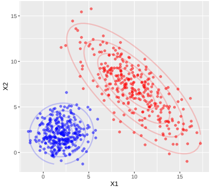
Introduction to Machine Learning

Classification: Discriminant Analysis



Learning goals

- Understand the ideas of linear and quadratic discriminant analysis
- Understand how parameteres are estimated for LDA and QDA
- Understand how decision boundaries are computed for LDA and QDA

LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA follows a generative approach

$$\pi_k(\mathbf{x}) = \mathbb{P}(y = k \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|y = k)\mathbb{P}(y = k)}{\mathbb{P}(\mathbf{x})} = \frac{p(\mathbf{x}|y = k)\pi_k}{\sum\limits_{j=1}^{g} p(\mathbf{x}|y = j)\pi_j},$$

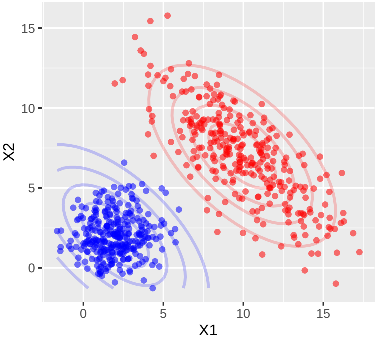
where we now have to pick a distributional form for $p(\mathbf{x}|y = k)$.

LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA assumes that each class density is modeled as a *multivariate Gaussian*:

$$p(\mathbf{x}|y=k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right)$$

with equal covariance, i. e. $\Sigma_k = \Sigma \quad \forall k$.



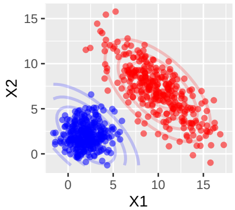
LINEAR DISCRIMINANT ANALYSIS (LDA)

Parameters $\boldsymbol{\theta}$ are estimated in a straightforward manner by estimating

$$\hat{\pi}_k = \frac{n_k}{n}, \text{ where } n_k \text{ is the number of class-}k\text{ observations}$$

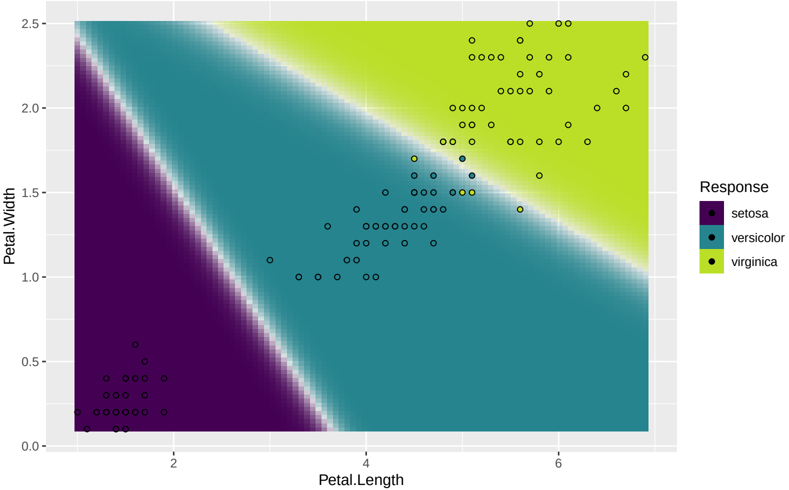
$$\hat{\boldsymbol{\mu}}_{\boldsymbol{k}} = \frac{1}{n_k} \sum_{i: y^{(i)}=k} \mathbf{x}^{(i)}$$

$$\hat{\Sigma} = \frac{1}{n-g} \sum_{k=1}^{g} \sum_{i: y^{(i)}=k} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\boldsymbol{k}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\boldsymbol{k}})^T$$



LDA AS LINEAR CLASSIFIER

Because of the equal covariance structure of all class-specific Gaussian, the decision boundaries of LDA are linear.



LDA AS LINEAR CLASSIFIER

We can formally show that LDA is a linear classifier, by showing that the posterior probabilities can be written as linear scoring functions - up to any isotonic / rank-preserving transformation.

$$\pi_k(\mathbf{x}) = \frac{\pi_k \cdot p(\mathbf{x}|y = k)}{p(\mathbf{x})} = \frac{\pi_k \cdot p(\mathbf{x}|y = k)}{\sum\limits_{j=1}^{g} \pi_j \cdot p(\mathbf{x}|y = j)}$$

As the denominator is the same for all classes we only need to consider

$$\pi_k \cdot p(\mathbf{x}|y = k)$$

and show that this can be written as a linear function of **x**.

