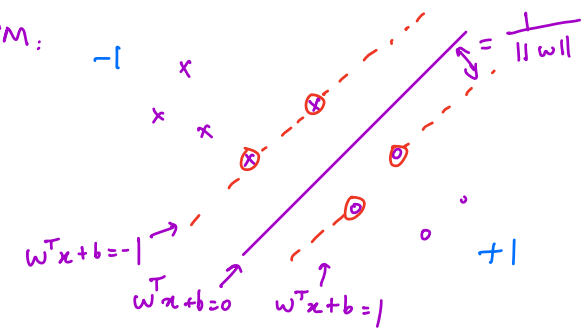
SVM:



Geometric margin

$\underline{y_i (w^T x_i + b)}$ functional margin

(Primary Problem)

$$\begin{cases} \underset{w}{\text{argmax}} \ \frac{1}{\|w\|} = \underset{w}{\text{argmin}} \ \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad \min \left(y_i (w^T x_i + b) \right) = 1 \end{cases}$$

(dual problem)

$$\begin{cases} L(w, b, \lambda) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \lambda_i \left(1 - y_i (w^T x_i + b) \right) \\ \max_{\lambda} \min_{w,b} L(w, b, \lambda) \\ \text{s.t.} \quad \lambda_i \geq 0, \ i = 1, 2, \dots n \end{cases}$$

Optimal Solution:

$$w^* = \sum_{i=1}^{n} \lambda_i^* y_i x_i$$

$$b^* = y_k - \sum_{i=1}^{n} \lambda_i^* y_i x_i^T x_k \quad k \text{ is the index for any support vector.}$$

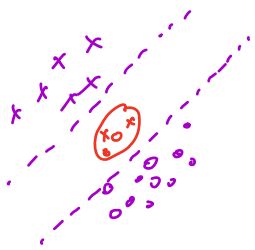
$\therefore$ Complementary slackness

$$\lambda_i \left(1 - y_i (w^T x_i + b) \right) = 0 \quad \text{and} \quad \lambda_i \geq 0$$

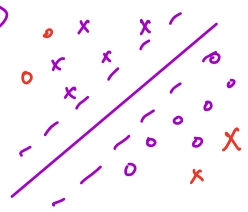
$\therefore$ for any non-Support vector, $\lambda_i = 0$

$\therefore$ $w^* = \sum_{i \in SV} \lambda_i y_i x_i \qquad b^* = y_k - \sum_{i \in SV} \lambda_i y_i x_i^T x_k$

Soft margin:

① 

will change hyperplane, ignore

② 

there's no hard-margin

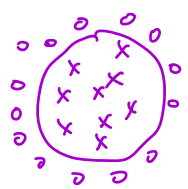
$$(P)\begin{cases} \min \frac{1}{2}\|w\|^2 + C\underbrace{\sum_{i=1}^{n} \max\{0,\ 1-y_i(w^T x+b)\}}_{\text{penalty}} \\ y_i(w^T x_i + b) \geq 1 \end{cases}$$

Why we can apply kernel?

$$f(x) = \text{Sign}\left(\sum_{i=1}^{n} \lambda_i^* y_i \langle x_i, x\rangle + b^*\right) = \text{Sign}\left(w^{*T} x + b^*\right)$$

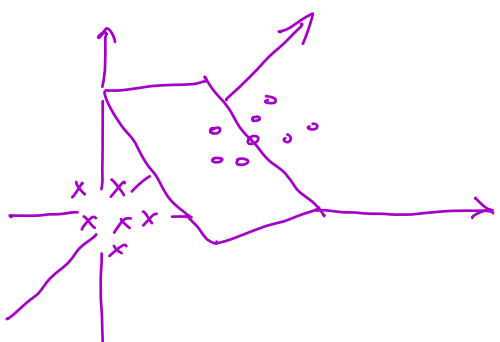
↑ inner product between samples

non-linear problem:



Linear fails

projection to 3 dimension



linearly separated

$$x_i \longrightarrow \varphi(x_i)$$

Kernel function

Kernel function is a function of form–

$$K(x, y) = (1 + \sum_{j=1}^{p} x_{ij} y_{ij})^d$$

, where d is the degree of polynomial.

Now the type of Kernel function we are going to use here is a **Radial kernel**.It is of form-

$$K(x, y) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - y_{ij})^2)$$

, and γ here is a tuning parameter which accounts for the smoothness of the decision boundary and controls the variance of the model.

If γ is very large then we get quiet fluctuating and wiggly decision boundaries which accounts for high variance and overfitting.
If γ is small, the decision line or boundary is smoother and has low variance.

