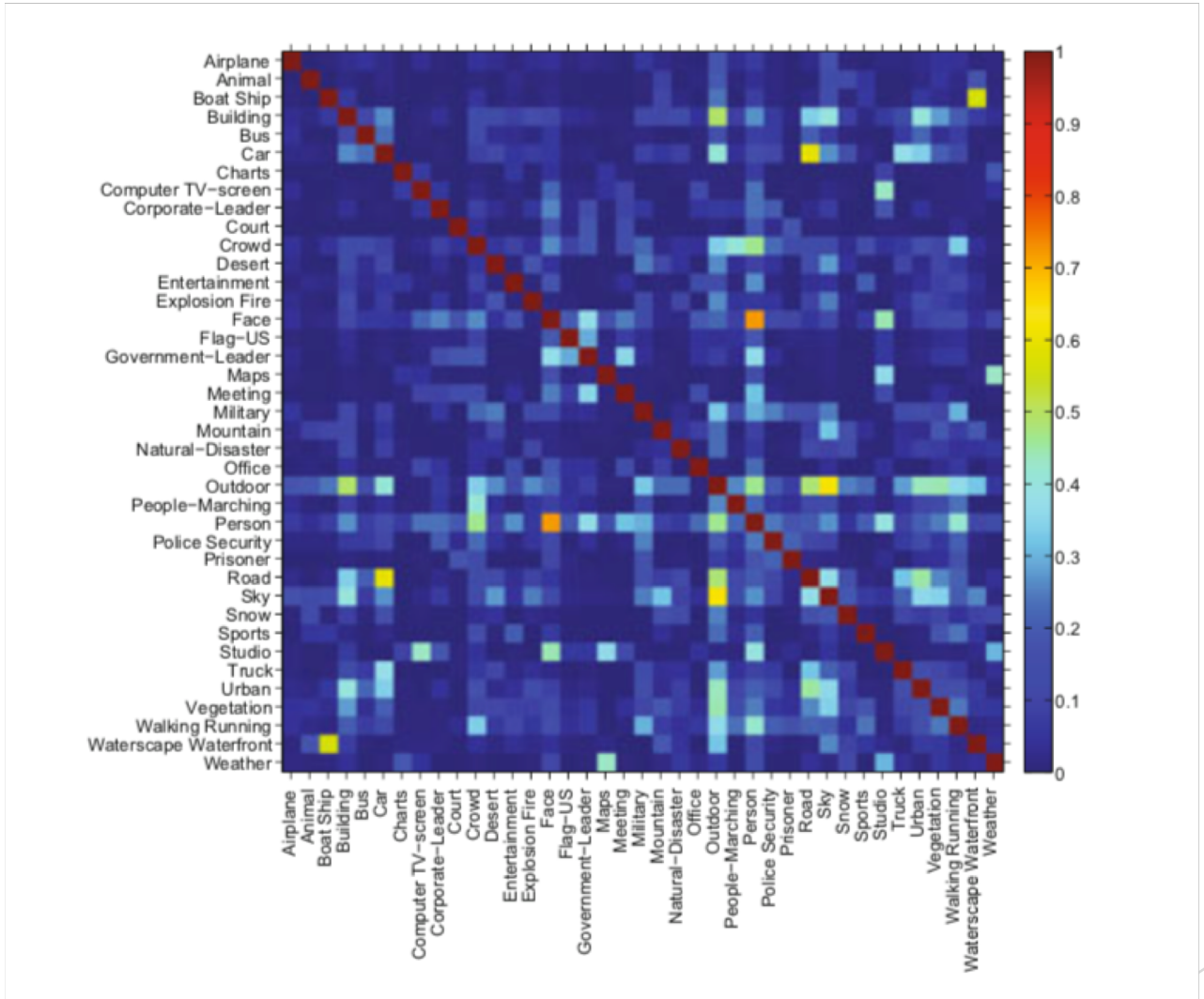
When applied into single-label classification, the multi-label scatter matrices are reduced to their corresponding counterparts in classical LDA.

- **Theorem 2:**

For multi-label class-wise scatter matrices as we defined before, we have ${S_b}^{(k)} + {S_w}^{(k)} = {S_t}^{(k)}$. Thus we still have $S_b + S_w = S_t$.

Multi-label Correlations

- Multi-label data provide a new opportunity to improve classification accuracy through label correlations, which are absent in single-label data.