LDA AS LINEAR CLASSIFIER

$$\begin{aligned}
& \pi_k \cdot p(\mathbf{x}|y=k) \\
\propto \quad & \pi_k \exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{k}}^T\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k}} + \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k}}\right) \\
= \quad & \exp\left(\log\pi_k - \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{k}}^T\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k}} + \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k}}\right)\exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right) \\
= \quad & \exp\left(\boldsymbol{\theta}_{0k} + \mathbf{x}^T\boldsymbol{\theta}_k\right)\exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right) \\
\propto \quad & \exp\left(\boldsymbol{\theta}_{0k} + \mathbf{x}^T\boldsymbol{\theta}_k\right)
\end{aligned}$$

by defining $\boldsymbol{\theta}_{0k} := \log\pi_k - \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{k}}^T\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k}}$ and $\boldsymbol{\theta}_k := \Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k}}$.

We have again left out all constants which are the same for all classes k, so the normalizing constant of our Gaussians and $\exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right)$.

By finally taking the log, we can write our transformed scores as linear:

$$f_k(\mathbf{x}) = \boldsymbol{\theta}_{0k} + \mathbf{x}^T\boldsymbol{\theta}_k$$

QUADRATIC DISCRIMINANT ANALYSIS (QDA)

QDA is a direct generalization of LDA, where the class densities are now Gaussians with unequal covariances Σ_k.

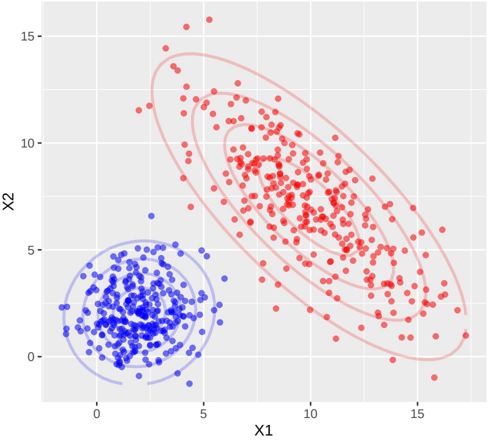
$$p(\mathbf{x}|y=k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right)$$

Parameters are estimated in a straightforward manner by:

$$\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n}, \text{ where } n_k \text{ is the number of class-}k\text{ observations} \\
\hat{\boldsymbol{\mu_k}} &= \frac{1}{n_k}\sum_{i:y^{(i)}=k} \mathbf{x}^{(i)} \\
\hat{\Sigma}_k &= \frac{1}{n_k-1}\sum_{i:y^{(i)}=k} (\mathbf{x}^{(i)}-\hat{\boldsymbol{\mu_k}})(\mathbf{x}^{(i)}-\hat{\boldsymbol{\mu_k}})^T
\end{aligned}$$

QUADRATIC DISCRIMINANT ANALYSIS (QDA)

- Covariance matrices can differ over classes.
- Yields better data fit but also requires estimation of more parameters.



QUADRATIC DISCRIMINANT ANALYSIS (QDA)

$$\begin{aligned}\pi_k(\mathbf{x}) &\propto \pi_k \cdot p(\mathbf{x}|y=k) \\ &\propto \pi_k|\Sigma_k|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^T\Sigma_k^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{k}}^T\Sigma_k^{-1}\boldsymbol{\mu}_{\boldsymbol{k}} + \mathbf{x}^T\Sigma_k^{-1}\boldsymbol{\mu}_{\boldsymbol{k}})\end{aligned}$$

Taking the log of the above, we can define a discriminant function that is quadratic in x.

$$\log \pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{k}}^T\Sigma_k^{-1}\boldsymbol{\mu}_{\boldsymbol{k}} + \mathbf{x}^T\Sigma_k^{-1}\boldsymbol{\mu}_{\boldsymbol{k}} - \frac{1}{2}\mathbf{x}^T\Sigma_k^{-1}\mathbf{x}$$

QUADRATIC DISCRIMINANT ANALYSIS (QDA)