Linear Discriminant Analysis

S. Marchand-Maillet

Computer Science – University of Geneva

1 Introduction

Linear Discrimination for classification in Machine Learning comes as a conjunction of several factors [2]:

1. The optimality of the Bayes classifier
2. The assumption of Gaussian models for class densities
3. The geometry of Gaussian models

which we detail next.

2 Bayesian theory for classification

Given a labeled information space $\Omega_X \times \Omega_y$, where $\Omega_X \subseteq \mathbb{R}^D$ is the data space and $\Omega_y \subseteq \mathbb{R}$ is the label space and where each $\mathbf{x} \in \Omega_X$ is associated with a label $y_{\mathbf{x}} \in \Omega_y$, we assume a (generally unknown) joint probability distribution $p(X, Y)$ for the related random variables $X \in \Omega_X$ and $Y \in \Omega_y$. Now classification is defined as the map ("the classifier", later "the learner") $\phi \in \mathbb{F}$:

$$\begin{array}{rccc} \phi: & \Omega_X & \rightarrow & \Omega_y \\ & \mathbf{x} & \mapsto & \phi(\mathbf{x}) = \widehat{y} \end{array}$$

which maps a data $\mathbf{x} \in \Omega_X$ onto a (predicted) label $\widehat{y} \in \Omega_y$. $\mathbb{F}$ is a given family of functions.

The evaluation of a classifier ϕ is defined by a loss function $\mathcal{L}(\Omega_X, \Omega_y, \phi) \in \mathbb{R}^+$. The loss function typically measures the ability of the classifier to map data $\mathbf{x}$ onto the true label $y_{\mathbf{x}}$, e.g:

$$\mathcal{L}(\mathbf{x}, y_{\mathbf{x}}, \phi) = \mathbb{1}_{\phi(\mathbf{x}) \neq y_{\mathbf{x}}}$$

Definition 1 (Risk)**.** The risk of classifier $\phi \in \mathbb{F}$ is its expected loss over the information space:

$$\mathcal{R}(\phi) \stackrel{\text{def}}{=} \mathbb{E}_{X,Y}\mathcal{L}(\Omega_X, \Omega_y, \phi) = \int_{\Omega_X \times \Omega_y} \mathcal{L}(\mathbf{x}, y_{\mathbf{x}}, \phi)\mu(d\mathbf{x})\mu(dy)$$

Definition 2 (Bayes classifier)**.** We define the Bayes classifier (as a tribute to its conditional structure) the map that affects the most likely label $y \in \Omega_y$ to any given data $\mathbf{x} \in \Omega_X$:

$$\phi_B(x) \stackrel{\text{def}}{=} \underset{y \in \Omega_y}{\operatorname{argmax}}\, p_{Y|X}(Y = y | X = \mathbf{x})$$

It is then possible to prove

Theorem 1. *Bayes classifier is an optimal classifier with minimal risk*

$$\phi_B = \underset{\phi \in \mathbb{F}}{\operatorname{argmin}}\, \mathcal{R}(\phi) \quad \textit{where} \quad \mathbb{F} = \{\phi : \Omega_X \rightarrow \Omega_y\}$$

Proof. (that Bayes classifier is optimal) TBC □

The optimality of the Bayes classifier motivates the analysis of the distribution $p_{Y|X}(Y = y|X = x)$. Of course in general $p_{Y|X}$ is unknown so that the Bayes classifier cannot operate directly. As a result, several strategies for learning a (empirical) surrogate for the Bayes classifier have been proposed.