Multi-label Correlations

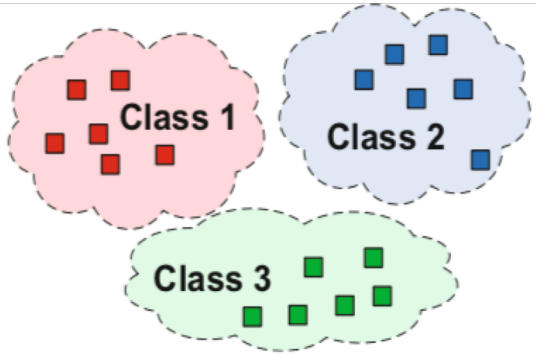
- Typically, the label correlation between two classes is formulated as following:

$$C_{kl} = \cos(\mathbf{y}_{(k)}, \mathbf{y}_{(l)}) = \frac{\langle \mathbf{y}_{(k)}, \mathbf{y}_{(l)} \rangle}{\|\mathbf{y}_{(k)}\| \, \|\mathbf{y}_{(l)}\|}$$

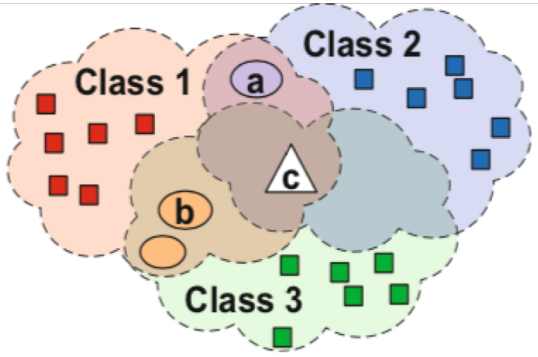
- ***C*** is a K x K symmetric matrix and ***C*** = ***I*** for single-label.
- For multi-label, more overlapped of two classes, more related they are.
- Correlated labels assignments: $Y^C = YC$, where $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T = [\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(K)}]$
- We replace Y by YC in calculation of class-wise scatter matrices to incorporate label correlations
- For single-label, ***YC*** = ***YI*** = ***Y*** , theorem 1 holds.
- Therorem 2 holds, since we introduce C in both sides of equations

Over-Counting Correction

- We can see the points in the overlap region are over counted.



(a) Single-label data.



(b) Multi-label data.

$$S_b = \sum_{k=1}^{K} S_b^{(k)}, \; S_b^{(k)} = \left(\sum_{i=1}^{n} Y_{ik} \right) (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T,$$

$$S_w = \sum_{k=1}^{K} S_w^{(k)}, \; S_w^{(k)} = \sum_{i=1}^{n} Y_{ik} \left(\mathbf{x}_i - \mathbf{m}_k \right) \left(\mathbf{x}_i - \mathbf{m}_k \right)^T$$

Over-Counting Correction

- In general, $\boldsymbol{x}_i$ with multiple labels is used $\sum_{k=1}^{K} \boldsymbol{y}_i(k)$ times.
- We use normalized matrix to correct the over-counting.
- A normalized matrix can be $\boldsymbol{z}_i = \frac{\boldsymbol{y}_i C}{\|\boldsymbol{y}_i\|_{l_1}}$, where $Z = [\mathbf{z}_1, \dots, \mathbf{z}_n]^T \in R^{n \times K}$, and after normalized we have $\sum_{k=1}^{K} \boldsymbol{z}_i(k) = 1$.

- Then we use normalized matrix $\boldsymbol{z}_i = \frac{\boldsymbol{y}_i}{\|\boldsymbol{y}_i\|_{l_1}}$, then $\sum_{k=1}^{K} \boldsymbol{z}_i(k) \geq 1$.

The more labels a data point are associated with, the more important it is .

