**STAT 383C: Statistical Modeling I** **Fall 2015**

# Lecture 13 — October 11

*Lecturer: Purnamrita Sarkar* *Scribe: Calvin Tsay*

**Disclaimer:** These scribe notes have been slightly proofread and may have typos etc.

**Note:** The latex template was borrowed from EECS, U.C. Berkeley.

## 13.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) approximates the Bayes classifier rule by modeling conditional class densities as multivariate normals. For the classes $C \in 1, ..., K$ and a feature vector $X \in \mathbb{R}^p$ this can be expressed:

$$P(X = x|C = j) = N(\mu_j, \Sigma) \tag{13.1}$$

Note each class $j$ has its own mean $\mu_j \in \mathbb{R}^p$, but the classes together share a covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. The LDA process can be decomposed into the following steps:

**Step 1: Center Data**

Given the data $X_1, ..., X_n$ and $y_1, ..., y_n$, we first center the data around their mean:

$$X'_i = X_i - \bar{X} \tag{13.2}$$

**Step 2: Calculate Estimators**

We can then estimate the parameters given from modeling the class densities as multivariate normals. The class means can be estimated:

$$\hat{\mu}_j = \frac{\sum\limits_{y_i=j} X'_j}{\sum\limits_{y_i=j} 1} \tag{13.3}$$

where the notation $\sum\limits_{y_i=j}$ represents $\sum\limits_{i} \forall\ i : y_i = j$. The pooled covariance matrix can be estimated:

$$\hat{\Sigma} = \frac{\sum\limits_{j}\sum\limits_{y_i=j} (X'_i - \hat{\mu}_j)(X'_i - \hat{\mu}_j)^T}{n-k} \tag{13.4}$$

Lastly, the probability of class $j$ can be estimated:

$$\hat{\pi}_j = \frac{1}{n}\sum_{y_i=j} 1 \tag{13.5}$$

We note that to classify some data point, we can then solve the following:

$$\underset{j}{argmin}\ \frac{1}{2}(X' - \hat{\mu}_j)^T \hat{\Sigma}^{-1}(X' - \hat{\mu}_j) - log(\hat{\pi}_j) \tag{13.6}$$

**Step 3: Sphere Variables**

We sphere the data points and means using the following:

$$\tilde{X} = \hat{\Sigma}^{-1/2}X' \tag{13.7}$$

$$\tilde{\mu_j} = \hat{\Sigma}^{-1/2}\hat{\mu}_j \tag{13.8}$$

Since $\Sigma$ is the covariance matrix, $X \sim N(0, \Sigma)$, and the sphering effectively standardizes the covariance of all the treated variables:

$$cov(\tilde{X}) = cov(\hat{\Sigma}^{-1/2}X') = \hat{\Sigma}^{-1/2}cov(X')\hat{\Sigma}^{-1/2} = \hat{\Sigma}^{-1/2}\hat{\Sigma}\hat{\Sigma}^{-1/2} = I \tag{13.9}$$

The equation (10.6) can then be rewritten:

$$\underset{j}{argmin}\ \frac{1}{2}||\tilde{X} - \tilde{\mu_j}||_2^2 - log(\hat{\pi}_j) \tag{13.10}$$

Now we consider that $X'$ and $\mu_j$ are $p$-dimensional, but there are only $K$ clusters. Therefore, we should only require $K$ dimensions to classify data. Additionally, if the data are centered, we should only require $K-1$ dimensions. We thus desire some projection matrix $P_m$ to project the variables into a $K-1$ dimensional subspace spanned by $\tilde{\mu}_1, ..., \tilde{\mu}_K$ (if the means $\mu_1, ..., \mu_K$ are linearly independent). In general, we can write the projection of a variable $X$ as:

$$X = P_m X + P_m^{\perp} X \tag{13.11}$$

where $P_m X$ is the variable $X$ projected down onto a new set of basis functions determined by $P_m$ and $P_m^{\perp} X$ is the un-projected, "left-over" features. We can then write:

$$||\tilde{X} - \tilde{\mu_j}||_2^2 = ||P_m(\tilde{X} - \tilde{\mu_j})||_2^2 + ||P_m^{\perp}\tilde{X}||_2^2 \tag{13.12}$$

