Skript zur Vorlesung

Maschinelles Lernen 1

Wintersemester 2006/2007

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: krm@cs.tu-berlin.de

Fisher Discriminant Analysis and Least Squares Regression

1 Fisher Discriminant Analysis (FDA)

[**subsumption**] As discussed, high dimensional data are ubiques in real world problems – but they are hard to handle. Effective ways to reduce the dimensionality are needed. We have seen that PCA is the linear projection to a lower dimensional feature space that best representas the data in a least-squares sense. In this lecture we will study Fisher (linear) Discriminant Analysis (FDA) which determines a projection that best separates multi-class data in a least-squares sense. Furthermore there is a close link to Linear Discriminant Analysis (LDA) and Fisher Discriminant is related to (the forthcoming) Support Vector Machines (SVMs).

[**reference**] Duda, Hart, Stork, "Pattern Classification" (2nd ed.), pp. 117.

Let be given d-dimensional data vectors

$$x_1, \ldots, x_n \in \mathbb{R}^d$$

which are devided into two subsets (classes) $\mathcal{D}_1$ and $\mathcal{D}_2$.

Missing Figure: Motivation of Fisher criterion – see separate sheet

Denoting the desired projection vector by $w \in \mathbb{R}^d$ the projected data points are

$$y_k = w^\top x_k \qquad (k = 1, \ldots, n)$$

which are devided into two classes $\mathcal{Y}_1$ and $\mathcal{Y}_2$ according to

$$\mathcal{Y}_i := \{w^\top x \,|\, x \in \mathcal{D}_i\} \qquad (i = 1, 2).$$

We now need to formalize how well an arbitrary projection vector w separates the classes $\mathcal{D}_1$ and $\mathcal{D}_2$ to formulate an according optimization problem. One obvious measure of separation is the distance of the projected sample means

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y = \frac{1}{n_i} \sum_{x \in \mathcal{D}_i} w^\top x = w^\top m_i$$

where $n_i = |\mathcal{D}_i| = \#$ of samples in class i and

$$m_i = \frac{1}{n_i} \sum_{x \in \mathcal{D}_i} x$$

is the mean of class i of the original data. The distance between the projected means is

$$|\tilde{m_1} - \tilde{m_2}| = |w^\top (m_1 - m_2)|.$$

The problem here is that this quantity is sensitive to a mere scaling of w: For instance, the vectors w_1 and $w_2 = 2w_1$ define the same separation, but the distance of the projected means is two times larger for w_2. Since scaling of w has no meaning for the goodness of the separation we could restrict the search for the best w to vectors of length 1. But we will take a different approach here, since the distance between the projected means alone is not a good measure of separation anyhow:

Missing Figure: Goodness-of-separation in projected data – see separate sheet

Although the distance between the projected means is the same, "case 2" represents a much better separation, because the smaller variance within the (projected) classes leads to a smaller overlap (= misclassification).
This observation leads to the refined approach of taking the ratio of the distance between the means by the sum of within class variances (of the projected data) as optimization criterion. The Fisher linear discriminant is defined as that vector w that maximizes

$$J(w) = \frac{|\tilde{m_1} - \tilde{m_2}|^2}{\tilde{s_1}^2 + \tilde{s_2}^2}$$

where

$$\tilde{s_i} := \sum_{y \in \mathcal{Y}_i} (y - \tilde{m_i})^2 \qquad (i = 1, 2)$$

is the *within-class scatter* of class i. In order to actually determine the w that maximizes $J(w)$, we need to formulate $J(w)$ explicitly as a function of w and determine the maximum. We define the *within-class scatter matrix*

$$S_W = S_1 + S_2 \qquad \text{where} S_i := \sum_{x \in \mathcal{D}_i} (x - m_i)(x - m_i)^\top$$

and the *between-class scatter matrix*

$$S_B := (m_1 - m_2)(m_1 - m_2)^\top.$$

Using these definitions we can formulate $J(w)$ as an explicit function of w in the following way:
numerator:

$$\begin{aligned} |\tilde{m_1} - \tilde{m_2}|^2 &= (w^\top m_1 - w^\top m_2)^2 \\ &= (w^\top (m_1 - m_2))^2 \\ &= w^\top (m_1 - m_2)(m_1 - m_2)^\top w \\ &= w^\top S_B w \end{aligned}$$

denominator:

$$\begin{aligned}\tilde{s}_i^2 &= \sum_{x\in\mathcal{D}_i}(w^\top x - w^\top m_i)^2 \\ &= \sum_{x\in\mathcal{D}_i} w^\top (x-m_i)(x-m_i)^\top w \\ &= w^\top S_i w \qquad (i=1,2)\end{aligned}$$

