# Analyse et Traitement de l'Information

TP5: Linear Discriminant Analysis (LDA) and Information Theory.

---

## 1 Fisher's Discriminant for Two Classes (50%)

Assume we are given a *training set* $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ where each $\boldsymbol{x}_i$ is a high dimensional observation in $\Re^d$ and each $y_i = 0, \ldots, (L-1)$ corresponds to a label specifying which of the $L$ possible classes $\boldsymbol{x}_i$ belongs to. For instance, each $\boldsymbol{x}_i$ may be an image where $y_i = 0, \ldots, (L-1)$ corresponds to the image class. The goal of *classification* is to learn a mapping $f: \boldsymbol{x} \to y$ based on this labeled training set. Predictions can then be made for any future sample $\boldsymbol{x}$ with an unknown label by setting it's predicted label equal to $f(\boldsymbol{x})$.

A simple approach to classification is Fisher's linear discriminant analysis (LDA). The basic idea of LDA is to project $\{\boldsymbol{x}_i\}_{i=1}^n$ into a low dimensional space where different classes are "best" separated. The classification is then performed on this low dimensional space. Here, without loss of generality, we only discuss LDA for two classes, i.e., the number of classes is $L = 2$ where $y_i$ can be either 0 or 1. In this case, the classifier is defined as

$$f(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{a}^T\boldsymbol{x} + b \geq 0; \\ 1 & \text{if } \boldsymbol{a}^T\boldsymbol{x} + b < 0, \end{cases} \tag{1}$$

where $\boldsymbol{a} \in \Re^d$ specifies a linear projection from $\Re^d$ to $\Re$ and $b \in \Re$ is a threshold to determine which of the two classes $\boldsymbol{x}$ belongs to. The learning task is to find the optimal $\boldsymbol{a}$ and $b$ based on a given training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$.

Rather than provide the complete derivation, we directly provide the solution of the optimal $\boldsymbol{a}$ and $b$ and refer the reader to the textbooks [1, 2] for the detailed analysis. The optimal $\boldsymbol{a}$ with respect to $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is given by

$$\boldsymbol{a} = S_W^{-1}(\boldsymbol{g}_0 - \boldsymbol{g}_1). \tag{2}$$

In eq.(2),

$$\boldsymbol{g}_0 = \frac{1}{N_0}\sum_{i:y_i=0} \boldsymbol{x}_i, \quad \boldsymbol{g}_1 = \frac{1}{N_1}\sum_{i:y_i=1} \boldsymbol{x}_i,$$

where $N_0$ and $N_1$ are the number of samples in class 0 and 1, respectively, and $\boldsymbol{g}_0$ and $\boldsymbol{g}_1$ are the corresponding mean vectors. $S_W$ is a $d \times d$ *within-class scatter matrix* defined as

$$S_W = \sum_{i:y_i=0} (\boldsymbol{x}_i - \boldsymbol{g}_0)(\boldsymbol{x}_i - \boldsymbol{g}_0)^T + \sum_{i:y_i=1} (\boldsymbol{x}_i - \boldsymbol{g}_1)(\boldsymbol{x}_i - \boldsymbol{g}_1)^T.$$

Since the given classes of data are balanced, the optimal $b$ with respect to $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is given by

$$b = -\frac{1}{2}\boldsymbol{a}^T(\boldsymbol{g}_0 + \boldsymbol{g}_1). \tag{3}$$

Consider the binary learning problem to classify images of "3" and "7". From the `MNIST` dataset, sample 2000 instances of "3"s and "7"s for training and use all instances for testing. Apply PCA to reduce the dimensionality of the input data to 50 before classification. Then perform the following tasks.

1. Apply LDA to classify the complete set of testing instances. Report the accuracy using the confusion matrix[1].
2. Use the $k$-NN method that you implemented in TP3 to classify the testing instances with $k = 1$. Report the accuracy using the confusion matrix.
3. Describe how the two classification methods compare.