Noting that the two norms can be split up because $P_m$ and $P_m^{\perp}$ are orthogonal and the term $P_m^{\perp}\tilde{\mu}_j = 0$ since $P_m = span(\mu_1, ..., \mu_K)$. We also note that since the second term does not include any dependence on the class $j$:

$$\underset{j}{argmin}\ \frac{1}{2}||\tilde{X} - \tilde{\mu_j}||_2^2 = \underset{j}{argmin}\ \frac{1}{2}||p_m(\tilde{X} - \tilde{\mu_j})||_2^2 \tag{13.13}$$

**Step 4: Project using $P_m$**

To determine the matrix $P_m$ we first write $\hat{M}_{K \times p}$ as the matrix containing all the estimated means row-wise, such that:

$$\hat{M} = \begin{bmatrix} - & \hat{\mu}_1^T & - \\ - & \hat{\mu}_2^T & - \\ \vdots & \vdots & \vdots \\ - & \hat{\mu}_K^T & - \end{bmatrix} \tag{13.14}$$

After sphering, the matrix is expressed:

$$\tilde{M} = \begin{bmatrix} - & \tilde{\mu}_1^T & - \\ - & \tilde{\mu}_2^T & - \\ \vdots & \vdots & \vdots \\ - & \tilde{\mu}_K^T & - \end{bmatrix} = \begin{bmatrix} - & (\hat{\Sigma}^{-1/2}\hat{\mu}_1)^T & - \\ - & (\hat{\Sigma}^{-1/2}\hat{\mu}_2)^T & - \\ \vdots & \vdots & \vdots \\ - & (\hat{\Sigma}^{-1/2}\hat{\mu}_K)^T & - \end{bmatrix} = \hat{M}\hat{\Sigma}^{-1/2} \tag{13.15}$$

We can then write:

$$\tilde{B} = \tilde{M}^T\tilde{M} = \hat{\Sigma}^{-1/2}\hat{M}^T\hat{M}\hat{\Sigma}^{-1/2} \tag{13.16}$$

with $\tilde{B}$ providing an indicator as to how separated the means $\mu_1, ..., \mu_K$ are. We then examine the eigenvalues and eigenvectors of matrix $\tilde{B}$, sorting the eigenvectors such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K$. Since the rank of $\hat{M}$ is $k-1$, the $\lambda_i \to 0 \; \forall \; i > k-1$, and we formulate $P_m$ as the top $k-1$ eigenvectors of $\tilde{B}$. Then $P_m$ projects $X$ into a $k-1$ dimensional subspace:

$$\tilde{X}_{K-1} = [\tilde{X}^T v_1, \tilde{X}^T v_2, ...\tilde{X}^T v_{K-1}] \tag{13.17}$$

where $v_i$ is the $i^{th}$ eigenvector of $\tilde{B}$. We can further introduce a new variable $w$ such that $w = \hat{\Sigma}^{-1/2}v_1$:

$$\tilde{X}^T v_1 = X'^T \underbrace{\hat{\Sigma}^{-1/2}v_1}_{w} = X'^T w \tag{13.18}$$

By the definition of eigenvalues/vectors we see that $w$ solves the following. Let $B = \hat{M}^T\hat{M}$.

$$\tilde{B}v_1 = \tilde{M}^T\tilde{M}v_1 = \hat{\Sigma}^{-1/2}\hat{M}^T\hat{M}\hat{\Sigma}^{-1/2}v_1 = \lambda_1 v_1 \tag{13.19}$$

$$\hat{\Sigma}^{-1}Bw = \lambda_1 w \tag{13.20}$$

$$w = \underset{||u||=1}{argmax} \frac{u^T B u}{u^T \hat{\Sigma} u} \tag{13.21}$$

The second direction will be obtained by optimizing over $u \perp w$. This is none other than a generalized eigenvalue problem. Moreover, this is also Fisher's discriminant analysis which Fisher arrived at by simply finding a direction such that when the data is projected on that direction, the inner cluster distance is maximized and in cluster variance is minimized so as to have maximum separation. Thus reduced rank LDA essentially just gives us the directions along which one can project the data points to maximally separate out the two clusters.

If the mean vectors were not linearly independent, then one can in fact find fewer than $k-1$ directions to project on.

This brings us to the full algorithm.

1. Center data
2. Estimate parameters including means, pooled variance and class proportions

3. Sphere data
4. Compute within cluster covariance matrix $B$ and project