Linear Discriminant Analysis (LDA)

Yang Xiaozhou

March 18, 2020

Industrial Systems Engineering and Management, NUS

Table of contents

LDA

LDA and its applications

LDA is used as a tool for classification.

- Bankruptcy prediction: Edward Altman's 1968 model
- Face recognition: learnt features are called Fisher faces
- Biomedical studies: discriminate different stages of a disease
- and many more

It has shown some really good results:

- Top 3 classifiers for 11 of the 22 datasets studied in the STATLOG project[1]

[1](Michie et al. 1994)

Classification by discriminant analysis

Consider a generic classification problem:

- K groups: $G = 1, \ldots, K$, each with a density $f_k(\mathbf{x})$ on $\mathbb{R}^p$.
- A discriminant rule divides the space into K disjoint regions $\mathbb{R}_1, \ldots, \mathbb{R}_K$ and

$$\text{allocate } \mathbf{x} \text{ to } \Pi_j \text{ if } \mathbf{x} \in \mathbb{R}_j \,.$$

Classification by discriminant analysis

Consider a generic classification problem:

- K groups: $G = 1, \ldots, K$, each with a density $f_k(\mathbf{x})$ on $\mathbb{R}^p$.
- A discriminant rule divides the space into K disjoint regions $\mathbb{R}_1, \ldots, \mathbb{R}_K$ and

$$\text{allocate } \mathbf{x} \text{ to } \Pi_j \text{ if } \mathbf{x} \in \mathbb{R}_j \,.$$

- Maximum likelihood rule:

$$\text{allocate } \mathbf{x} \text{ to } \Pi_j \text{ if } j = \arg\max_i f_i(\mathbf{x}) \,,$$

- Bayesian rule with class priors $\pi_1, \ldots, \pi_K$:

$$\text{allocate } \mathbf{x} \text{ to } \Pi_j \text{ if } j = \arg\max_i \pi_i f_i(\mathbf{x}) \,.$$

Gaussian as class density

If we assume data comes from Gaussian distribution:
$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)}$

- Parameters are estimated using training data: $\hat{\pi}_k, \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k$.
- Looking at the log-likelihod:

$$\text{allocate } \mathbf{x} \text{ to } \Pi_j \text{ if } j = \arg\max_i \delta_i(\mathbf{x}) \,.$$

$\delta_i(\mathbf{x}) = \log f_i(\mathbf{x}) + \log \pi_i$ is called discriminant function.

Gaussian as class density

If we assume data comes from Gaussian distribution:
$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)}$

- Parameters are estimated using training data: $\hat{\pi}_k, \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k$.
- Looking at the log-likelihod:

$$\text{allocate } \mathbf{x} \text{ to } \Pi_j \text{ if } j = \arg\max_i \delta_i(\mathbf{x})\,.$$

$\delta_i(\mathbf{x}) = \log f_i(\mathbf{x}) + \log \pi_i$ is called discriminant function.

Assume equal covariance among K classes: LDA

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

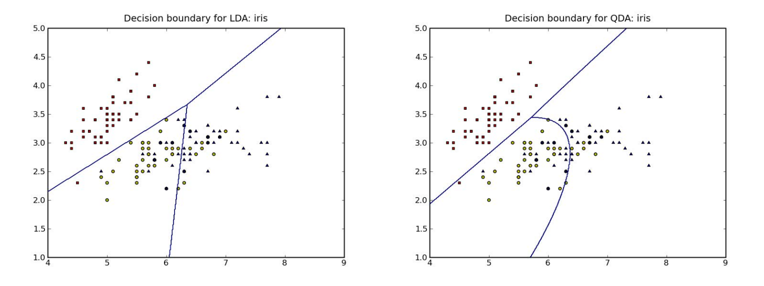
Without that assumption on class covariance: QDA.

Decision boundary: LDA vs QDA

- Between any pair of classes k and ℓ, the decision boundary is:

$$\{\mathbf{x} : \delta_k(\mathbf{x}) = \delta_\ell(\mathbf{x})\}$$

- LDA: linear boundary; QDA: quadratic boundary.



- Number of parameters to estimate rises quickly in QDA:
  - LDA: $(K-1)(p+1)$
  - QDA: $(K-1)\{p(p+3)/2+1\}$

Reduced-rank LDA

Reduced-rank LDA

Computation for LDA:

- Sphere the data:

$$\mathbf{x}^* \leftarrow \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{x}\,,$$

  where $\hat{\boldsymbol{\Sigma}} = \mathbf{UDU}^T$.