## 2 k-means Clustering (50%)

Classification is a *supervised* problem, where labeled (training) data is used to label the unlabeled (testing) data.
Clustering is an *unsupervised* problem, where the objective is to group a given (unlabeled) dataset $\{\boldsymbol{x}_i\}_{i=1}^n$ into $k$ clusters, so that similar samples appear in the same cluster, and dissimilar samples are in different clusters.

The most basic clustering algorithm, $k$-means, aims to partition the input data into $k$ clusters in the most compact way, in the sense that the following objective function is minimized

$$J(\boldsymbol{y}) = \sum_{j=1}^{k} \sum_{i:y_i=j} ||\boldsymbol{x}_i - \boldsymbol{g}_j||^2, \tag{4}$$

where

- $\boldsymbol{y} = (y_1, \ldots, y_n)$ represents a clustering scheme. Each $y_i$ is an integer taking values from $\{1, 2, \ldots, k\}$, indicating which cluster the corresponding sample $\boldsymbol{x}_i$ belongs to;
- $\boldsymbol{g}_j$ is the center of cluster $j$ so that

$$\boldsymbol{g}_j = \frac{1}{N_j} \sum_{i:y_i=j} \boldsymbol{x}_i \quad (N_j \text{ is the number of samples in cluster } j)$$

Therefore, $J(\boldsymbol{y})$ means the sum of the square distances from each sample to its associated cluster center. Minimizing $J(\boldsymbol{y})$ means to take each cluster *compact*, in the sense that a sample should be near to its cluster center. To minimize this $J(\boldsymbol{y})$, the $k$-means algorithm is given as follows

- Randomly sample 2000 instances "3"s and "7"s from the `MNIST` dataset. Perform $k$-means with $k = 2, 3, \ldots, 10$. Compute $J$ defined in eq.(4) for each $k$. Plot $J$ varying with $k$. Describe the effect of $k$ over $J$.

[1] `https://en.wikipedia.org/wiki/Confusion_matrix`

**procedure** K-MEANS($\{\boldsymbol{x}_i\}_{i=1}^n, k$) ▷ *As inputs:* a set of samples and the number of clusters
    ▷ *As outputs:* $\{y_i\}_{i=1}^n$. $y_i \in \{1, 2, ..., k\}$ specifies $\boldsymbol{x}_i$ belongs to which of the $k$ clusters.
Randomly initialize each $y_i^0$ to one of $\{1, 2, ..., k\}$; t=0;
**repeat**
    **for** $j \leftarrow 1 \rightarrow k$ **do** ▷ Compute the centroid of each cluster
        $\boldsymbol{g}_j \leftarrow \sum_{i:y_i^t=j} \boldsymbol{x_i}/N_j$;
    **for** $i \leftarrow 1 \rightarrow n$ **do** ▷ Assign each point to the closest centroid
        $y_i^{t+1} \leftarrow \text{argmin}_j ||\boldsymbol{x}_i - \boldsymbol{g}_j||$;
    $t \leftarrow t + 1$;
**until** $\forall i, y_i^t = y_i^{t-1}$
Output $\{y_i\}_{i=1}^n$

- For the case $k = 2$ in the above experiment, give the confusion matrix.
- Sample 2000 instances of "3"s and "5"s (rather than "3"s and "7"s). Perform $k$-means with $k = 2$. Build the confusion matrix. Based on your results, explain the difference between these two clustering problems: (1)"3" vs. "7"; (2) "3" vs. "5".

### Assessment

Please archive your report and codes in "Prénom Nom.zip" (replace "Prénom" and "Nom" with your real name), and upload to "`Upload TPs/TP4 LDA and k-means`" on `https://moodle.unige.ch` before Monday, November 11 2019, 23:59 PM. Note, the assessment is mainly based on your report, which should include your answers to all questions and the experimental results.

## 3 Supplements

1. Explain the k-means algorithm.
2. Explain LDA.

## References


[1] Christopher M.Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NY, USA, 2006.

[2] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, Cambridge, MA, 2012.