These constructions for learning essentially rely on the use of Bayes theorem to emerge the possibility to use empirical knowledge about the information space $\Omega_x \times \Omega_y$ to set the structure of the classifier. In other words, noting that

$$p_{Y|X}(Y = y|X = x) = \frac{1}{Z} p_{X|Y}(X = x|Y = y)\, p_Y(Y = y) \quad \text{where} \quad Z = p_X(X) = \int_{\Omega_y} p_{X|Y}(X = x|Y = y)\, p_Y(Y = y)\mu(dy)$$

one can address the following terms:

- $p_{X|Y}(X = x|Y = y)$: label-conditional (class-conditional) density. This is the density of all data x receiving label y
- $p_Y(Y = y)$ label density. This is the density of label y in label space Ω_y
- $p_X(X)$ data density. This is the density of the data. As evident from above it is used as a normalizer to preserve $\int_{\Omega_x \times \Omega_y} p_{Y|X}(Y = y|X = x)\mu(dx)\mu(dy) = 1$. It is generally not estimated explicitly but implicitly via normalization, which corresponds to using the law of total probabilities (sum rule over all labels).

  Density ratios also circumvent the computation of the normalizer.

In other words, the data density over Ω_x is seen as a mixture of class-conditional densities:

$$pX(x) = \int_{\Omega_y} p_{X|Y}(X = x|Y = y)\, p_Y(y)dy$$

which in case of a discrete label space ($y \in \Omega_y = \{0, \cdots, K-1\}$) becomes

$$p_X(x) = \sum_{y \in \Omega_y} \mathbb{P}_Y(Y = y)\, p_{X|Y}(X = x|Y = y) = \sum_{k=0}^{K-1} \pi_k\, p_k(x)$$

which is the typical mixture model with class-conditional density p_k and mixture weight π_k.

3 Gaussian model for class-conditional densities

One natural model for the class density (given label y) is the Gaussian model. The Gaussian distribution

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

considers a main prototype μ and a likelihood that decreases exponentially with the Malahanobis distance (shaped by Σ) from μ. The Gaussian model can therefore be characterized by its 2 first centered moments μ and Σ, others being zero.

The estimation of the class k parameters follows MLE classical estimates under Maximum Likelihood for Gaussian models[a]. Following MLE, the class mean μ_k can be computed as the empirical mean of the class data. The covariance matrix Σ_k can also be computed as MLE estimate. In other words,

$$N_k = \sum_{i=1}^{N} [y_i = k\ ?] \quad \mu_k = \frac{1}{N_k} \sum_{i:y_i=k} x_i \quad \Sigma_k = \frac{1}{N_k} \sum_{i:y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T \quad \pi_k = \frac{N_k}{N} \tag{1}$$

The relationship between Gaussian models and log likelihood[a] creates a direct link between probabilistic discrimination and its geometry.

3.1 Geometry of Gaussian models

Bayes classifier is a maximum likelihood classifier (Definition 2): it allocates the label y with highest likelihood to data x. Classes (of labels y) will therefore compete over Ω_x for label allocation. As it turns out, the geometry of the separating set can be studied in the binary case ($K = 2$) as follows.

[a] Maximum Likelihood Estimate (MLE) and the Gaussian model

Proposition 1. Given $y_0, y_1 \in \Omega_y$ and $p(X = x|Y = y_k) = \mathcal{N}(x|\mu_k, \Sigma_k)$ $(k = 0, 1)$ then the separating set between class y_0 and y_1

- is linear if $\Sigma_0 = \Sigma_1$ (homoscedasticity)
- is quadratic otherwise (heteroscedasticity)

Class populations, as class priors ($\mathbb{P}(Y = k)$) guide the location of the separating set with respect to class means.

Proof. TBC. See proof at [(b)]. □

4 Discriminant Analysis

From the above, effective classifiers may be constructed and learned, based on an empirical estimation of their parameters [3]. In a discrete supervised classification setup, given $\mathcal{X} \times \mathcal{Y} = \{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \Omega_\mathcal{X} \subseteq \mathbb{R}^D$ and $y_i \in \mathcal{Y} = \{0, \cdots, K-1\} \subset \Omega_y$ we seek to train a classifier

$$\begin{array}{rccl} \phi_\theta : & \Omega_\mathcal{X} & \to & \mathcal{Y} \\ & x & \mapsto & \phi_\theta(x) = y \end{array} \qquad \text{where } \phi_\theta \in \mathbb{F} \text{ and } \mathbb{F} \text{ is a family of functions parameterized by } \theta$$

Note that classifier ϕ_θ maps data to a finite discrete set of labels $\mathcal{Y}$ although the above theory is deployed for continuous labels ($y \in \Omega_y \subseteq \mathbb{R}$) and is therefore also fit for a regression setup.