- Classify $\mathbf{x}$ to the closest centroid in the transformed space:

$$\delta_k(\mathbf{x}^*) = \mathbf{x}^{*T}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\mu}_k + \log\pi_k\,.$$

Reduced-rank LDA

Computation for LDA:

- Sphere the data:

$$\mathbf{x}^* \leftarrow \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{x}\,,$$

  where $\hat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

- Classify $\mathbf{x}$ to the closest centroid in the transformed space:

$$\delta_k(\mathbf{x}^*) = {\mathbf{x}^*}^T\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\mu}_k + \log\pi_k\,.$$

Inherent dimension reduction in LDA:

- K centroids lie in a subspace of dimension at most $(K-1)$:

$$H_{K-1} = \mu_1 \oplus \text{span}\,\{\mu_i - \mu_1, 2 \le i \le K\}$$

- Classification is done by distance comparison in H_{K-1}.
  - $p \to K-1$ dimension reduction assuming $p > K$.

Reduced-rank LDA

We can look for an even smaller subspace $H_L \subseteq H_{K-1}$:

- Rule: Class centroids of sphered data have maximum separation in this subspace in terms of variance.

Reduced-rank LDA

We can look for an even smaller subspace $H_L \subseteq H_{K-1}$:

- Rule: Class centroids of sphered data have maximum separation in this subspace in terms of variance.

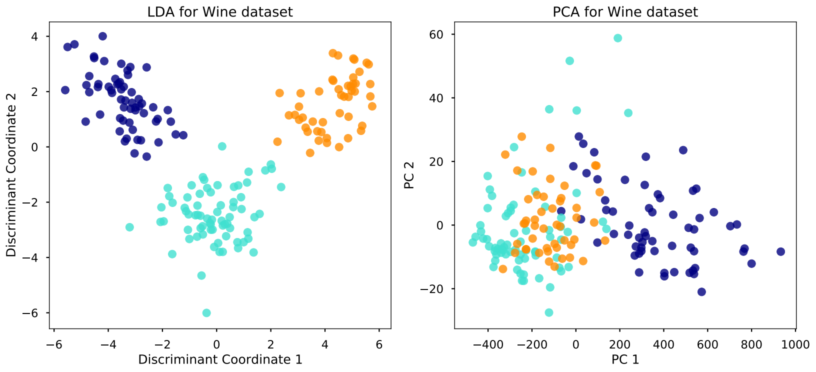
PCA on class centroids to find coordinates of H_L.

1. Find class mean and pooled var-cov: $\mathbf{M}, \mathbf{W}$.
2. Sphere the centroids: $\mathbf{M}^* = \mathbf{M}\mathbf{W}^{-\frac{1}{2}}$.
3. Obtain eigenvectors $(\mathbf{v}_\ell^*)$ in $\mathbf{V}^*$ of $\mathrm{cov}(\mathbf{M}^*) = \mathbf{V}^*\mathbf{D_B}\mathbf{V}^{*T}$.
4. Obtain new (discriminant) variables $Z_\ell = (\mathbf{W}^{-\frac{1}{2}}\mathbf{v}_\ell^*)^T X$, $\ell = 1, \ldots, L$.

Dimension reduction: $\mathbf{X}_{N\times p} \to \mathbf{Z}_{N\times L}$.

Reduced-rank LDA vs PCA

Wine dataset: 13 variables to distinguish three types of wines.



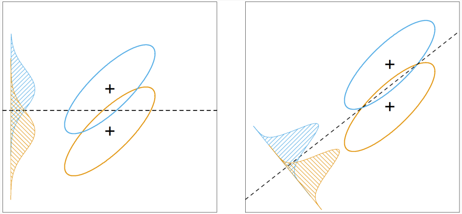
Fisher's LDA

Fisher's LDA

The previous rule is proposed by Fisher:

- Find a linear combination $Z = \mathbf{a}^T X$ that has maximum between-class variance relative to its within-class variance:

$$\mathbf{a} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}}.$$

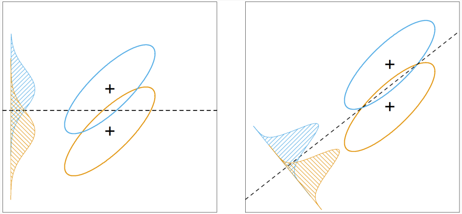


Fisher's LDA

The previous rule is proposed by Fisher:

