Notes on Statistical and Machine Learning

Flexible Discriminants

Chapter: *17* **Prepared by:** *Chenxi Zhou*

This note is prepared based on *Chapter 12, Support Vector Machines and Flexible Discriminants* in Hastie, Tibshirani, and Friedman (2009).

I. Generalizing Linear Discriminant Analysis

1. **Advantages of LDA:** The classic LDA has the following advantages:
   (a) LDA is a simple prototype classifier. By saying a "prototype" classifier, we mean each class is represented by its centroid and a new observation is classified to the class with the closest centroid;
   (b) LDA is the estimated classifier if the observations are multivariate Gaussian in each class with a common variance;
   (c) The decision boundaries created by LDA are linear, leading to decision rules are simple to describe and implement;
   (d) LDA provides natural low-dimensional views of the data;
   (e) LDA produces satisfactory classification results, because of its simplicity and low variance.
2. **Disadvantages of LDA:** LDA can fail in some situations:
   (a) Often linear boundaries do *not* adequately separate the classes. We want to model irregular boundaries;
   (b) A *single* prototype per class is insufficient. In many situations, several prototypes per class are more appropriate;
   (c) In the case of having many predictors, LDA uses too many parameters, which are estimated with high variance, and its performance suffers.
3. **Three Ideas of Generalizing LDA:**
   (a) *Flexible Discriminant Analysis (FDA):* Recast the LDA problem as a nonparametric regression problem;
   (b) *Penalized Discriminant Analysis (PDA):* Fit an LDA model, and penalize its coefficients to be smooth or coherent in the spatial domain (e.g., an image);
   (c) *Mixture Discriminant Analysis (MDA):* Model each class by a mixture of two or more Gaussians with different centroids, but with every component Gaussian, both within and between classes, sharing the same covariance matrix.

II. Flexible Discriminant Analysis

1. **Main Idea:** The main idea of FDA is to perform LDA using linear regression on derived responses, which leads to nonparametric and flexible alternatives to LDA.

2. **Setup:** Assume the quantitative response variable G belonging to one of W classes $\mathcal{W} := \{1, \cdots, W\}$, and the feature variable is $\mathbf{x} \in \mathbb{R}^p$.

3. **Single Scoring:** Suppose $\theta : \mathcal{W} \to \mathbb{R}$ is a function that assigns scores to the class labels such that the transformed class labels are optimally predicted by linear regression on X. If the training sample has the form $\{(\mathbf{x}_i, g_i)\}_{i=1}^n$, where $g_i \in \mathcal{W}$ for all $i = 1, 2, \cdots, n$, we then solve
$$\underset{\boldsymbol{\beta},\theta}{\text{minimize}} \left\{ \sum_{i=1}^n \big(\theta(g_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\big)^2 \right\},$$
with certain restrictions on θ to avoid a trivial solution. Note that the preceding minimization problem produces a one-dimensional separation between the two classes.

4. **Multiple Scorings:** We can find up to $L \leq W - 1$ sets of independent scorings for the class labels, $\theta_1, \theta_2, \cdots, \theta_L$, and L corresponding linear maps
$$\eta_\ell(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_\ell, \qquad \text{for all } \ell = 1, \cdots, L,$$
chosen to be optimal for multiple regression in $\mathbb{R}^p$.

   Then, $\{\theta_\ell\}_{\ell=1}^L$ and $\{\boldsymbol{\beta}_\ell\}_{\ell=1}^L$ are chosen by minimizing the average squared residual, i.e.,
$$\text{ASR}(\theta_1, \cdots, \theta_L, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_L) := \frac{1}{n} \sum_{\ell=1}^L \left[\sum_{i=1}^n \big(\theta_\ell(g_i) - \mathbf{x}_i^\top \boldsymbol{\beta}_\ell\big)^2 \right].$$
   The set of scores is assumed to be mutually orthogonal and normalized with respect to an appropriate inner product to prevent trivial zero solutions.

