67577 – Intro. to Machine Learning **Fall semester, 2004/5**

Lecture 10: Algebraic Representations II: LDA, CCA

Lecturer: Amnon Shashua *Scribe: Amnon Shashua* [1]

In the previous lecture we introduced algebraic considerations for dimensionality reduction which preserve variance. The dimensionality reduction, called PCA, represents a data point $\mathbf{x} \in R^n$ in a new q-dimensional coordinate system $\mathbf{y} = U^\top(\mathbf{x} - \mu)$ where U is an orthonormal $n \times q$ matrix whose columns consist of the leading q eigenvectors of the (scaled) sample covariance matrix $\sum_i(\mathbf{x} - \mu)(\mathbf{x}_i - \mu)^\top$ of the training data points $\mathbf{x}_1, ..., \mathbf{x}_m$ with mean $\mu = (1/m)\sum_i \mathbf{x}_i$. We saw that variance preserving dimensionality reduction is equivalent to (i) de-correlating the training sample data, and (ii) seeking the q-dimensional subspace of R^n which is the closest (in least-squares sense) possible to the original training sample.

In this lecture we extend the variance preserving approach for data representation for *labeled* data sets. We will focus on 2-class sets and look for a separating hyperplane:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b,$$

such that $\mathbf{x}$ belongs to the first class if $f(\mathbf{x}) > 0$ and $\mathbf{x}$ belongs to the second class if $f(\mathbf{x}) < 0$. In the statistical literature this type of function is called a *linear discriminant function.* The decision boundary is given by the set of points satisfying $f(\mathbf{x}) = 0$ which is a hyperplane. Fisher's (1936) Linear Discriminant Analysis (LDA) is a variance preserving approach for finding a linear discriminant function.

We will then introduce another popular statistical technique called Canonical Correlation Analysis (CCA) for learning the mapping between input and output vectors using the notion "angle" between subspaces.

What is common in the three techniques PCA, LDA and CCA is the use of spectral matrix analysis — i.e., what can you do with eigenvalues and eigenvectors of matrices representing subspaces of the data? These techniques produce optimal results for normally distributed data and are very easy to implement. There is a large variety of uses of spectral analysis in statistical and learning literature including spectral clustering, Multi Dimensional Scaling (MDS) and data modeling in general.

10.1 Fisher's LDA: Basic Idea

To appreciate the general idea behind Fisher's LDA consider Fig. 10.1. Let the centers of classes one and two be denoted by μ_1 and μ_2 respectively. A linear discriminant function is a projection onto a 1D subspace such that the classes would be separated the most in the 1D subspace. The obvious first step in this kind of analysis is to make sure that the projected centers $\hat{\mu}_1, \hat{\mu}_2$ would be separated as much as possible. We can easily see that the direction of the 1D subspace should be proportional to $\mu_1 - \mu_2$ as follows:

$$(\hat{\mu}_1 - \hat{\mu}_2)^2 = \left(\frac{\mathbf{w}^\top \mu_1}{\|\mathbf{w}\|} - \frac{\mathbf{w}^\top \mu_2}{\|\mathbf{w}\|}\right)^2 = \left(\frac{\mathbf{w}^\top}{\|\mathbf{w}\|}(\mu_1 - \mu_2)\right)^2.$$

[1]

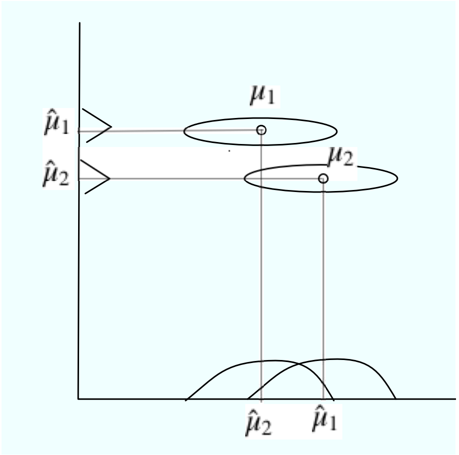

Figure 10.1: Linear discriminant analysis based on class centers alone is not sufficient. Seeking a projection which maximizes the distance between the projected centers will prefer the horizontal axis over the vertical, yet the two classes overlap on the horizontal axis. The projected distance along the vertical axis is smaller yet the classes are better separated. The conclusion is that the sample variance of the two classes must be taken into consideration as well.