- Find a linear combination $Z = \mathbf{a}^T X$ that has maximum between-class variance relative to its within-class variance:

$$\mathbf{a} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}} .$$

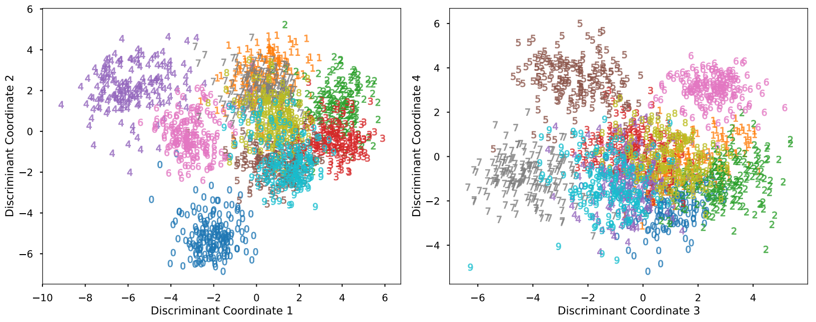


- The optimization is solved by a generalized eigenvalue problem: $\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda \mathbf{a}$.
- Eigenvectors ($\mathbf{a}_\ell$) of $\mathbf{W}^{-1}\mathbf{B}$ are the same as $(\mathbf{W}^{-\frac{1}{2}}\mathbf{v}_\ell^*)$. Fisher arrives at this without Gaussian assumption.

Fisher's LDA

Digit dataset: 64 variables to distinguish 10 written digits.

- Top 4 of Fisher's discriminant variables are shown.
- For example, coordinate 1 contrasts 4's and 2/3's.



Summary of LDA

Virtues of LDA:

1. Simple prototype classifier: simple to interpret.
2. Decision boundary is linear: simple to describe and implement.
3. Dimension reduction: provides informative low-dimensional view on data.

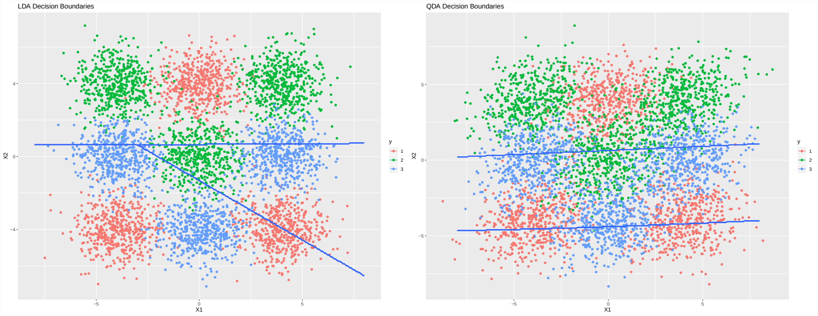
Shortcomings of LDA:

1. Linear decision boundaries may not adequately separate the classes. Support for more general boundaries is desired.
2. In high-dimensional setting, LDA uses too many parameters. Regularized version of LDA is desired.

Flexible Discriminant Analysis

Beyond linear boundaries: FDA

Flexible discriminant analysis (FDA) can tackle the first shortcoming.



Idea: Recast LDA as a regression problem, apply the same techniques generalizing linear regression.

LDA as a regression problem

We can recast LDA as a regression problem via optimal scoring.

Set up:

- Response G falls into one of K classes, $\mathcal{G} = \{1, \ldots, K\}$.
- X is the p-dimensional feature vector.

Suppose a scoring function:

$$\theta : \mathcal{G} \mapsto \mathbb{R}^1$$

such that scores are optimally predicted by regressing on X, e.g. a linear map $\eta(X) = X^T \beta$.

LDA as a regression problem

We can recast LDA as a regression problem via optimal scoring.

Set up:

- Response G falls into one of K classes, $\mathcal{G} = \{1, \ldots, K\}$.
- X is the p-dimensional feature vector.

Suppose a scoring function:

$$\theta : \mathcal{G} \mapsto \mathbb{R}^1$$

such that scores are optimally predicted by regressing on X, e.g. a linear map $\eta(X) = X^T \boldsymbol{\beta}$.

In general, select $L \leq K - 1$ such scoring functions and find the optimal $\{$score, linear map$\}$ pairs that minimize:

$$ASR = \frac{1}{N} \sum_{\ell=1}^{L} \left[\sum_{i=1}^{N} \left(\theta_\ell \left(g_i \right) - \boldsymbol{x}_i^T \boldsymbol{\beta}_\ell \right)^2 \right]$$

LDA via optimal scoring

Procedures of LDA via optimal scoring:

1. **Initialize.** Build response indicator matrix $\mathbf{Y}_{N\times K}$ where $\mathbf{Y}_{ij} = 1$ if ith samples comes from jth class, and 0 otherwise.
2. **Multivariate regression.** Regress $\mathbf{Y}$ on $\mathbf{X}$ using *ASR* to get $\mathbf{P}_X$ where $\hat{\mathbf{Y}} = \mathbf{P}_X\mathbf{Y}$, and regression coefficients $\mathbf{B}$.
3. **Optimal scores.** Obtain the L largest eigenvectors $\mathbf{\Theta}$ of $\mathbf{Y}^T\mathbf{P}_X\mathbf{Y}$.
4. **Update.** Update the coefficients: $\mathbf{B} \leftarrow \mathbf{B\Theta}$