LDA: X: normal distributed
Common Σ

Bayesian Thm: $$\Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$$

dist of x: $$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k)}$$

log ratio:

$$\begin{aligned}
&\log \frac{\Pr(G = k \mid X = x)}{\Pr(G = \ell \mid X = x)} \\
=&\log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell} \\
=&\log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T\mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell) + x^T\mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell)
\end{aligned}$$

linear in x

Classification function: $\underset{k}{\operatorname{argmax}}\ \delta_k(x) = x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + \log \pi_k$

linear in x

To find the covariance matrix, we simply compute

$$\hat{\Sigma} = \sum_{k=1}^{K} \frac{1}{N-K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T .$$

The means of the classes, which are also called centroids, are defined by

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{g_i=k} x_i .$$

The priors π_k are set to the prevalence ratio of the class-specific observations:

$$\hat{\pi}_k = \frac{N_k}{N} .$$

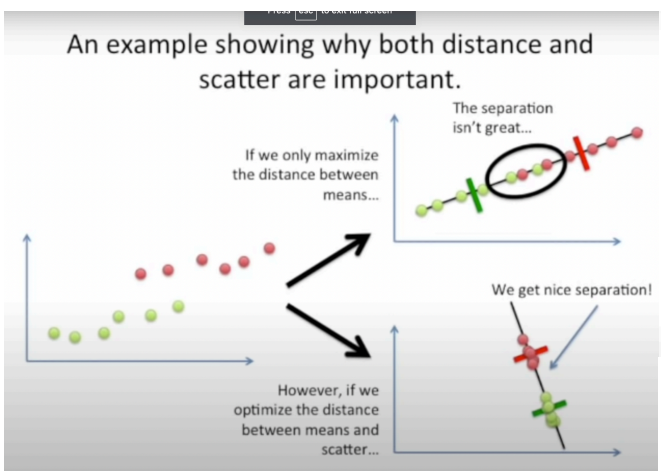
Fisher's view

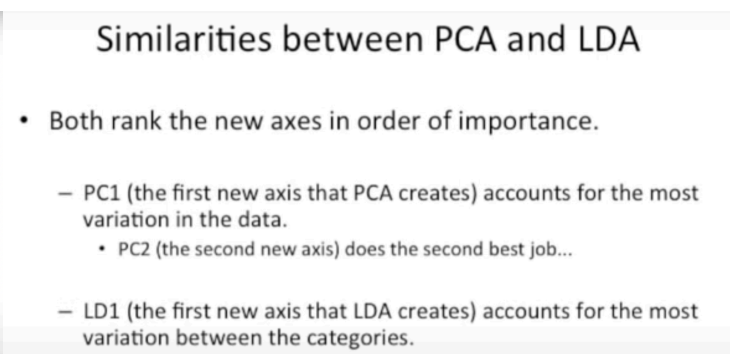
According to Fisher, LDA can be understood as a dimensionality reduction technique where each successive transformation is orthogonal and maximizes the between-class variance relative to the within-class variance. This procedure transforms the feature space to an affine space with $K - 1$ dimensions. After sphering the input data, new points can be classified by determining the closest centroid in the affine space under consideration of the class priors.

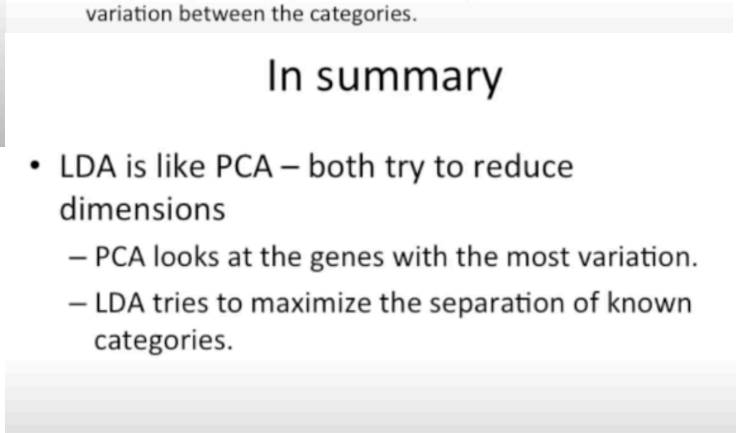
Properties of LDA

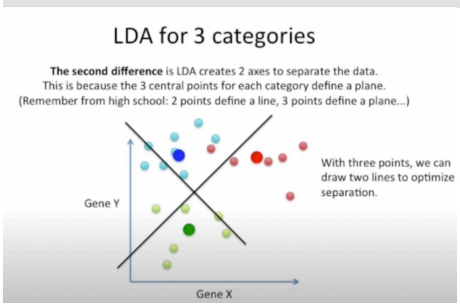
LDA has the following properties:

- LDA assumes that the data are Gaussian. More specifically, it assumes that all classes share the same covariance matrix.
- LDA finds linear decision boundaries in a $K - 1$ dimensional subspace. As such, it is not suited if there are higher-order interactions between the independent variables.
- LDA is well-suited for multi-class problems but should be used with care when the class distribution is imbalanced because the priors are estimated from the observed counts. Thus, observations will rarely be classified to infrequent classes.
- Similarly to PCA, LDA can be used as a dimensionality reduction technique. Note that the transformation of LDA is inherently different to PCA because LDA is a supervised method that considers the outcomes.