Note: l_p space is : $\|x\|p=(|x1|^p+|x2|^p+...+|xn|^p)^{1/p}$

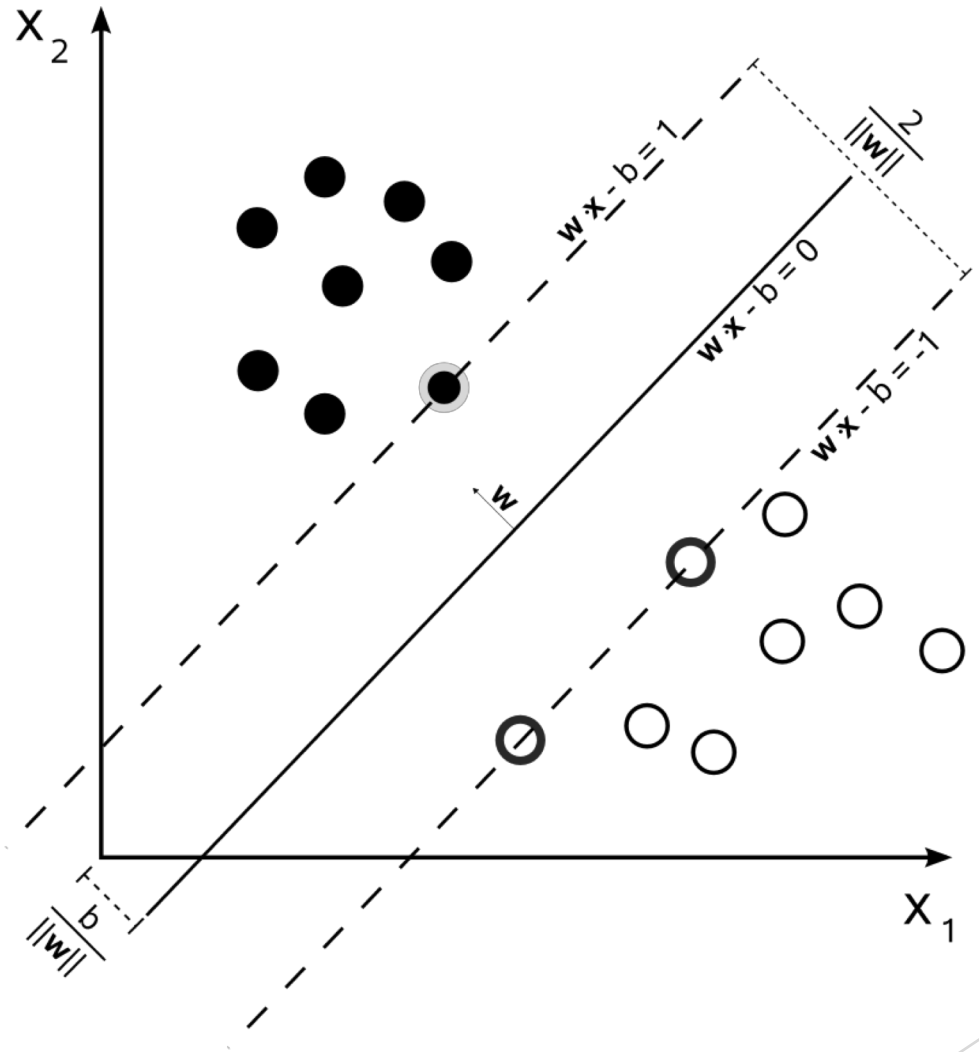
MLDA for Multi-label Classification

- Replace Y by Z in previous MLDA scatter matrices.
- Centering input $\tilde{X} = \mathrm{X} - \mathbf{m}\boldsymbol{e}^T$, where $\boldsymbol{e}$ = [1,,,1]$^\mathrm{T}$ ($\tilde{X} = \mathrm{X} - \boldsymbol{e}\boldsymbol{e}^T/\mathrm{n}$ for LDA)
- $S_b = \tilde{X}ZW^{-1}Z^T\tilde{X}^T$, where W = diag($w_1$,…$w_K$), and $w_k = \sum_{i=1}^{n} Z_{ik}$
- $S_t = \tilde{X}L\tilde{X}^T$ where L = diag. (l_1,…l_K) and $l_i = \sum_{i=1}^{K} Z_{ik}$
- Finally, we get a ratio similar to LDAc which is

$$\arg\max_{G} \mathbf{tr}\left(\frac{G^T S_b G}{G^T S_w G}\right)$$

Related Works

- Support Vector Machine
- In machine learning, **support vector machines** are supervised learning models with associated learning algorithms that analyze data use
for classification and regression analysis.



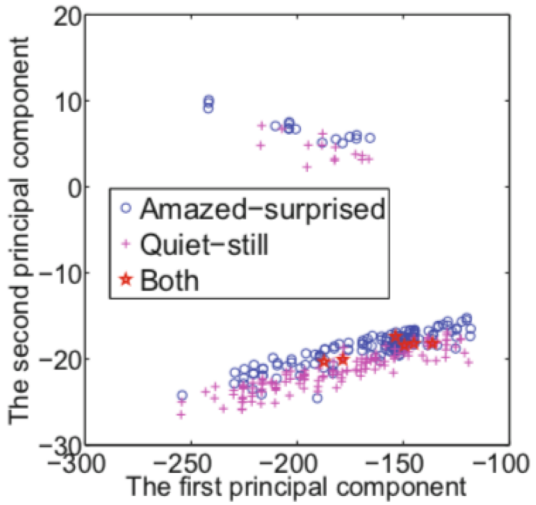
Related Works

- MLSI: extended unsupervised latent semantic indexing to make use of supervision information, called Multi-label informed Latent Semantic Indexing.
- MDDM: proposed Multi-label Dimensionality reduction via Dependence Maximization method
- MLLS: Multi-Label Least Squre
- Flaws:
- 1.$S_b = \sum_{k=1}^{K} w_k \boldsymbol{m}_k {\boldsymbol{m}_k}^T$, which is a coarse approcimation of S_b.
- 2.They didn't think of the label correlations.

