Linear classification algorithms

There are several different approaches to linear classification.

The chapter *Linear Methods for Classification* in *The Elements of Statistical Learning* by Hastie et al. discusses three groups of algorithms: (1) logistic regression, (2) linear discrimininant analysis, and (3) separating hyperplanes (perceptron and linear support vector machine).

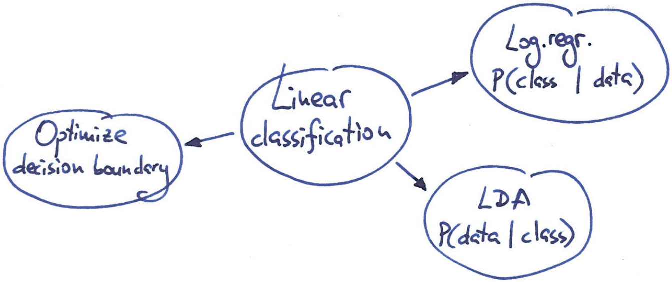
In logistic regression, we modeled $P(\text{class} \mid \text{data})$.

In discriminant analysis, we will model $P(\text{data} \mid \text{class})$.

One can also directly optimize a linear decision boundary, without any probabilistic model. We will not cover such methods in this course.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Linear classification algorithms



$P(\text{class} \mid \mathbf{x})$ vs. $P(\mathbf{x} \mid \text{class})$

What we want, is $P(\text{class} \mid \mathbf{x})$. This is what logistic regression directly estimates.

Alternatively, we can assume some model for $P(\mathbf{x} \mid \text{class})$ and some prior $P(\text{class})$. Then, using Bayes rule, we can get $P(\text{class} \mid \mathbf{x})$:

$$P(\text{class} = k \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \text{class} = k)P(\text{class} = k)}{\sum_i P(\mathbf{x} \mid \text{class} = i)P(\text{class} = i)}$$

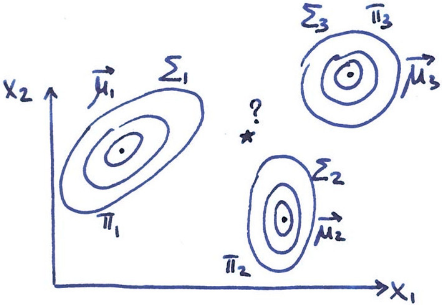
$$P(\text{class} = k \mid \mathbf{x}) \sim f_k(\mathbf{x})\pi_k$$

Recall that linear regression models $P(y \mid \mathbf{x})$. A question to think about: would it make sense to assume a model for $P(\mathbf{x} \mid y)$ together with a prior $P(y)$?..

Gaussian densities

Suppose that $f_k(\mathbf{x})$ are all multivariate Gaussians:

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \det \boldsymbol{\Sigma}_k}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\Big].$$



Quadratic discriminant analysis (QDA)

Let us consider a binary classification problem with $\pi_1 = \pi_2 = \frac{1}{2}$ and

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \det \boldsymbol{\Sigma}_k}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\Big].$$

Then the *decision boundary* is given by $P(\text{class } 1 \mid \mathbf{x}) = P(\text{class } 2 \mid \mathbf{x})$:

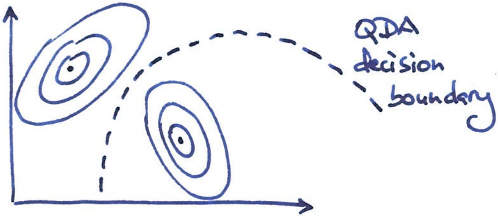
$$\cancel{-\frac{p}{2}\log(2\pi)} - \frac{1}{2}\log\det\boldsymbol{\Sigma}_1 - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) =$$
$$\cancel{-\frac{p}{2}\log(2\pi)} - \frac{1}{2}\log\det\boldsymbol{\Sigma}_2 - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2).$$

This is called *quadratic discriminant analysis* (QDA).