LDA via optimal scoring

Procedures of LDA via optimal scoring:

1. **Initialize.** Build response indicator matrix $\mathbf{Y}_{N \times K}$ where $\mathbf{Y}_{ij} = 1$ if ith samples comes from jth class, and 0 otherwise.
2. **Multivariate regression.** Regress $\mathbf{Y}$ on $\mathbf{X}$ using ASR to get $\mathbf{P}_X$ where $\hat{\mathbf{Y}} = \mathbf{P}_X \mathbf{Y}$, and regression coefficients $\mathbf{B}$.
3. **Optimal scores.** Obtain the L largest eigenvectors $\mathbf{\Theta}$ of $\mathbf{Y}^T \mathbf{P}_X \mathbf{Y}$.
4. **Update.** Update the coefficients: $\mathbf{B} \leftarrow \mathbf{B\Theta}$

- The optimal linear map is: $\boldsymbol{\eta}(X) = \mathbf{B}^T X$.
- Columns of $\mathbf{B}$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_\ell$, are the same as $\mathbf{a}_\ell$'s in LDA up to a constant.

This equivalence with regression problem provides a starting point for generalizing LDA to a more flexible and nonparametric version.

From LDA to FDA

Extend LDA by generalizing the linear map:

$$\boldsymbol{\eta}(X) = \mathbf{B}^T X$$

to

$$\boldsymbol{\eta}(X) = \mathbf{B}^T h(X).$$

- Generalized additive fits
- Spline functions
- MARS models
- Projection pursuits
- Neural networks

The idea behind FDA: LDA in an enlarged space.

FDA via optimal scoring

The procedures of FDA is the same as LDA via optimal scoring with one change:

- Replace $\mathbf{P}_X$ with $\mathbf{S}_{h(X)}$, the nonparametric regression operator.

Initialize $\rightarrow$ **Multivariate regression** $\rightarrow$ **Optimal scores** $\rightarrow$ **Update.**

FDA via optimal scoring

The procedures of FDA is the same as LDA via optimal scoring with one change:

- Replace $\mathbf{P}_X$ with $\mathbf{S}_{h(X)}$, the nonparametric regression operator.

Initialize $\rightarrow$ **Multivariate regression** $\rightarrow$ **Optimal scores** $\rightarrow$ **Update.**

- Optimal fit: $\boldsymbol{\eta}(\mathbf{X})$.
- Fitted class centroids: $\overline{\boldsymbol{\eta}}^k = \sum_{g_i=k} \boldsymbol{\eta}\left(\mathbf{x}_i\right) / \boldsymbol{N}_k$.

A new observation X is classified to class k that minimizes:

$$\delta(\mathbf{x}, k) = \left\|\mathbf{D}\left(\boldsymbol{\eta}(\mathbf{x}) - \overline{\boldsymbol{\eta}}^j\right)\right\|^2$$

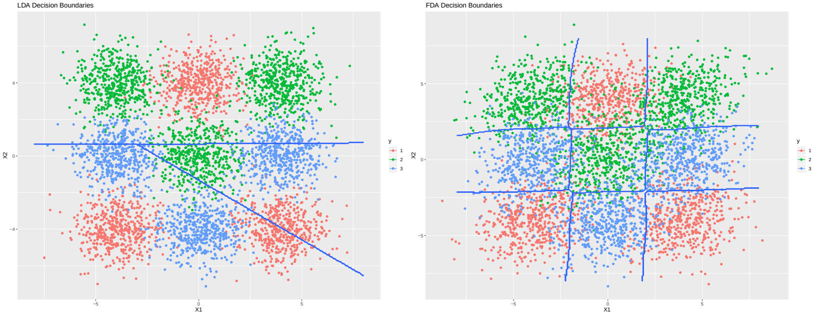
where $\mathbf{D}$ is the constant factor linking optimal fits and LDA coordinates.

LDA vs FDA

Data: three classes with mixture Gaussian densities.

FDA uses an additive model using smoothing splines of the form:

$$\alpha + \sum_{1}^{p} f_j\left(X_j\right)$$



References

1. Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
2. Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. Journal of the American statistical association, 89(428), 1255-1270.
3. Hastie, T., Tibshirani, R., & Buja, A. (1995). Flexible discriminant and mixture models.
4. Mardia, K. V., Kent, J. T., & Bibby, J. M. Multivariate analysis. 1979. Probability and mathematical statistics. Academic Press Inc.

Questions?

Contact: xiaozhou.yang@u.nus.edu