Lecture 4: Regularized and Flexible Discriminant Analysis

Rebecka Jörnsten, Mathematical Sciences

MSA220/MVE441 Statistical Learning for Big Data

28th March 2022

CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Extensions of Discriminant Analysis

Recap: The setting of Discriminant Analysis

Apply Bayes' law

$$p(i|\mathbf{x}) = \frac{p(\mathbf{x}|i)p(i)}{\sum_{j=1}^{K} p(\mathbf{x}|j)p(j)}$$

Instead of specifying $p(i|\mathbf{x})$ we can specify

$$p(\mathbf{x}|i) \quad \text{and} \quad p(i)$$

The main assumption of Discriminant Analysis (DA) is

$$p(\mathbf{x}|i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^p$ is the mean vector for class i and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{p \times p}$ the corresponding covariance matrix.

Finding the parameters of DA

- Notation: Write $p(i) = \pi_i$ and consider them as unknown parameters
- Given data $(i_l, \mathbf{x}_l)$ the likelihood maximization problem is

$$\arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}} \prod_{l=1}^{n} N(\mathbf{x}_l | \boldsymbol{\mu}_{i_l}, \boldsymbol{\Sigma}_{i_l}) \pi_{i_l} \quad \text{subject to} \quad \sum_{i=1}^{K} \pi_i = 1.$$

- Can be solved using a Lagrange multiplier (try it!) and leads to

$$\widehat{\pi}_i = \frac{n_i}{n}, \quad \text{with} \quad n_i = \sum_{l=1}^{n} \mathbb{1}(i_l = i)$$

$$\widehat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \sum_{i_l = i} x_l$$

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1} \sum_{i_l = i} (x_l - \widehat{\boldsymbol{\mu}}_i)(x_l - \widehat{\boldsymbol{\mu}}_i)^\mathsf{T}$$

Performing classification in DA

Bayes' rule implies the classification rule

$$c(\mathbf{x}) = \underset{1\le i\le K}{\arg\max}\, N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\pi_i$$

Note that since $\log$ is strictly increasing this is equivalent to

$$c(\mathbf{x}) = \underset{1\le i\le K}{\arg\max}\, \delta_i(\mathbf{x})$$

where

$$\begin{aligned}\delta_i(\mathbf{x}) &= \log N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \log \pi_i \\ &= \log \pi_i - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| \quad (+\,C)\end{aligned}$$

This is a quadratic function in $\mathbf{x}$.

Different levels of complexity

- This method is called **Quadratic Discriminant Analysis (QDA)**
- **Problem:** Many parameters that grow quickly with dimension
  - $K-1$ for all π_i
  - $p \cdot K$ for all $\boldsymbol{\mu}_i$
  - $p(p+1)/2 \cdot K$ for all $\boldsymbol{\Sigma}_i$ (most costly)
- **Solution:** Replace covariance matrices $\boldsymbol{\Sigma}_i$ by a pooled estimate

$$\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^{K} \widehat{\boldsymbol{\Sigma}}_i \frac{n_i - 1}{n - K} = \frac{1}{n-K} \sum_{i=1}^{K} \sum_{i_l = i} (x_l - \widehat{\boldsymbol{\mu}}_i)(x_l - \widehat{\boldsymbol{\mu}}_i)^\top$$

- **Simpler correlation and variance structure:** All classes are assumed to have the same correlation structure between features

Performing classification in the simplified case

As before, consider

$$c(\mathbf{x}) = \underset{1 \leq i \leq K}{\arg\max}\, \delta_i(\mathbf{x})$$

where

$$\delta_i(\mathbf{x}) = \log \pi_i + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i \quad (+\, C)$$

This is a linear function in $\mathbf{x}$. The method is therefore called **Linear Discriminant Analysis (LDA)**.

Even more simplifications

Other simplifications of the correlation structure are possible

- Ignore all correlations between features but allow different variances, i.e. $\mathbf{\Sigma}_i = \mathbf{\Lambda}_i$ for a diagonal matrix $\mathbf{\Lambda}_i$ (**Diagonal QDA** or **Naive Bayes' Classifier**)
- Ignore all correlations and make feature variances equal, i.e. $\mathbf{\Sigma}_i = \mathbf{\Lambda}$ for a diagonal matrix $\mathbf{\Lambda}$ (**Diagonal LDA**)
- Ignore correlations and variances, i.e. $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}_{p \times p}$ (**Nearest Centroids adjusted for class frequencies π_i**)