Quadratic discriminant analysis (QDA)

QDA decision boundary:

$$\log\det \boldsymbol{\Sigma}_1 + (\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) = \log\det \boldsymbol{\Sigma}_2 + (\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2).$$



Linear discriminant analysis (LDA)

QDA decision boundary:

$$\log\det\boldsymbol{\Sigma}_1 + (\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) = \log\det\boldsymbol{\Sigma}_2 + (\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2).$$

Let us assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Then:

$$(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) = (\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$$

$$2\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = \underbrace{\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2}_{\text{const}}$$

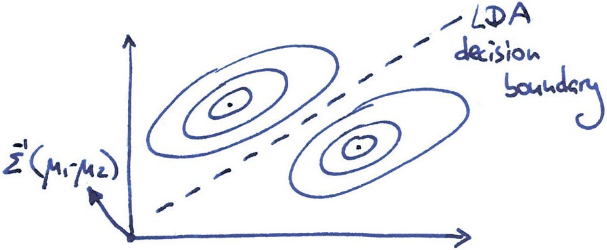
$$\boxed{\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = \text{const}}$$

This is linear projection of $\mathbf{x}$ onto the $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ direction.

Linear discriminant analysis (LDA)

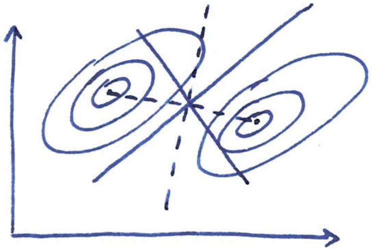
LDA decision boundary:

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2).$$



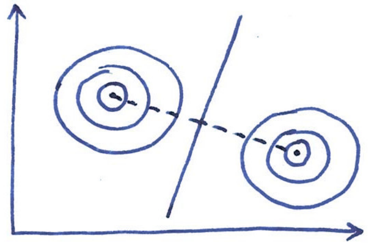
The role of $\mathbf{\Sigma}^{-1}$ in LDA

Why does LDA use projection on $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and not simply on $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$?



Nearest centroid classifier

Under additional assumption that the covariance matrix is spherical, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, LDA reduces to the *nearest centroid* classifier:



But note that the nearest centroid classifier is non-probabilistic, whereas 'spherical LDA' makes probabilistic predictions.

Estimating Gaussian parameters

For QDA / LDA / nearest centroid, we need to know π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$. They can be estimated from the training data using standard formulas:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \mathbf{x}_i,$$

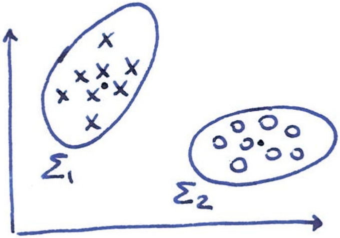
$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k - 1} \sum_{i \in \mathcal{C}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top,$$

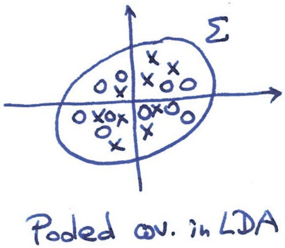
$$\hat{\pi}_k = \frac{n_k}{n}.$$

For LDA, one uses the *pooled* covariance estimator:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top.$$

Estimating Gaussian parameters





$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top.$$

Overfitting and ridge regularization in LDA

The $\mathbf{\Sigma}^{-1}$ factor in $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ may cause high-variance overfitting problems unless $n \gg p$.

Recall that in linear regression, ridge regularization replaces $(\mathbf{X}^\top \mathbf{X})^{-1}$ with $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$. Similarly, we can construct regularized LDA by replacing $\mathbf{\Sigma}^{-1}$ with $(\mathbf{\Sigma} + \lambda \mathbf{I})^{-1}$.

This can be alternatively written as $\big((1-\lambda)\mathbf{\Sigma} + \lambda \mathbf{I}\big)^{-1}$: interpolating between LDA and nearest centroid classifier.

Similar interpolation can be used between QDA and LDA: $\big((1-\lambda)\mathbf{\Sigma}_k + \lambda \mathbf{\Sigma}\big)^{-1}$.