5. **Generalizing the Linear Maps:** We can replace the linear regression fits $\eta_\ell(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_\ell$ by more flexible nonparametric fits, such as additive fits, spline models and MARS, in order to achieve a more flexible classifier than LDA.

   In the more general form, the regression problem is defined via the criterion
$$\widetilde{\text{ASR}}(\theta_1, \cdots, \theta_L, \eta_1, \cdots, \eta_L) = \frac{1}{n} \sum_{\ell=1}^L \left[\sum_{i=1}^n \big(\theta_\ell(g_i) - \eta_\ell(\mathbf{x}_i)\big)^2 + \lambda \cdot J(\eta_\ell) \right], \tag{1}$$
   where J is a regularizer appropriate for some forms of nonparametric regression.

6. **Computing the FDA Estimates:** We suppose that the nonparametric regression procedure can be represented by a linear operator; that is, there exists a linear operator $\mathbf{S}_\lambda$ such that the fitted value vector $\mathbf{Y}$ and the response vector $\widehat{\mathbf{Y}}$ are related by $\widehat{\mathbf{Y}} = \mathbf{S}_\lambda \mathbf{Y}$, where λ is a penalty parameter.

   The procedure of computing the FDA estimates is the following:

(1) *Create response matrix:* Create an $n \times W$ indicator response matrix $\mathbf{Y}$ from the responses g_i such that $y_{i,w} = 1$ if $g_i = w$, otherwise $y_{i,w} = 0$, for all $i = 1, 2, \cdots, n$ and $w = 1, 2, \cdots, W$;

(2) *Multivariate nonparametric regression:* Fit a multi-response, adaptive nonparametric regression of $\mathbf{Y}$ on $\mathbf{X}$, giving fitted values $\widehat{\mathbf{Y}}$. Let $\boldsymbol{\eta}^*$ be the vector of fitted regression functions;

(3) *Compute optimal scores:* Compute the eigen-decomposition of $\mathbf{Y}^\top \widehat{\mathbf{Y}} = \mathbf{Y}^\top \mathbf{S}_\lambda \mathbf{Y}$, where the eigenvectors $\boldsymbol{\Theta}$ are normalized so that $\boldsymbol{\Theta}^\top \mathbf{D}_\pi \boldsymbol{\Theta} = \mathbf{I}_W$. Here, $\mathbf{D}_\pi := \mathbf{Y}^\top \mathbf{Y}/n$ is a diagonal matrix of the estimated class prior probabilities;

(4) *Update the model from Step (1) using the optimal scores:* $\boldsymbol{\eta}(\mathbf{x}) = \boldsymbol{\Theta}^\top \boldsymbol{\eta}^*(\mathbf{x})$.

III. Penalized Discriminant Analysis

1. **More on FDA:** In (1), if we choose $\eta_\ell(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\top \boldsymbol{\beta}_\ell$ to be a function of transformed features $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^M$ and the penalty functional J to be quadratic, we can rewrite (1) as

$$\widetilde{\mathrm{ASR}}(\theta_1, \cdots, \theta_L, \eta_1, \cdots, \eta_L) = \frac{1}{n} \sum_{\ell=1}^{L} \left[\sum_{i=1}^{n} \big(\theta_\ell(g_i) - \mathbf{h}(\mathbf{x})^\top \boldsymbol{\beta}_\ell\big)^2 + \lambda \boldsymbol{\beta}_\ell^\top \boldsymbol{\Omega} \boldsymbol{\beta}_\ell \right], \quad (2)$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{M \times M}$ depends on the problem and the function space $\mathbf{h}$ resides.