Regularized Discriminant Analysis

You can let the amount/type of regularization be controlled by tuning parameters

- $\widehat{\Sigma}_i(\lambda) = (1-\lambda)\widehat{\Sigma}_i + \lambda\widehat{\Sigma}$: shrinking individual covariance matrices toward the same shape.
- $\widehat{\Sigma}_i(\lambda,\gamma) = (1-\gamma)\widehat{\Sigma}_i(\lambda) + \gamma\frac{1}{d}Trace(\widehat{\Sigma}_i)I$: shrinks the regularized estimate toward a diagonal.
- Use CV to estimate the optimal tuningparameter values γ^*, λ^* - implemented in the caret package.
- With $\gamma,\lambda = 0,0$ this is QDA, with $\gamma,\lambda = 0,1$ it's LDA and $\gamma,\lambda = 1,1$ is nearest centroid, $\gamma,\lambda = 1,0$ Naive Bayes.

More flexible DA methods

What if the model assumption for DA is too simple?

- Mixture Discriminant Analysis where each class is modeled with several Gaussian distributions, i.e.

$$p(\mathbf{x}|i) \sim \sum_{c=1}^{C_i} \pi_c N(\mathbf{x}|\boldsymbol{\mu}_{i,c}, \boldsymbol{\Sigma})$$

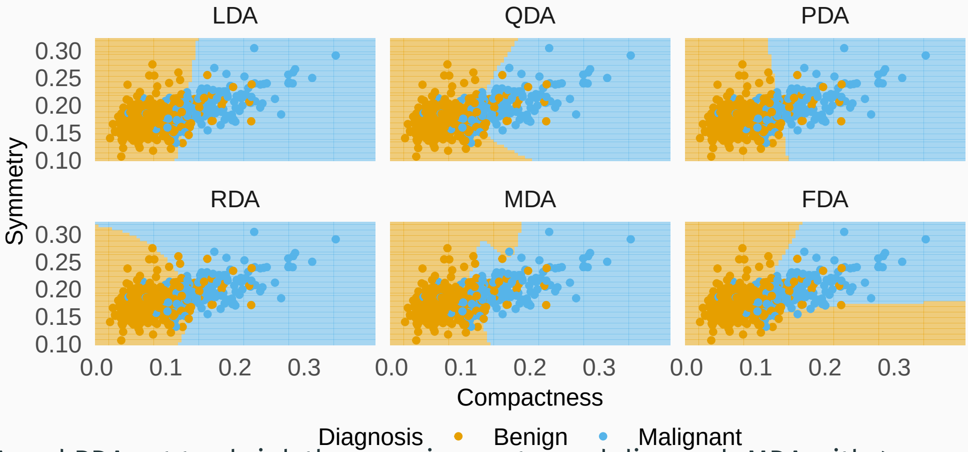
- Use a simple, common shape for all the class components
- You can use a different number of components for each class. This could be very useful in practice when some classes are more complex than others
- In order to fit the class-mixtures you use an iterative algorithm called *Expectation-Maximization* - we will revisit this in upcoming lecture when we switch focus to *clustering/unsupervised learning*

More flexible DA methods

Another way to increase flexibility to make the classification boundaries depend on more transformations/richer representations of x

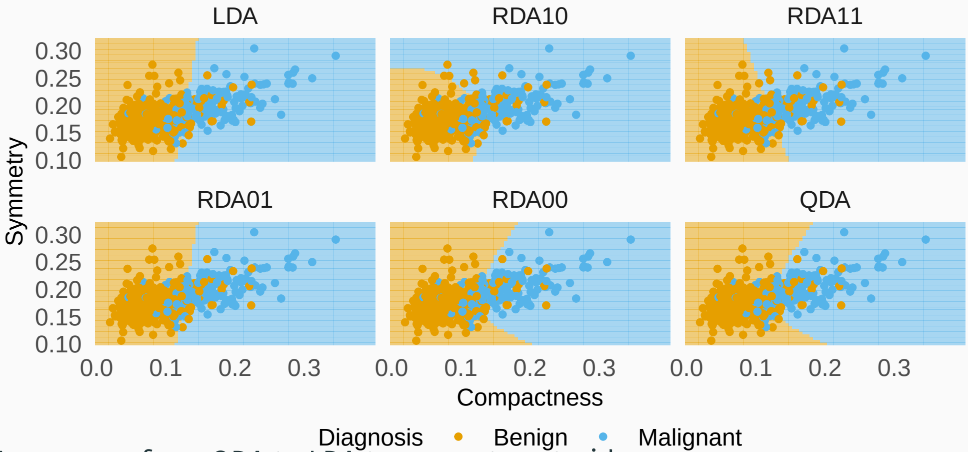
- Flexible Discrimimant Analysis is method based on LDA but extended through regression modeling and optimal scoring.
- The scores, $\theta_k, k = 1, \cdots, L <= K - 1$, are maps from a regression prediction $x^T\beta_i$ to classes.
- LDA = regression fits + nearest centroid classification in the fitted space.
- Now you extend this to more general forms of regression: splines, polynomial OR, if you want to regularize instead, ridge regression. Lots of possibilities.

