ISYE 7405

Homework 7

Jacob Aguirre *Email: aguirre@gatech.edu*
Instructor: Dr. Shihao Yang

1. (Weighting for clustering) Show that weighted Euclidean distance

$$d^{(w)}(x, x') = \sum_{j=1}^{p} w_j (x_j - x'_j)^2 / \sum_{j=1}^{p} w_j$$

satisfies

$$d^{(w)}(x, x') = d(z, z') = \sum_{j=1}^{p} (z_j - z'_j)^2,$$

where $z_j = x_j (w_j / \sum_{i=1}^{p} w_j)^{1/2}$, and z'_j is similarly defined. Thus weighted Euclidean distance based on x is equivalent to unweighted Euclidean distance based on z.
Solution. The conclusion is obvious:

$$d(z, z') = \sum_{j=1}^{p} (z_j - z'_j)^2 = \sum_{j=1}^{p} (x_j - x'_j)^2 w_j / \sum_{i=1}^{p} w_j = d^{(w)}(x, x'). \square$$

2. Consider a two-class classification problem. The predictors (features) are $x \in \mathcal{R}^p$. Among the n observed data points, n_1 are in class 1 and n_2 are in class 2. The two classes are coded as $y = -n/n_1$ and n/n_2 respectively.

1. Show that the LDA classifies to class 2 if

   $$x^\top \widehat{\Sigma}^{-1} (\widehat{\mu}_2 - \widehat{\mu}_1) > \text{a threshold}$$

   and class 1 otherwise.

   Solution. We have shown in class that $x^\top \widehat{\Sigma}^{-1} (\widehat{\mu}_2 - \widehat{\mu}_1)$ is the estimated linear discriminant function. Therefore, the conclusion follows. Alternatively, the classification rule is equivalent to the ML classification rule, which is also established in class. $\square$

2. Consider minimization of the least squares criterion

   $$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1^\top x_i)^2.$$

   Show that the Solution $\widehat{\beta}_1$ satisfies

   $$\left\{ (n-2)\widehat{\Sigma} + \frac{n_1 n_2}{n} \widehat{\Sigma}_B \right\} \widehat{\beta}_1 = n(\widehat{\mu}_2 - \widehat{\mu}_1),$$

where $\widehat{\Sigma}_B = (\widehat{\mu}_2 - \widehat{\mu}_1)(\widehat{\mu}_2 - \widehat{\mu}_1)^\top$.

Solution. The Least Squares Criterion leads to

$$\begin{aligned}
\widehat{\beta} &= (X'X)^{-1}X'y, \\
\widehat{\beta} &= \begin{pmatrix} n & 1'X \\ X'1 & X'X \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ n(\widehat{\mu}_2 - \widehat{\mu}_1) \end{pmatrix}.
\end{aligned}$$

The lower corner of the block matrix inverse is the inverse of the following matrix:

$$\begin{aligned}
X'X - \frac{1}{n}X'11'X &= \sum (x_i - \widehat{\mu})(x_i - \widehat{\mu})' \\
&= \sum_1 (x_i - \widehat{\mu}_1)(x_i - \widehat{\mu}_1)' + n_1(\widehat{\mu} - \widehat{\mu}_1)(\widehat{\mu} - \widehat{\mu}_1)' \\
&\quad + \sum_2 (x_i - \widehat{\mu}_2)(x_i - \widehat{\mu}_2)' + n_2(\widehat{\mu} - \widehat{\mu}_2)(\widehat{\mu} - \widehat{\mu}_2)' \\
&= (n-2)\widehat{\Sigma} + n_1\frac{n_2^2}{n^2}(\widehat{\mu}_2 - \widehat{\mu}_1)(\widehat{\mu}_2 - \widehat{\mu}_1)' + n_2\frac{n_1^2}{n^2}(\widehat{\mu}_2 - \widehat{\mu}_1)(\widehat{\mu}_2 - \widehat{\mu}_1)' \\
&= (n-2)\widehat{\Sigma} + \frac{n_1 n_2}{n}\widehat{\Sigma}_B.
\end{aligned}$$

Therefore, we have

$$\left\{(n-2)\widehat{\Sigma} + \frac{n_1 n_2}{n}\widehat{\Sigma}_B\right\}\widehat{\beta}_1 = n(\widehat{\mu}_2 - \widehat{\mu}_1).\square$$

3. Hence show that $\widehat{\Sigma}_B\widehat{\beta}_1$ is in the direction $(\widehat{\mu}_2 - \widehat{\mu}_1)$ and thus $\widehat{\beta}_1 \propto \widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$. Therefore, the least square regression coefficient is identical to the LDA coefficient up to a scale multiple.

