Link between LDA and OLS

Jean-Philippe Vert

June 9, 2011

This is the solution to exercise 4.2 of [1] which shows a link between linear discriminant analysis (LDA) and ordinary least squares (OLS) in the binary case.

We have features $x \in \mathbb{R}^p$ and a two-class response, with class sizes N_1, N_2. The training patterns are denoted $x_1, \ldots, x_N \in \mathbb{R}^p$, stored in the $n \times p$ matrix X. We encode the class of each training point in the real number $y_i = -N/N_1$ for patterns x_i in class 1, and $y_i = N/N_2$ for patterns x_i in class 2.

(a) From equation (4.11) in [1] we know that, in the binary case, the LDA rule classifies a pattern x to class 2 if

$$x^\top \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^\top \hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^\top \hat{\Sigma}^{-1}\hat{\mu}_1 + \log\left(\frac{N_1}{N}\right) - \log\left(\frac{N_2}{N}\right) , \quad (1)$$

and class 1 otherwise.

(b) Let us introduce a few more notations. Let $U_i \in \mathbb{R}^n$ be the class indicator vector of class i, and $U = U_1 + U_2$ be the vector with all entries equal to 1. When we encode class 1 (resp. class 2) by the real number $a_1 = -N/N_1$ (resp. $a_2 = N/N2$), the vector of labels becomes $Y = a_1 U_1 + a_2 U_2$.

We consider the minimization of the least square criterion for $\beta \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$:

$$RSS(\beta, \beta_0) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta^\top x_i\right)^2 = \left(Y - \beta_0 U - X\beta\right)^\top \left(Y - \beta_0 U - X\beta\right) .$$

This convex criterion is minimized when its gradient vanishes, which gives:

$$\nabla_\beta RSS = 2X^\top X\beta - 2X^\top Y + 2\beta_0 X^\top U = 0 ,$$

and

$$\nabla_{\beta_0} RSS = 2U^\top U\beta_0 - 2U^\top \left(Y - X\beta\right) = 2N\beta_0 - 2U^\top \left(Y - X\beta\right) = 0 .$$

From the second condition we obtain:

$$\hat{\beta}_0 = \frac{1}{N} U^\top (Y - X\beta) \;, \tag{2}$$

which we can plug in the first one to obtain the optimality condition for β:

$$\left(X^\top X - \frac{1}{n} X^\top U U^\top X \right) \hat{\beta} = X^\top Y - \frac{1}{N} X^\top U U^\top Y \,. \tag{3}$$

Let us now try to simplify the left- and right-hand sides of (3). Notice that, with our notations, we have $X^\top U_i = N_i \hat{\mu}_i$ for $i = 1, 2$.

- *Left-hand side.* Because $X^\top U = X^\top (U_1 + U_2) = N_1\hat{\mu}_1 + N_2\hat{\mu}_2$, we can rewrite the matrix on the l.h.s. of (3) as:

  $$X^\top X - \frac{1}{N} \left(N_1^2 \hat{\mu}_1 \hat{\mu}_1^\top + N_2^2 \hat{\mu}_2 \hat{\mu}_2^\top + N_1 N_2 \hat{\mu}_1 \hat{\mu}_2^\top + N_1 N_2 \hat{\mu}_2 \hat{\mu}_1^\top \right) \,. \tag{4}$$

  The estimate of the covariance matrix used in LDA is given by:

  $$\begin{aligned} (N-2)\hat{\Sigma} &= \sum_{i:y_i = a_1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^\top + \sum_{i:y_i = a_2} (x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^\top \\ &= X^\top X - N_1 \hat{\mu}_1 \hat{\mu}_1^\top - N_1 \hat{\mu}_2 \hat{\mu}_2^\top \end{aligned}$$

  Defining $\hat{\Sigma}_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$, we deduce:

  $$\begin{aligned} &(N-2)\hat{\Sigma} + \frac{N_1 N_2}{N} \hat{\Sigma}_B \\ &\quad = X^\top X + \left(\frac{N_1 N_2}{N} - N_1 \right) \hat{\mu}_1 \hat{\mu}_1^\top + \left(\frac{N_1 N_2}{N} - N_2 \right) \hat{\mu}_2 \hat{\mu}_2^\top - \frac{N_1 N_2}{N} \hat{\mu}_2 \hat{\mu}_1^\top - \frac{N_1 N_2}{N} \hat{\mu}_1 \hat{\mu}_2^\top \\ &\quad = X^\top X - \frac{N_1^2}{N} \hat{\mu}_1 \hat{\mu}_1^\top - \frac{N_2^2}{N} \hat{\mu}_2 \hat{\mu}_2^\top - \frac{N_1 N_2}{N} \hat{\mu}_2 \hat{\mu}_1^\top - \frac{N_1 N_2}{N} \hat{\mu}_1 \hat{\mu}_2^\top \,, \end{aligned}$$

  which is exactly equal to (4)