Discriminant Analysis Methods



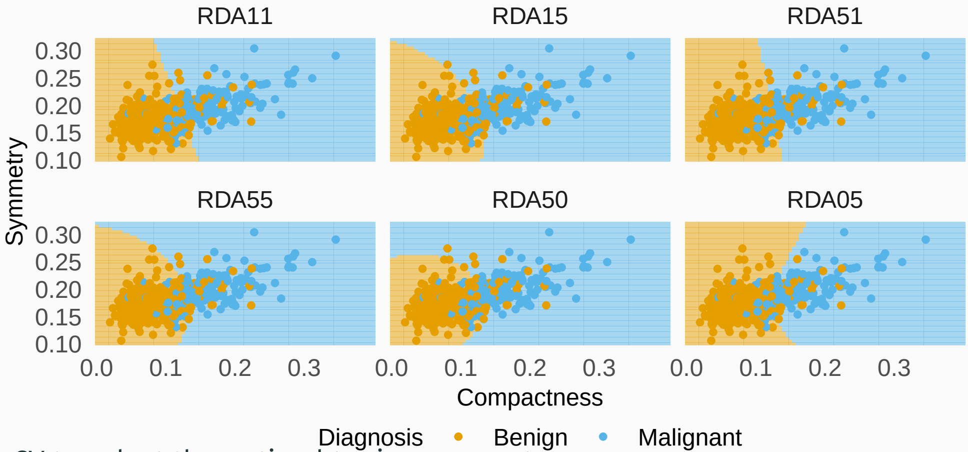
RDA and PDA set to shrink the covariances toward diagonal. MDA with 4 components and FDA with 3-degree splines.

Regularized Discriminant Analysis



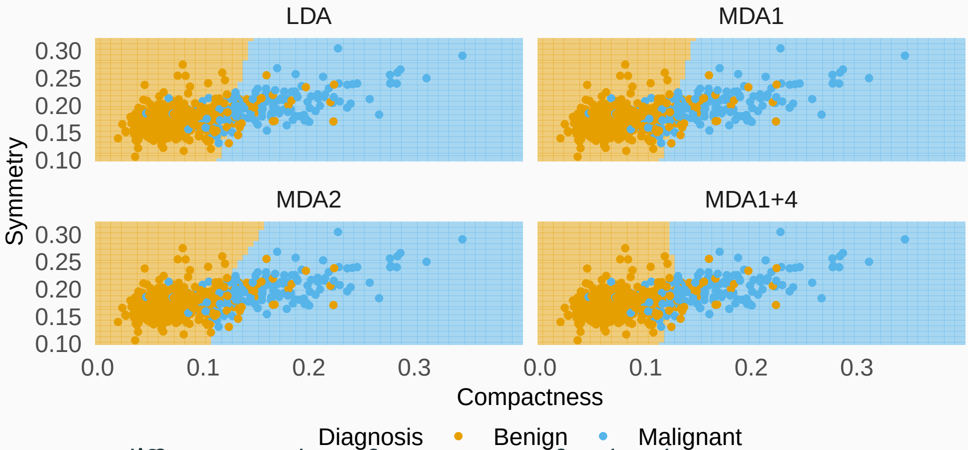
RDA can span from QDA to LDA to nearest centroid.

Regularized Discriminant Analysis



Use CV to select the optimal tuning parameters.

Mixture Discriminant Analysis



You can use different number of components for the classes.

Take-home message

- Discriminant Analysis can be extended to handle very complex decision boundaries.
  - MDA
  - FDA - goes through regression which opens up for use of lots if regression type methods for creating both flexible and interpretable classification (more later on sparse DA)
- There is a range of assumptions in DA about the correlation structure in feature space $\rightarrow$ trade-off between numerical stability and flexibility
- High-dimensional case: simplify or regularize
- Let the data tell you which modeling assumptions provide the best generalization properties (CV)