2. **Penalized Discriminant Analysis:** The *penalized discriminant analysis*, or PDA, follows from the following steps:

(a) Enlarge the set of predictors $\mathbf{x} \in \mathbb{R}^p$ via a basis expansion $\mathbf{h} : \mathbb{R}^p \to \mathbb{R}^M$;

(b) Use the penalized LDA in the enlarged space, where the penalized Mahalanobis distance is given by

$$D(\mathbf{x}, \boldsymbol{\mu}) := (\mathbf{h}(\mathbf{x}) - \mathbf{h}(\boldsymbol{\mu}))^\top (\boldsymbol{\Sigma}_W + \lambda \boldsymbol{\Omega})^{-1} (\mathbf{h}(\mathbf{x}) - \mathbf{h}(\boldsymbol{\mu})),$$

where $\boldsymbol{\Sigma}_W$[1] is the within-class covariance matrix of the derived variables $\{\mathbf{h}(\mathbf{x}_i)\}_{i=1}^n$;

(c) Decompose the classification subspace using a penalized metric

$$\begin{aligned} &\text{maximize } \mathbf{u}^\top \boldsymbol{\Sigma}_B \mathbf{u} \\ &\text{subject to } \mathbf{u}^\top (\boldsymbol{\Sigma}_W + \lambda \boldsymbol{\Omega}) \mathbf{u} = 1, \end{aligned}$$

where $\boldsymbol{\Sigma}_B$ denotes the between-class covariance matrix.

[1] Note that here the subscript "W" is to denote this is the *w*ithin-class covariance matrix, and has nothing to do with the total number of classes W.

IV. Mixture Discriminant Analysis

1. **Motivation:** Linear discriminant analysis can be viewed as a *prototype* classifier — each class is represented by its centroid, and we classify an observation to the closest centroid using an appropriate metric.

   In many situations, a *single* prototype for each class is *not* sufficient to represent inhomogeneous classes. Mixture models are more appropriate.

2. **Gaussian Mixture Models:** A *Gaussian mixture model* for the w-th class has density

$$f(\mathbf{x} \,|\, G = w) = \sum_{r=1}^{R_w} \pi_{w,r} \phi(\mathbf{x}; \boldsymbol{\mu}_{w,r}, \boldsymbol{\Sigma}),$$

   where the mixing proportions $\{\pi_{w,r}\}_{r=1}^{R_w}$ sum to one. This has R_w prototypes for Class w and the same covariance matrix $\boldsymbol{\Sigma}$. Given such a model for each class, the class posterior probabilities are given by

$$\mathbb{P}(G = w \,|\, X = \mathbf{x}) = \frac{\sum_{r=1}^{R_k} \pi_{w,r} \phi(\mathbf{x}; \boldsymbol{\mu}_{w,r}, \boldsymbol{\Sigma}) \Pi_w}{\sum_{\ell=1}^{W} \sum_{s=1}^{R_w} \pi_{\ell,s} \phi(\mathbf{x}; \boldsymbol{\mu}_{\ell,s}, \boldsymbol{\Sigma}) \Pi_\ell},$$

   where Π_k represents the prior probabilities of Class k.

3. **Parameter Estimation:** We estimate the parameters by the method of maximum likelihood, i.e., we maximize

$$\sum_{w=1}^{W} \sum_{\{i | g_i = w\}} \log \left[\sum_{r=1}^{R_w} \pi_{w,r} \phi(\mathbf{x}_i; \boldsymbol{\mu}_{w,r}, \boldsymbol{\Sigma}) \Pi_w \right]. \tag{3}$$

   We can use the EM algorithm to compute the maximizer of (3), which alternates between the following two steps:

   (a) *E-step*: Given the current parameters, compute the responsibility of subclass $c_{w,r}$ within Class w for each of the class-w observations ($g_i = w$):

$$\frac{\pi_{w,r} \phi(\mathbf{x}_i; \boldsymbol{\mu}_{w,r}, \boldsymbol{\Sigma})}{\sum_{\ell=1}^{R_w} \pi_{w,\ell} \phi(\mathbf{x}_i; \boldsymbol{\mu}_{w,\ell}, \boldsymbol{\Sigma})}.$$

   (b) *M-step*: Compute the weighted MLEs for the parameters of each of the component Gaussians within each of the classes, using the weights from the E-step.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.