These calculations imply

$$J(w) = \frac{w^\top S_B w}{w^\top S_W w}$$

In order to determine the w that maximizes $J(w)$ we use the following theorem known from linear algebra (it was probably already used in the lecture on PCA):

Theorem: 1 *Let $A \in \mathbb{R}^{m\times m}$ be a symmetric matrix. Then the function*

$$f : \mathbb{R}^m \to \mathbb{R}; \quad x \mapsto x^\top A x$$

attains its maximum on the set $\{z \in \mathbb{R}^m \mid \|z\| = 1\}$ for an $x \in \mathbb{R}^m$ satisfying

$$\exists \lambda \in \mathbb{R} \quad Ax = \lambda x$$

(i.e., x is an Eigenvector of A).

Since $J(w) = J(\alpha w)$ for any $\alpha \in \mathbb{R}^{\neq 0}$ we can restrict the search for the optimal w to those satisfying $w^\top S_W w = 1$, i.e., we regard the problem

$$\max_w w^\top S_B w \qquad \text{s.t.} \quad w^\top S_W w = 1 \tag{1}$$

Note that the scatter matrices S_W and S_B are symmetric. While S_B is (at most) of rank 1, S_W may have full rank d if $n > d$. From now on we assume S_W to be invertible. If this is not the case, one needs regularization which is the topic of a later lecture. Defining $x := S_W^{1/2} w$ implies $w = S_W^{-1/2} x$ and $x^\top x = 1$ (if $S_W^{1/2}$ is symmetric, which is possible) such that eqn. (1) becomes equivalent to

$$\max_x x^\top S_W^{-1/2} S_B S_W^{-1/2} x \qquad \text{s.t.} \quad x^\top x = 1$$

for which we know by above theorem, the solution satisfies

$$\exists \lambda \in \mathbb{R} \quad S_W^{-1/2} S_B S_W^{-1/2} x = \lambda x$$

Multiplying by $S_W^{1/2}$ from the left yields

$$\begin{aligned}S_B S_W^{1/2} x &= \lambda S_W^{1/2} x \\ &= \lambda S_W^{1/2} S_W^{1/2} S_W^{-1/2} x\end{aligned}$$

where we inserted the identity matrix in the form $S_W^{1/2} S_W^{-1/2}$. Now we substitute back $x = S_W^{1/2} w$ and obtain

$$S_B w = \lambda S_W w \tag{2}$$

So far we have shown that the w which maximizes $J(w)$ satisfies above equation for some $\lambda \in \mathbb{R}$. Equation (2) is known as *generalized Eigenvalue problem*. There are good ways to solve it for general matrices but in our case we can obtain a much simpler solution. Due to the special form of S_B the vector $S_B w$ has always the same direction as $m_1 - m_2$: An arbitray vector $x \in \mathbb{R}^d$ which is perpendicular to $m_1 - m_2$ is also perpendicular to $S_B w$ which can be seen as follows: $\langle x, m_1 - m_2 \rangle = 0$ implies

$$\begin{aligned} \langle x, S_B w \rangle &= \langle x, (m_1 - m_2)(m_1 - m_2)^\top w \rangle \\ &= (m_1 - m_2)^\top w \langle x, m_1 - m_2 \rangle \\ &= 0 \end{aligned} \tag{3}$$

That means there exists an $\alpha \in \mathbb{R}$ such that

$$S_B w = \alpha(m_1 - m_2)$$

holds for the w that maximizes $J(w)$. Plugging this into equation (2) we obtain

$$\begin{aligned} w &= \frac{1}{\lambda} S_W^{-1} S_B w \\ &= \frac{\alpha}{\lambda} S_W^{-1}(m_1 - m_2) \end{aligned}$$

Since the scaling of w is irrelevant for the value of $J(w)$, we can conclude that (also)

$$w := S_W^{-1}(m_1 - m_2) \quad \text{maximizes } J(w)$$

[**final remark**] Since the within-class scatter matrix is proportional to the sample covariance matrix, it is obvious that the Fisher discriminant vector w is equivalent to the projection vector of Linear Discriminant Analysis.

