Gaussian Discriminant Analysis

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)
Université d'Antananarivo

Discriminative vs Generative classifiers

Discriminative classifier

A discriminative classifier directly models the posterior $p(y|\boldsymbol{x})$. It can only be used to discriminate between classes.

Generative classifier

In contrast, a generative classifier models the class conditional density $p(\boldsymbol{x}|y)$. It can be used to generate examples $\boldsymbol{x}$ from each class y.
We can obtain the posterior using Bayes rule

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})}$$

We do not even need to calculate the denominator to make predictions:

$$\hat{y} = \arg\max_y p(\boldsymbol{x}|y)p(y)$$

Model

In Gaussian discriminant analysis, the class confitional densities are multivariate Gaussians:

$$p(\boldsymbol{x}|y=c;\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$$

The posterior has the following form:

$$p(y=c|\boldsymbol{x};\boldsymbol{\theta}) \propto \pi_c\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$$

where $\pi_c = p(y=c)$ is the prior probability of label c

Quadratic decison boundaries

The log posterior over class label is

$$\log p(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = \log \pi_c - \frac{1}{2}\log|\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c) + cst$$

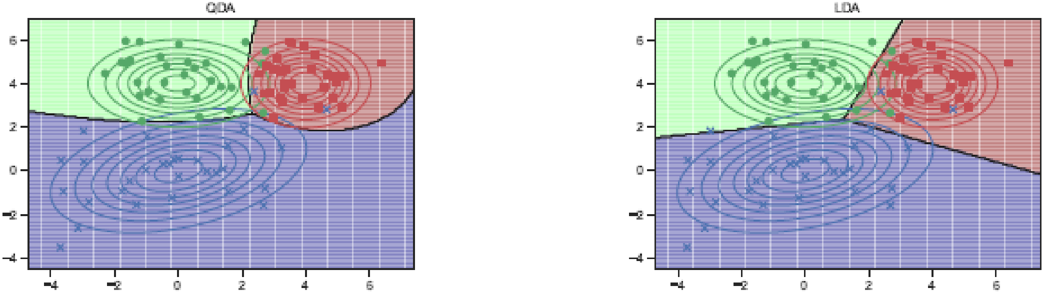
This is the discriminant function which is a quadratic function of $\boldsymbol{x}$. The model is called **quadratic discriminant analysis** or **QDA**.

Linear decison boundaries

If the covariance matrices are shared across classes, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$, the log posterior becomes

$$\begin{aligned}
\log p(y = c | \boldsymbol{x}; \boldsymbol{\theta}) &= \log \pi_c - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c) + cst \\
&= \log \pi_c - \frac{1}{2}\boldsymbol{\mu}_c^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + cst \\
&= a_c + \boldsymbol{x}^\top b_c + cst
\end{aligned}$$

The discriminant function is a linear function of $\boldsymbol{x}$. This is called **Linear discriminant analysis** or **LDA**.



(a) *(b)*

Figure: QDA and LDA fit to data from 3 classes

Fitting the model

Using MLE, the likelihood function is

$$\prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \prod_{i=1}^{N} \mathrm{Cat}(y_i|\boldsymbol{\pi}) \prod_{c=1}^{C} \mathcal{N}(\boldsymbol{x}_i;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)^{\mathbb{I}(y_i=c)}$$

$$= \prod_{i=1}^{N} \prod_{c=1}^{C} \pi_c^{\mathbb{I}(y_i=c)} \prod_{c=1}^{C} \mathcal{N}(\boldsymbol{x}_i;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)^{\mathbb{I}(y_i=c)}$$

the log-likelihood is

$$\left[\sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{I}(y_i = c)\log \pi_c\right] + \sum_{c=1}^{C}\left[\sum_{i:y_i=c} \log \mathcal{N}(\boldsymbol{x}_i;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\right]$$

The parameters $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ can be optimized separately. The MLE for the prior is

$$\hat{\pi}_c = \frac{N_c}{N}$$

The MLE for the Gaussians are as follows :

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \boldsymbol{x}_i$$

$$\hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c} \sum_{i:y_i=c} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_c)^\top$$

where N_c is the number of observations whose class is c.
For LDA, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$, the covariance matrix estimate is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i:y_i=c} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_c)^\top$$

Regularized discriminant analysis

MLE $\hat{\Sigma}_c$ can overfit if N_c is small compared to D. Forving the $\hat{\Sigma}_c$ to be diagonal can overcome the problem.

We can also use a MAP estimate of a shared full covariance matrix with an inverse Wishart prior (distribution over positive definite matrices). The MAP estimate is

$$\hat{\Sigma}_{MAP} = \lambda diag(\hat{\Sigma}_{MLE}) + (1 - \lambda)\hat{\Sigma}_{MLE}$$

where λ controls the regularization. This is called **Regularized discriminant analysis**. There are robusts ways to invert $\hat{\Sigma}_{MAP}$.

There is more to Gaussian discriminant analysis.

Explore further

- Diagonal covariances (Naive Bayes assumption)
- Shared diagonal covariance matrix (Diagonal LDA)
- Nearest centroid classifier (nearest class mean classifier)
- Nearest class mean metric learning
- Fisher's linear discriminant analysis
- Connection between LDA and logistic regression