QDA: different Σ_k for each group

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

Regularized discriminant analysis

RDA is a compromise between LDA and QDA as it shrinks Σ_k to a pooled variance Σ by defining

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

and replacing $\hat{\Sigma}_k$ with $\hat{\Sigma}_k(\alpha)$ in the discriminant functions. Here, $\alpha \in [0, 1]$ is a tuning parameter determining whether the covariances should be estimated independently ($\alpha = 1$) or should be pooled ($\alpha = 0$).

Additionally, $\hat{\Sigma}$ can also be shrunk toward the scalar covariance by requiring

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 I$$

where $\gamma = 1$ leads to the pooled covariance and $\gamma = 0$ leads to the scalar covariance. Replacing $\hat{\Sigma}_k$ by $\hat{\Sigma}(\alpha, \gamma)$ leads to a more general notion of covariance.

Since RDA is a regularization technique, it is particularly useful when there are many features that are potentially correlated. Let us now evaluate RDA on the phoneme data set.

PENALIZED DISCRIMINANT ANALYSIS

By Trevor Hastie, Andreas Buja and Robert Tibshirani[1]

Stanford University, AT & T Bell Laboratories and University of Toronto

Fisher's linear discriminant analysis (LDA) is a popular data-analytic tool for studying the relationship between a set of predictors and a categorical response. In this paper we describe a penalized version of LDA. It is designed for situations in which there are many highly correlated predictors, such as those obtained by discretizing a function, or the grey-scale values of the pixels in a series of images. In cases such as these it is natural, efficient and sometimes essential to impose a spatial smoothness constraint on the coefficients, both for improved prediction performance and interpretability. We cast the classification problem into a regression framework via optimal scoring. Using this, our proposal facilitates the use of any penalized regression technique in the classification setting. The technique is illustrated with examples in speech recognition and handwritten character recognition.

fused lasso:

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i=1}^{p-1} \left|\boldsymbol{\beta}_{i+1} - \boldsymbol{\beta}_i\right|,$$

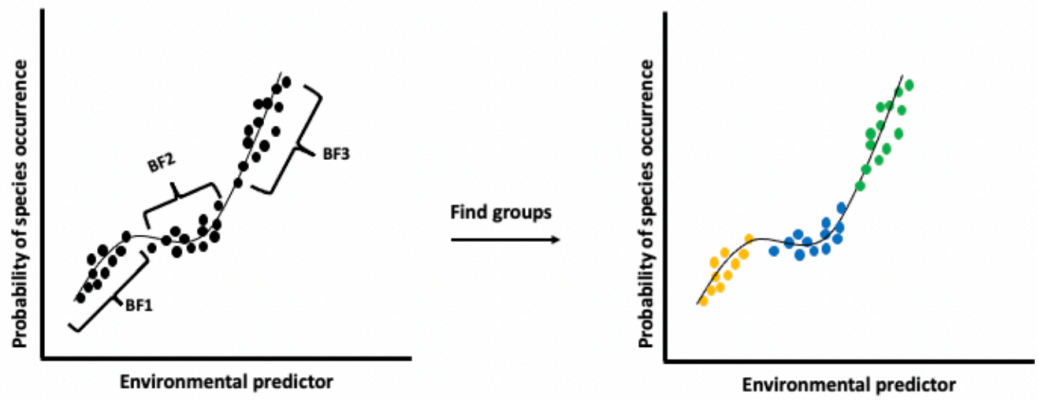
We propose the 'fused lasso', a generalization that is designed for problems with features that can be ordered in some meaningful way. The fused lasso penalizes the L_1 -norm of both the coefficients and their successive differences. Thus it encourages sparsity of the coefficients and also sparsity of their differences—i.e. local constancy of the coefficient profile.

That is, it's called that because adjacent parameters may be set equal -- i.e. "fused" (somewhat akin to aligning broken bones, which ultimately fuse together).