4.1 Binary case

Assume $K = 2$, we use Bayes classification theory to train a classifier discriminating 2 classes, each having a Gaussian model:

$$p(X = x|Y = 0) = \mathcal{N}(x|\mu_0, \Sigma_0) \quad \text{and} \quad p(X = x|Y = 1) = \mathcal{N}(x|\mu_1, \Sigma_1)$$

Recall that

$$\mathbb{P}(Y = k) = \pi_k \qquad k \in \{0, 1\}$$

Bayes classification therefore leads to check whether (Definition 2)

$$\mathbb{P}(Y = 0)\, p(X = x|Y = 0) > \mathbb{P}(Y = 1)\, p(X = x|Y = 1) \quad \text{i.e} \quad \pi_0 \mathcal{N}(x|\mu_0, \Sigma_0) > \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1)$$

which can also be qualified using the log odd ratio:

$$\log \frac{\mathbb{P}(Y = 0)\, p(X = x|Y = 0)}{\mathbb{P}(Y = 1)\, p(X = x|Y = 1)} > 0 \quad \text{i.e} \quad \log \frac{\pi_0 \mathcal{N}(x|\mu_0, \Sigma_0)}{\pi_1 \mathcal{N}(x|\mu_1, \Sigma_1)} > 0$$

4.1.1 Linear Discriminant Analysis: LDA

Here, LDA consists in estimating separate means μ_0 and μ_1 and a unique covariance matrix Σ for all classes. While the mean and mixture weight estimation follows equation (1), the shared covariance matrix is estimated as the weighted average of individual class covariance matrices:

$$\Sigma = \frac{1}{N}(N_0 \Sigma_0 + N_1 \Sigma_1) \tag{2}$$

As mentioned above, classes with Gaussian conditional model and equal covariance matrices Σ are separated by a linear set (hyperplane)[(b)].

It is easy to show that in the binary case, the separating hyperplane is orthogonal to the line joining the 2 means (μ_0, μ_1)[(b)], located closer to the largest class $k \in \{0, 1\}$ proportionally to π_k.

4.1.2 Quadratic Discriminant Analysis: QDA

Again as shown from the geometry of Gaussian models, Gaussian class-conditional densities with individual covariance matrices Σ_k show quadratic separating sets. Here, the estimation again follows MLE for μ_k and preserves individual covariance matrices Σ_k.

Again, in the binary case, one can show that the separating set is "bent" around the mean of smallest covariance norm[(b)] and "attracted" by the most populated class according to π_k.

[(b)] Geometry of Gaussian models

4.2 General case

In the general case of $K > 2$, equations (1) remain valid but there is no simple interpretation of the geometry of class competition. For multiclass LDA, equation (2) becomes

$$\boldsymbol{\Sigma} = \sum_{k=0}^{K-1} \frac{N_k}{N} \boldsymbol{\Sigma}_k$$

5 Extension: Fisher criterion

The above technique emphasizes class separation. This is visible in the binary case where the separating hyperplane is orthogonal to the line joining the two means. This would be called the between criterion in Fisher discrimination theory [1]. This criterion is mirrored by the within criterion, using the direction of minimal variance within the class as separating criterion. Fisher criterion proposes a trade-off between these two competing notions

$$\text{class separation} = \underset{\text{direction}}{\text{argmax}} \frac{\text{between criterion}}{\text{within criterion}}$$

References

[1] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg, 2006. (available online).

[2] Luc Devroye, Laszlo Györfi and Gabor Lugosi. A probabilistic theory of pattern recognition. Springer, 1996.

[3] Trevor Hastie, Robert Tibshirani and Jerome Friedman. The Elements of Statistical Learning. Data mining, Inference and Prediction. Springer Series in Statistics. Springer New York Inc., 2nd edition, 2008.