LDA/QDA flavours

Other choices are possible by constraining $\mathbf{\Sigma}_k$ in various ways:

Covariance matrices	Separate	Shared
Full	QDA	LDA
Diagonal	Naive Bayes	Diagonal LDA
Spherical	'Spherical QDA'	Nearest centroid

Exercise: for a binary classification in dimensionality p, how many parameters does each of these covariance models use?

Note: one can interpolate between any of them using λ.

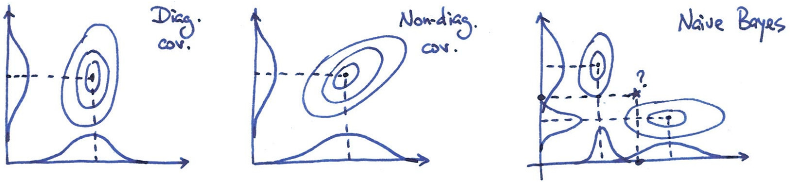
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Diagonal QDA aka Gaussian Naive Bayes

Here correlations between features are ignored and each feature is treated independently:

$$f_k(\mathbf{x}) = \prod_{j=1}^{p} f_{kj}(x_j).$$

Exercise: if $f_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with diagonal $\boldsymbol{\Sigma}_k = \mathrm{diag}\{\sigma_{k1}^2, \sigma_{k2}^2, \ldots, \sigma_{kp}^2\}$, then $f_{kj} \sim \mathcal{N}(\mu_{kj}, \sigma_{kj}^2)$.



Fisher's discriminant analysis*

Fisher (1936) derived LDA via a different route. He posed the following problem: find a linear projection that would maximize the ratio of the between-class 'spread' to the within-class 'spread'.

Let the means and the variances of each class after projection be m_1 and m_2, and $s_1^2/(n_1 - 1)$ and $s_2^2/(n_2 - 1)$. Then the ratio can be written as

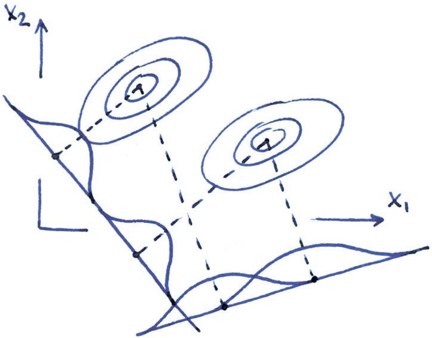
$$\mathcal{R} = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} \sim \frac{\big(\mathbf{w}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big)^2}{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}.$$

Define $\mathbf{v} = \boldsymbol{\Sigma}^{1/2}\mathbf{w}$. Then

$$\mathcal{R} = \frac{\big(\mathbf{v}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big)^2}{\mathbf{v}^\top \mathbf{v}},$$

so $\hat{\mathbf{v}} \sim \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and hence $\hat{\mathbf{w}} \sim \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Fisher's discriminant analysis



LDA vs. logistic regression

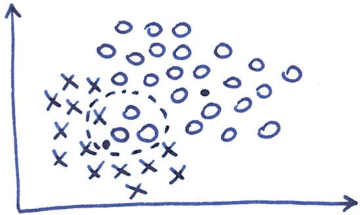
LDA and logistic regression are both very popular. When appropriately regularized, they often perform similarly well in practice.

If the data are truly Gaussian, LDA is optimal.

But logistic regression can perform better when the data are non-Gaussian (and can be more robust to outliers).

Nearest centroid vs. k nearest neighbours

Instead of using the nearest centroid for classification, one can use the majority vote among the k nearest neighbours.



The value of k controls the bias–variance tradeoff: low bias with $k = 1$, low variance with $k \gg 1$.

This is a *non-parametric* method. The entire training set needs to be available at test time. But it can be given a probabilistic interpretation as a non-parametric estimate of $p(\mathbf{x} \mid \text{class} = i) \sim c_i/n_i$.