The right-hand term is maximized when $\mathbf{w} \approx \mu_1 - \mu_2$. As illustrated in Fig. 10.1, this type of consideration is not sufficient to capture separability in the projected subspace because the spread (variance) of the data points around their centers also play an important role. For example, the horizontal axis in the figure separates the centers better than the vertical axis but on the other hand does a worse job in separating the classes themselves because of the way the data points are spread around their centers. The argument in favor of separating the centers would work if the data points were living in a hyper-sphere around the centers, but will not be sufficient otherwise.

The basic idea behind Fisher's LDA is to consider the sample covariance matrix of the individual classes as well as their centers, in the following way. The optimal 1D projection would that which maximizes the variance of the projected centers while *minimizes* the variance of the projected data points of each class separately. Mathematically, this idea can be implemented by maximizes the following ratio:

$$\max_{\mathbf{w}} \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{s_1^2 + s_2^2},$$

where s_1^2 is the scaled variance of the projected points of the first class:

$$s_1^2 = \sum_{\mathbf{x}_i \in C_1} (\hat{\mathbf{x}}_i - \hat{\mu}_1)^2,$$

and likewise,

$$s_2^2 = \sum_{\mathbf{x}_i \in C_2} (\hat{\mathbf{x}}_i - \hat{\mu}_2)^2,$$

where $\hat{\mathbf{x}} = \frac{\mathbf{w}^\top}{\|\mathbf{w}\|}\mathbf{x}_i + b$.

We will now formalize this approach and derive its solution. We will begin with a general description of a multiclass problem where the sample data points belong to q different classes, and later focus on the case of $q = 2$.

10.2 Fisher's LDA: General Derivation

Let the sample data points S be members of q classes $C_1, ..., C_q$ where the number of points belonging to class C_i is denoted by l_i and the total number of the training set is $l = \sum_i l_i$. Let μ_j denote the center of class C_i and μ denote the center of the complete training set S:

$$\mu_j = \frac{1}{l_j} \sum_{bfx_i \in C_j} \mathbf{x}_i$$

$$\mu = \frac{1}{l} \sum_{\mathbf{x}_i \in S} \mathbf{x}_i$$

Let A_j be the matrix associated with class C_j whose columns consists of the mean shifted data points:

$$A_j = [\mathbf{x}_1 - \mu_j, ..., \mathbf{x}_{l_j} - \mu_j] \qquad \mathbf{x}_i \in C_j.$$

Then, $\frac{1}{l_j} A_j A_j^\top$ is the covariance matrix (see Lecture 9) associated with class C_j. Let S_w (where "w" stands for "within") be the sum of the class covariance matrices:

$$S_w = \sum_i^q \frac{1}{l_j} A_j A_j^\top.$$

From the discussion in the previous section, it is $\frac{1}{\|\mathbf{w}\|^2}\mathbf{w}^\top S_w \mathbf{w}$ which we wish to minimize. To see why this is so, note

$$\sum_{\mathbf{x}_i \in C_j} (\hat{\mathbf{x}}_i - \hat{\mu}_j)^2 = \sum_{\mathbf{x}_i \in C_j} \frac{\mathbf{w}^\top(\mathbf{x}_i - \mu_j)^2}{\|\mathbf{w}\|^2} = \frac{1}{\|\mathbf{w}\|^2}\mathbf{w}^\top A_j A_j^\top \mathbf{w}.$$

Let B be the matrix holding the class centers:

$$B = [\mu_1 - \mu, ..., \mu_q - \mu],$$

and let $S_b = \frac{1}{q} B B^\top$ (where "b" stands for "between"). From the discussion above it is $\frac{1}{\|\mathbf{w}\|^2}\mathbf{w}^\top S_b \mathbf{w} = \sum_i (\hat{\mu}_i - \hat{\mu})^2$ which we wish to *maximize*. Taken together, we wish to maximize the ratio (called "Rayleigh's quotient"):

$$\max_{\mathbf{w}} J(\mathbf{w}) = \frac{\mathbf{w}^\top S_b \mathbf{w}}{\mathbf{w}^\top S_w \mathbf{w}}.$$

The necessary condition for optimality is:

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{S_b \mathbf{w}(\mathbf{w}^\top S_w \mathbf{w}) - S_w \mathbf{w}(\mathbf{w}^\top S_b \mathbf{w})}{(\mathbf{w}^\top S_w \mathbf{w})^2} = 0,$$

From which we obtain the generalized eigensystem:

$$S_b \mathbf{w} = J(\mathbf{w}) S_w \mathbf{w}. \tag{10.1}$$

That is, $\mathbf{w}$ is the leading eigenvector of $S_w^{-1} S_b$ (assuming S_w is invertible). The general case of finding q such axes involves finding the leading generalized eigenvectors of (S_b, S_w) — the derivation is out of scope of this lecture. Note that since $S_w^{-1} S_b$ is not symmetric there may be no real-value solution, which is a complication will not pursue further in this course. Instead we will focus now on the 2-class ($q = 2$) setting below.