2 Least Squares Regression (LSR)

[**subsumption**] Having discussed some basic classification methods, we will introduce now the regression problem and investigate its simplest case of linear regression with minimum squared error as optimization criterion. It will turn out that it is linked closely to Fisher Discriminant (and also to linear perceptrons).

[**reference**] Duda, Hart, Stork, "Pattern Classification" (2nd ed.), pp. 240.

Given input vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ and output values $y_1, \ldots, y_n \in \mathbb{R}$ the problem of linear regression is to find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that the outputs of the linear regression function

$$\hat{y_k} = w^\top x_k + b \qquad (k = 1, \ldots, n)$$

fit well the true outputs y_k. The goodness-of-fit is most often measured as least squared error:

$$\sum_{k=1}^{n} (y_k - \hat{y_k})^2 \,. \tag{4}$$

In order to determine the optimal solution $(w; b)$ in this sense, we will introduce some notions:

$$X = \begin{pmatrix} x_1^\top & 1 \\ \vdots & \vdots \\ x_n^\top & 1 \end{pmatrix} \in \mathbb{R}^{n\times(d+1)}$$
$$y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$$
$$v = \left(\begin{smallmatrix} w \\ b \end{smallmatrix}\right) \in \mathbb{R}^{d+1}$$

Using these notions minimizing eq. (4) with respect to w and b is equivalent to

$$\min_v \|Xv - y\|^2. \tag{5}$$

We determine the minimum of this error function by setting the gradient to zero:

$$\begin{aligned} \frac{\partial}{\partial v}\|Xv - y\|^2 = 2X^\top(Xv - y) &\stackrel{!}{=} 0 \\ \Leftrightarrow 2X^\top Xv - 2X^\top y &= 0 \\ \Leftrightarrow X^\top Xv &= X^\top y \end{aligned} \tag{6}$$

If $X^\top X$ is invertible, the last equation is equivalent to

$$v = (X^\top X)^{-1} X^\top y$$

The quantity $X^\dagger := (X^\top X)^{-1} X^\top$ is called pseudo-inverse of X. It can also be defined in the (rare) case that $X^\top X$ is singular such that $X^\dagger y$ is the least squares solution. Since the Hessian

$$\frac{\partial^2}{\partial^2 v}\|Xv - y\|^2 = 2X^\top X$$

is positive semi-definite, the found extreme at v is a minimum.

Linear regression can also be used for classification if the output values y_k are taken as the labels of the classification problem. We will show in the follwing that for a proper choice of y Least Squares Regression (LSR) leads to the same solution as Fisher Discriminant Analysis. We use the same notions as in Sec. 1 (Fisher Discriminant) and start with the linear regression problem of mapping samples $x \in \mathcal{D}_1$ to $\frac{n}{n_1}$ and samples $x \in \mathcal{D}_2$ to $-\frac{n}{n_2}$: Let X_1 be the $d \times n_1$ matrix whose columns are the elements of $\mathcal{D}_1$ and X_2 the matrix of elements of $\mathcal{D}_2$. (The ordering of the columns in X_1 and X_2 is irrelevant.) Let $\mathbb{1}_m$ be the column vector of m ones. Then our regression problem (5) is given by

$$X = \begin{pmatrix} X_1^\top & \mathbb{1}_{n_1} \\ X_2^\top & \mathbb{1}_{n_2} \end{pmatrix} \qquad \text{and} \qquad y = \begin{pmatrix} \frac{n}{n_1} \cdot \mathbb{1}_{n_1} \\ -\frac{n}{n_2} \cdot \mathbb{1}_{n_2} \end{pmatrix}$$