- *Right-hand side.* The first term is equal to:

  $$\begin{aligned} X^\top Y &= X^\top (a_1 U_1 + a_2 U_2) \\ &= a_1 N_1 \hat{\mu}_1 + a_2 N_2 \hat{\mu}_2 \,. \end{aligned}$$

  The second term is equal to:

  $$\begin{aligned} \frac{1}{N} X^\top U U^\top Y &= \frac{1}{N} (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)(a_1 N_1 + a_2 N_2) \\ &= \frac{a_1 N_1^2 + a_2 N_1 N_2}{N} \hat{\mu}_1 + \frac{a_2 N_2^2 + a_1 N_1 N_2}{N} \hat{\mu}_2 \,. \end{aligned}$$

Combining both terms (and using again the fact that $N = N_1 + N_2$) we obtain that the r.h.s. of (3) is equal to:

$$\frac{a_1 N_1 N_2 - a_2 N_1 N_2}{N}\hat{\mu}_1 + \frac{a_2 N_1 N_2 - a_1 N_1 N_2}{N}\hat{\mu}_2 = \frac{N_1 N_2}{N}(a_1 - a_2)(\hat{\mu}_1 - \hat{\mu}_2) .$$

Combining the simplifications for the l.h.s. and r.h.s. of (3) shows that $\hat{\beta}$ minimizes RSS if and only if it satisfies:

$$\left[(N-2)\hat{\Sigma} + \frac{N_1 N_2}{N}\hat{\Sigma}_B\right]\hat{\beta} = \frac{N_1 N_2}{N}(a_1 - a_2)(\hat{\mu}_1 - \hat{\mu}_2) . \quad (5)$$

Taking the encoding $a_1 = -N/N_1$ and $a_2 = N/N_2$, we get

$$a_1 - a_2 = -\frac{N}{N_1} - \frac{N}{N_2} = -\frac{N(N_1 + N_2)}{N_1 N_2} = -\frac{N^2}{N_1 N_2} ,$$

so the optimality condition (5) becomes

$$\left[(N-2)\hat{\Sigma} + \frac{N_1 N_2}{N}\hat{\Sigma}_B\right]\hat{\beta} = N(\hat{\mu}_2 - \hat{\mu}_1) . \quad (6)$$

(c) Let the real number $c = (\hat{\mu}_2 - \hat{\mu}_1)^\top \hat{\beta}$. Then we immediately get:

$$\hat{\Sigma}_B\hat{\beta} = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^\top \beta = c(\hat{\mu}_2 - \hat{\mu}_1) ,$$

showing that $\hat{\Sigma}_B\hat{\beta}$ is in the direction of $(\hat{\mu}_2 - \hat{\mu}_1)$. Combined with (6), this shows that $\hat{\Sigma}\hat{\beta}$ is also in the direction of $(\hat{\mu}_2 - \hat{\mu}_1)$ (as a difference of two terms in this direction), i.e.,

$$\hat{\beta} \sim \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) .$$

This shows that the OLS estimator is identical to the LDA coefficient, up to a scalar multiple.

(d) Since (5) holds for any encoding a_1 and a_2, the result also holds for any encoding.

(e) Note that with the encoding $a_1 = -N/N_1$ and $a_2 = N/N_2$ we have

$$U^\top Y = a_1 N_1 + a_2 N_2 = -N + N = 0 .$$

We deduce from the optimality condition (2) the value of $\hat{\beta}_0$:

$$\hat{\beta}_0 = -\frac{1}{N}U^\top X\hat{\beta} = -\left(\frac{N_1}{N}\hat{\mu}_1^\top + \frac{N_2}{N}\hat{\mu}_2^\top\right)\hat{\beta} .$$

The decision function for a pattern $x \in \mathbb{R}^p$ is

$$f(x) = x^\top \hat{\beta} + \hat{\beta}_0 = \left(x^\top - \frac{N_1}{N} \hat{\mu}_1^\top - \frac{N_2}{N} \hat{\mu}_2^\top \right) \hat{\beta}$$

Since we have $\hat{\beta} = \lambda \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$ for some $\lambda \in \mathbb{R}$, the decision whether or not:

$$x^\top \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \left(\frac{N_1}{N} \hat{\mu}_1^\top + \frac{N_2}{N} \hat{\mu}_2^\top \right) \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

When $N_1 = N_2$, this simplifies to the LDA decision function (1).

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2001.