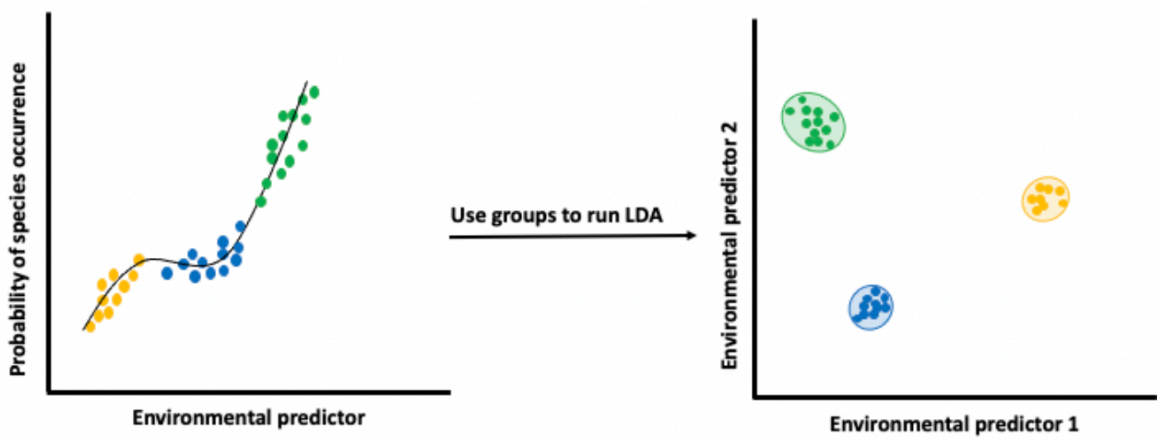
Introduction

Flexible discriminant analysis (FDA) is a general methodology which aims at providing tools for multigroup non-linear classification. It is a classification model based on a mixture of non-parametric regression models e.g. MARS and linear discriminant analysis.

The first step of an FDA is a non-parametric regression, which uses optimal scoring to transform the response variable so that the data are in a better form for linear separation. It builds multiple regression models, so called basis functions (BF), across the range of predictor values. In this procedure, the range of predictor values is partitioned in several groups/ categories.



In the second step of an FDA the groups identified in the first step are used to run a linear discriminant analysis. Linear discriminant analysis focuses on maximising the seperatibility among groups, while minimising the variance within each group.

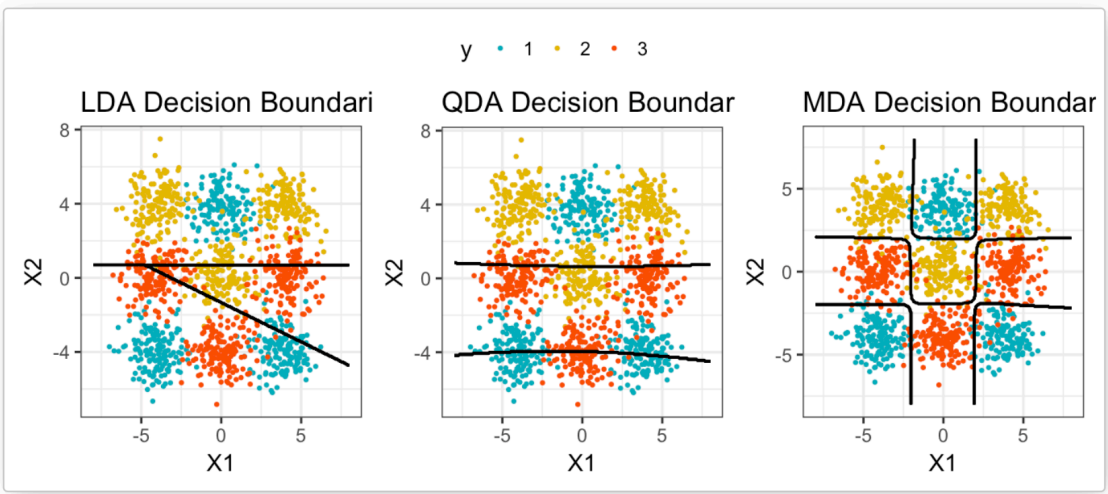


The first axis that LDA creates (environmental predictor 1) accounts for the most variation between the groups. The second axis (environmental predictor 2) accounts for the second most variation between the groups. This continues until every predictor is ranked. For simplicity reason only a 2-dimensional graph with 2 predictors (axis) is displayed at one time.

Advantages

- Works well with a large number of predictor variables
- Automatically detects interactions between variables
- It is an efficient and fast algorithm, despite its complexity
- Robust to outliers

MDA might outperform LDA and QDA is some situations, as illustrated below. In this example data, we have 3 main groups of individuals, each having 3 no adjacent subgroups. The solid black lines on the plot represent the decision boundaries of LDA, QDA and MDA. It can be seen that the MDA classifier have identified correctly the subclasses compared to LDA and QDA, which were not good at all in modeling this data.



Mixture Discriminant Analysis

- A single Gaussian to model a class, as in LDA, is too restricted.
- Extend to a mixture of Gaussians. For class k, the within-class density is:

$$f_k(x) = \sum_{r=1}^{R_k} \pi_{kr} \phi(x|\mu_{kr}, \Sigma)$$

- A common covariance matrix is still assumed.