10.3 Fisher's LDA: 2-class

The general derivation is simplified when there are only two classes. The covariance matrix $BB^\top$ becomes a rank-1 matrix:

$$BB^\top = (\mu_1 - \mu)(\mu_1 - \mu)^\top + (\mu_2 - \mu)(\mu_2 - \mu)^\top = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top.$$

As a result, $BB^\top \mathbf{w}$ is a vector in direction $\mu_1 - \mu_2$. Therefore, the solution for $\mathbf{w}$ from eqn. 10.1 is:

$$\mathbf{w} \cong S_w^{-1}(\mu_1 - \mu_2).$$

The decision boundary $\mathbf{w}^\top(\mathbf{x} - \mu) = 0$ becomes:

$$\mathbf{x}^\top S_w^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^\top S_w^{-1}(\mu_1 - \mu_2) = 0. \qquad (10.2)$$

This decision boundary will surface again in the course when we consider Bayseian inference. It will be shown that this decision boundary is the Maximum Likelihood solution in the case where the two classes are normally distributed with means μ_1, μ_2 and with the same covariance matrix S_w.

10.4 LDA versus SVM

Both LDA and SVM search for a so called "optimal" linear discriminant function, what is the difference? The heart of the matter lies in the definition of what constitutes a sufficient compact representation of the data. In LDA the assumption is that each class can be represented by its mean vector and its spread (i.e., covariance matrix). This is true for normally distributed data — but not true in general. This means that we should expect that LDA will produce the optimal discriminant linear function when each of the classes are normally distributed.

With SVM, on the other hand, there is no assumption on how the data is distributed. Instead, the emerging result is that the data is represented by the subset of data points which lie on the boundary between the two classes (the so called support vectors). Rather than making a parametric assumption on how the data can be captured (i.e., mean and covariance) the theory shows that the data can be captured by a special subset of points. The tools, as a result, are naturally more complex (quadratic linear programming versus spectral matrix analysis) — but the advantage is that optimality is guaranteed without making assumptions on the distribution of the data (i.e., distribution free). It can be shown that SVM and LDA would produce the same result if the class data is normally distributed.

10.5 Canonical Correlation Analysis

CCA is a technique for learning a mapping $f(\mathbf{x}) = \mathbf{y}$ where $\mathbf{x} \in R^k$ and $\mathbf{y} \in R^s$ using the notion of subspace similarity (an extension of the inner product between two vectors) from a training set of $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, ..., n$. Such a mapping, where $\mathbf{y}$ can be any point in R^k as opposed to a discrete set of labels, is often referred to as a "regression" (as opposed to "classification").

Like in PCA and LDA, the approach would be to look for projection axes such that the projection of the input and output vectors on those axes satisfy certain requirements — and like PCA and LDA the tools we would be using is matrix spectral analysis.

It will be convenient to stack our vectors as rows of an input matrix A and output matrix B. Let A be an $n \times k$ matrix whose rows are $\mathbf{x}_1^\top, ..., \mathbf{x}_n^\top$ and B is the $n \times s$ matrix whose rows are $\mathbf{y}_1^\top, ..., \mathbf{y}_n^\top$. Consider vectors $\mathbf{u} \in R^k$ and $\mathbf{v} \in R^s$ and project the input and output data onto them producing $A\mathbf{u} = (\mathbf{x}_1^\top \mathbf{u}, ..., \mathbf{x}_n^\top \mathbf{u})$ and $B\mathbf{v}$. The requirement we would like to place on the projection axes is that $A\mathbf{u} \approx B\mathbf{v}$, or in other words that $(A\mathbf{u})^\top (B\mathbf{v})$ is maximal. The requirement therefore is that the projection of the input points onto the $\mathbf{u}$ axis is similar to the projection of the output points onto the $\mathbf{v}$ axis. If we extend this notion to multiple axes $\mathbf{u}_1, ..., \mathbf{u}_q$ (not necessarily orthogonal) and $\mathbf{v}_1, ..., \mathbf{v}_q$ where $q \leq \min(k, s)$ our requirement becomes that the new coordinates of the input points projected onto the subspace spanned by the $\mathbf{u}$ vectors are *similar* to the new coordinates of the output points projected onto the subspace spanned by the $\mathbf{v}$ vectors. In other words, we wish to find two q-dimensional subspaces one of R^k and the other of R^s such that the two sets of projected points are as aligned as possible.