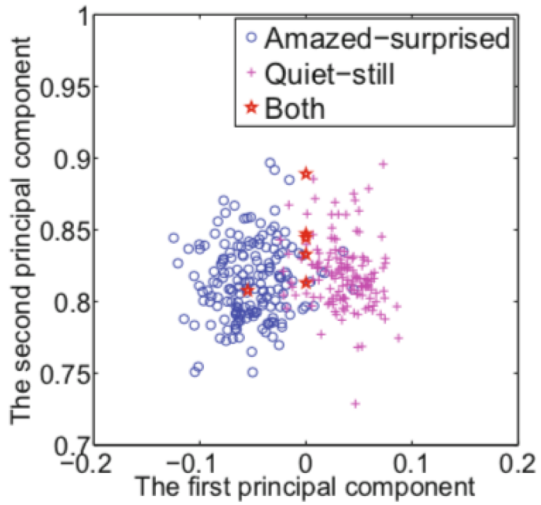
Experiments

- Multi-label data
- TRECVID 2005: data set contains 61901 images and labeled with 39 con- cepts (labels).
- MSRC: data set is provided by the computer vision group at Microsoft Re- search Cambridge, which has 591 images annotated by 22 classes.
- Mediamill: data set includes 43907 sub-shots with 101 classes, where each image is characterized by a 120-dimensional vector.
- Music emotion: data set comprises 593 songs with 6 emotions (labels). The dimensionality of the data points is 72.
- Yahoo data set described in came from the “yahoo.com” domain. Each web page is described as a 37187-dimensional feature vector.

Experiment Results



(a) Visualization on 2D plane in the original space ($p = 72$).



(b) Visualization on 2D plane in the reduced subspace ($l = 5$) by MLDA.

Fig. 4. Visualization on 2D plane for the data points from the two classes in music emotion data set, in original space and projected space by MLDA, respectively

Experiment Results

- F1 score:
- the **F_1 score** is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score.
- *p* is the number of correct positive results divided by the number of all positive results returned by the classifier,
- *r* is the number of correct positive results divided by the number of all relevant samples
- All s amples that should have been identified as positive

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Experiment Results

- The micro average: the summation of contingency matrices for all binary classifiers.
- The macro average is the mean of the values of a standard class-wise metric over all the label.

Table 1. Performance evaluations of six compared methods by 5-fold cross validations

Data	Evaluation metrics		Compared methods					
			LDA-C1	SVM	MLSI	MDDM	MLLS	MLDA
TREC05	Macro average	Precision	0.282	0.269	0.247	0.366	0.248	**0.420**
		F1 score	0.190	0.286	0.275	0.370	0.276	**0.399**
	Micro average	Precision	0.274	0.252	0.234	0.352	0.241	**0.418**
		F1 score	0.408	0.399	0.293	0.491	0.295	**0.528**
MSRC	Macro average	Precision	0.291	0.274	0.252	0.370	0.255	**0.431**
		F1 score	0.201	0.295	0.287	0.392	0.290	**0.410**
	Micro average	Precision	0.288	0.262	0.253	0.363	0.255	**0.420**
		F1 score	0.415	0.406	0.301	0.504	0.302	**0.533**
MediaMill	Macro average	Precision	0.337	0.302	0.207	0.385	0.206	**0.410**
		F1 score	0.349	0.322	0301	0.418	0.311	**0.430**
	Micro average	Precision	0.335	0.297	0.207	0.382	0.205	**0.388**
		F1 score	0.518	0.398	0.341	0.440	0.340	**0.443**
Music emotion	Macro average	Precision	0.507	0.434	0.329	0.509	0.311	**0.614**
		F1 score	0.453	0.418	0.323	0.506	0.471	**0.618**
	Micro average	Precision	0.504	0.501	0.328	0.507	0.308	**0.613**
		F1 score	0.477	0.441	0.339	0.518	0.475	**0.626**
Yahoo (Science)	Macro average	Precision	0.458	0.414	0.396	0.463	0.421	**0.501**
		F1 score	0.227	0.302	0.296	0.481	0.443	**0.498**
	Micro average	Precision	0.447	0.416	0.395	0.458	0.420	**0.499**
		F1 score	0.226	0.218	0.209	0.484	0.519	**0.544**
MSRC (SIFT)	Macro average	Precision	0.415	0.408	0.428	0.520	0.424	**0.612**
		F1 score	0.367	0.358	0.381	0.471	0.376	**0.531**
	Micro average	Precision	0.408	0.403	0.412	0.515	0.407	**0.597**
		F1 score	0.612	0.611	0.620	0.671	0.617	**0.698**

Experiment Results

Thank you!