According to eqn. (6) the LSR solution satisfies

$$\begin{aligned} & X^\top X \left(\begin{smallmatrix} w \\ b \end{smallmatrix}\right) = X^\top y \\ \Leftrightarrow \quad & \begin{pmatrix} X_1 & X_2 \\ \mathbb{1}_{n_1}^\top & \mathbb{1}_{n_2}^\top \end{pmatrix} \begin{pmatrix} X_1^\top & \mathbb{1}_{n_1} \\ X_2^\top & \mathbb{1}_{n_2} \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \\ \mathbb{1}_{n_1}^\top & \mathbb{1}_{n_2}^\top \end{pmatrix} \begin{pmatrix} \frac{n}{n_1} \cdot \mathbb{1}_{n_1} \\ -\frac{n}{n_2} \cdot \mathbb{1}_{n_2} \end{pmatrix} \\ \Leftrightarrow & \begin{pmatrix} X_1 X_1^\top + X_2 X_2^\top & X_1 \mathbb{1}_{n_1} + X_2 \mathbb{1} n_2 \\ \mathbb{1}_{n_1}^\top X_1^\top + \mathbb{1}_{n_2}^\top X_2^\top & \mathbb{1}_{n_1}^\top \mathbb{1}_{n_1} + \mathbb{1}_{n_2}^\top \mathbb{1}_{n_2} \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} n\left(\frac{1}{n_1} X_1 \mathbb{1}_{n_1} - \frac{1}{n_2} X_2 \mathbb{1}_{n_2}\right) \\ \frac{n}{n_1} n_1 - \frac{n}{n_2} n_2 \end{pmatrix} \end{aligned}$$

and due to $X_i \mathbb{1}_{n_i} = n_i m_i$ we can continue

$$\Leftrightarrow \begin{pmatrix} S_W + n_1 m_1 m_1^\top + n_2 m_2 m_2^\top & n_1 m_1 + n_2 m_2 \\ n_1 m_1^\top + n_2 m_2^\top & n_1 n_2 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} n(m_1 - m_2) \\ 0 \end{pmatrix} \tag{7}$$

The second equation calculates to

$$(n_1 m_1^\top + n_2 m_2^\top) w + nb = 0 \quad \Leftrightarrow \quad b = -m^\top w$$

where $m = \frac{1}{n} \sum x = \frac{1}{n}(n_1 m_1 + n_2 m_2)$ is the mean of all samples $x \in \mathcal{D}_1 \cup \mathcal{D}_2$. We plug $b = -m^\top w$ into the first equation of (7) to get

$$\left(S_W + n_1 m_1 m_1^\top + n_2 m_2 m_2^\top - n_1 m_1 m^\top - n_2 m_2 m^\top\right) w = n(m_1 - m_2)$$

Since

$$n_1 m_1 m^\top = n_1 m_1 \tfrac{1}{n}(n_1 m_1^\top + n_2 m_2^\top) = \tfrac{n_1^2}{n} m_1 m_1^\top + \tfrac{n_1 n_2}{n} m_1 m_2^\top$$

(analog for $n_2 m_2 m^\top$) and

$$n_1 = \frac{n_1(n_1 + n_2)}{n_1 + n_2} = \frac{n_1^2 + n_1 n_2}{n}$$

we can continue

$$\begin{aligned} &\left(\tfrac{1}{n} S_W + \tfrac{n_1}{n} m_1 m_1^\top + \tfrac{n_2}{n} m_2 m_2^\top - \tfrac{n_1^2}{n^2} m_1 m_1^\top - \tfrac{n_1 n_2}{n^2} m_1 m_2^\top - \tfrac{n_2^2}{n^2} m_2 m_1^\top - \tfrac{n_1 n_2}{n^2} m_2 m_2^\top\right) w \\ &\qquad = m_1 - m_2 \\ \Leftrightarrow &\left(\tfrac{1}{n} S_W + \tfrac{n_1 n_2}{n}(m_1 - m_2)(m_1 - m_2)^\top\right) w = m_1 - m_2 \\ \Leftrightarrow &\left(\tfrac{1}{n} S_W + \tfrac{n_1 n_2}{n} S_B\right) w = m_1 - m_2 \end{aligned}$$

As we have seen in FDA, eq. (3), $S_B w$ is always (i.e., for all w) in the direction of $m_1 - m_2$. So we can find an $\alpha \in \mathbb{R}$ such that

$$S_B w = \alpha(m_1 - m_2)$$

We obtain

$$\begin{aligned} &\tfrac{1}{n} S_W w + \tfrac{n_1 n_2}{n} \alpha(m_1 - m_2) = m_1 - m_2 \\ \Leftrightarrow \quad &\tfrac{1}{n} S_W w = (m_1 - m_2)\left(1 - \tfrac{n_1 n_2}{n} \alpha\right) \\ \Leftrightarrow &w = n\left(1 - \tfrac{n_1 n_2}{n} \alpha\right) S_W^{-1} w \end{aligned}$$

which is identical to the solution of Fisher's linear discriminant, except for the irrelevant scaling factor $n(1 - \frac{n_1 n_2}{n} \alpha)$.