CCA goes a step further and makes the assumption that the input/output relationship is solely determined by the relation (angles) between the column spaces of A, B. In other words, the particular columns of A are not really important, what is important is the space U_A spanned by the columns. Since $\mathbf{g} = A\mathbf{u}$ is a point in U_A (a linear combination of the columns of A) and $\mathbf{h} = B\mathbf{v}$ is a point in U_B, then $\mathbf{g}^\top \mathbf{h}$ is the cosine angle, $\cos(\phi)$ between the two axes provided that we normalize the vectors $\mathbf{g}$ and $\mathbf{h}$. If we continue this line of reasoning recursively, we obtain a set of angles $0 \leq \theta_1 \leq ... \leq \theta_q \leq (\pi/2)$, called "principal angles", between the two subspaces uniquely defined as:

$$cos(\theta_j) = \max_{\mathbf{g} \in U_A} \max_{\mathbf{h} \in U_B} \mathbf{g}^\top \mathbf{h} \tag{10.3}$$

subject to:

$$\mathbf{g}^\top \mathbf{g} = \mathbf{h}^\top \mathbf{h} = 1, \quad \mathbf{h}^\top \mathbf{h}_i = 0, \mathbf{g}^\top \mathbf{g}_i = 0, \quad i = 1, ..., j-1$$

As a result, we obtain the following optimization function over axes $\mathbf{u}, \mathbf{v}$:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^\top A^\top B \mathbf{v} \quad \text{s.t.} \quad \|A\mathbf{u}\|^2 = 1, \quad \|B\mathbf{v}\|^2 = 1.$$

To solve this problem we first perform a "QR" factorization of A and B. A "QR" factorization of a matrix A is a Grahm-Schmidt process resulting in an orthonormal set of vectors arranged as the columns of a matrix Q_A whose column space is equal to the column space of A, and a matrix R_A which contains the coefficients of the linear combination of the columns of Q_A such that $A = Q_A R_A$. Since orthoganilzation is not unique, the Grahm-Schmidt process perfroms the orthogonalization such that R_A is an upper-diagonal matrix. Likewise let $B = Q_B R_B$. Because the column spaces of A and Q_A are the same, then for every $\mathbf{u}$ there exists a $\hat{\mathbf{u}}$ such that $A\mathbf{u} = Q_A \hat{\mathbf{u}}$. Our optimization problem now becomes:

$$\max_{\hat{\mathbf{u}},\hat{\mathbf{v}}} \hat{\mathbf{u}}^\top Q_A^\top Q_B \hat{\mathbf{v}} \quad \text{s.t.} \quad \|\hat{\mathbf{u}}\|^2 = 1, \quad \|\hat{\mathbf{v}}\|^2 = 1.$$

The solution of this problem is when $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are the leading singular vectors of $Q_A^\top Q_B$. The singular value decomposition (SVD) of any matrix E is a decomposition $E = UDV^\top$ where the columns of U are the leading eigenvectors of $EE^\top$, the rows of $V^\top$ are the leading eigenvectors of $E^\top E$ and D is a diagonal matrix whose entries are the corresponding square eigenvalues (note that the eigenvalues of $EE^\top$ and $E^\top E$ are the same). The SVD decomposition has the property that if we keep only the first q leading eigenvectors then $UDV^\top$ is the closest (in least squares sense) rank q matrix to E.

Therefore, let $\hat{U}D\hat{V}^\top$ be the SVD of $Q_A^\top Q_B$ using the first q eigenvectors. Then, our sought after axes $U = [\mathbf{u}_1, ..., \mathbf{u}_q]$ is simply $R_A^{-1}\hat{U}$ and likewise and the axes $V = [\mathbf{v}_1, ..., \mathbf{v}_q]$ is equal to $R_B^{-1}\hat{V}$. The axes are called "canonical vectors", and the vectors $\mathbf{g}_i = A\mathbf{u}_i$ (mutually orthogonal) are called "variates". The concept of principal angles is due to Jordan in 1875, where Hotelling in 1936 is the first to introduce the recursive definition above.

Given a new vector $\mathbf{x} \in R^k$ the resulting vector $\mathbf{y}$ can be found by solving the linear system $U^\top \mathbf{x} = V^\top \mathbf{y}$ (since our assumption is that in the new basis the coordinates of $\mathbf{x}$ and $\mathbf{y}$ are similar).

To conclude, the relationship between A and B is captured by creating similar variates, i.e., creating subspaces of dimension q such that the projections of the input vectors and the output vectors have similar coordinates. The process for obtaining the two q-dimensional subspaces is by performing a QR factorization of A and B followed by an SVD. Here again the spectral analysis of the input and output data matrices plays a pivoting role in the input/output association.