   Solution. Since $(\widehat{\mu}_2 - \widehat{\mu}_1)'\widehat{\beta}_1$ is a scalar, we know

   $$\widehat{\Sigma}_B\widehat{\beta}_1 = (\widehat{\mu}_2 - \widehat{\mu}_1)(\widehat{\mu}_2 - \widehat{\mu}_1)'\widehat{\beta}_1 \propto (\widehat{\mu}_2 - \widehat{\mu}_1),$$

   and therefore, $\widehat{\Sigma}\widehat{\beta}_1 \propto (\widehat{\mu}_2 - \widehat{\mu}_1)$, and $\widehat{\beta}_1 \propto \widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$. $\square$

4. Show that this results holds for any distinct coding of the two classes.

   Solution. For any distinct coding of $y \in \{A, B\}$, there is a linear and one-to-one mapping from $z \in \{-n_1/n, n_2/n\}$ to $y = c + dz \in \{A, B\}$. The least squares criterion becomes

   $$\min_{\beta_0,\beta_1}\sum_{i=1}^n (y_i - \beta_0 - \beta_1^\top x_i)^2 \Leftrightarrow \min_{\beta_0,\beta_1}\sum_{i=1}^n (c + dz_i - \beta_0 - \beta_1^\top x_i)^2 \Leftrightarrow \min_{\gamma_0,\gamma_1}\sum_{i=1}^n (z_i - \gamma_0 - \gamma_1^\top x_i)^2,$$

   where $\gamma_0 = (\beta_0 - c)/d$ and $\gamma_1 = \beta_1/d$. Therefore, the problem reduce to the original problem, implying that $\gamma_1 \propto \widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$ and therefore $\widehat{\beta}_1 \propto \widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$. $\square$

3. Consider the LDA procedure. Suppose we transform the original predictors X to $\widehat{Y} = X(X^\top X)^{-1}X^\top Y = X\widehat{\beta}$, the linear regression fit. Similarly, for any input x, we get a transformed scalar $\widehat{y} = x^\top\widehat{\beta}$. Show that LDA using $\widehat{Y}$ is identical to LDA in the original space.

Solution. Now X should contain the constant 1. According to problem 2, we know that LDA is equivalent to using the discriminant function $x^\top\widehat{\beta} >$ threshold.

If we transform our data and get a scalar $\widehat{y} = x^\top\widehat{\beta}$, then LDA based on this scalar is simply $\widehat{y} >$ threshold, which is the same as the original LDA. $\square$

4. Show that the criterion

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \{1 - y_i f(x_i)\}_+ + \frac{\lambda}{2}||\beta||^2$$

is equivalent to the original SVM criterion of

$$\begin{aligned} &\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \xi_i \\ &s.t.\ \xi_i \geq 0, \quad y_i(x_i^\top\beta + \beta_0) \geq 1 - \xi_i, \forall i. \end{aligned}$$

Solution. In the SVM solution, we have

$$\xi_i \geq \max\{0, 1 - y_i f(x_i)\}.$$

We argue that in order to attain the minimum, it must be true that

$$\xi_i = \max\{0, 1 - y_i f(x_i)\} = \{1 - y_i f(x_i)\}_+.$$

Otherwise, we can reduce the objective function by letting the inequality being the equality above. Consequently, the SVM criterion is equivalent to

$$\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \{1 - y_i f(x_i)\}_+.$$

By taking $C = \lambda^{-1}$, we can prove the conclusion. $\square$