Análisis Lineal Discriminante

Motivación

Preliminares

- El Análisis Lineal Discriminante (ALD) es otra variante de algoritmo generativo.
- En Bayes Ingenuo evaluamos la ocurrencia de una clase, condicional a la distribución de la verosimilitud.
- En ALD, la ocurrencia de una clase se evalúa en función a distribuciones multivariadas y una frontera de decisión.
- Busca generar una nueva reexpresión de los datos en una dimensión en base a distribuciones multivariadas normales.

Distribución Multivariada Normal

Distribución Multivariada Normal

$$\mathsf{MultiNormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{\mathbb{N}/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) \right]$$

- La distribución multivariada normal permite evaluar el comportamiento de un conjunto de atributos y datos en cuanto a una distribución más compleja.
- A grandes rasgos, es una expansión de la distribución normal ya estudiada, considerando la inclusión de dos o más variables.

Distribución Multivariada Normal

$$\mathrm{MultiNormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{\mathbb{N}/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right]$$

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

Vector de medias para
cada uno de los atributos

Distribución Multivariada Normal

$$\mathsf{MultiNormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{\mathbb{N}/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) \right]$$

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}^2 & \sigma_{m2} & \cdots & \sigma_{mn}^2 \end{bmatrix}$$

Matriz de varianza (elementos diagonales) y covarianza (elementos en los triángulos)

Distribución Multivariada Normal

$$\mathsf{MultiNormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{\mathbb{N}/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu})\right]$$

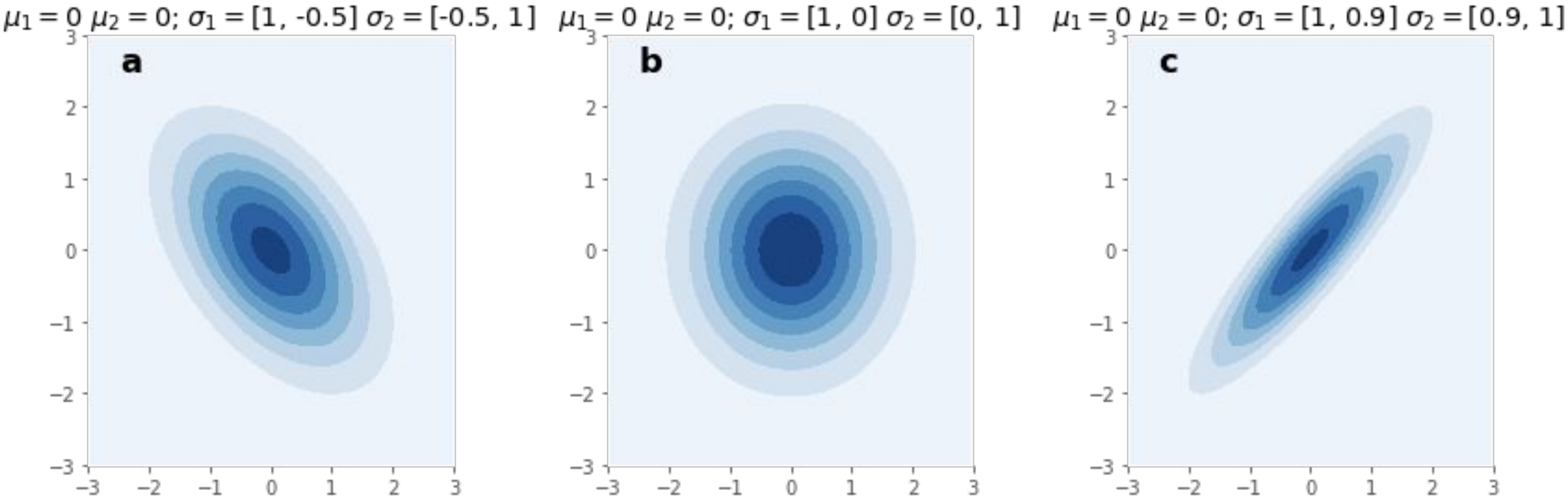
Evalúa la distancia de un punto específico respecto a media de todos los atributos existentes.

Distribución Multivariada Normal

$$\mathsf{MultiNormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{\mathbb{N}/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right]$$

Regula la forma de la elipse.

Distribución Multivariada Normal



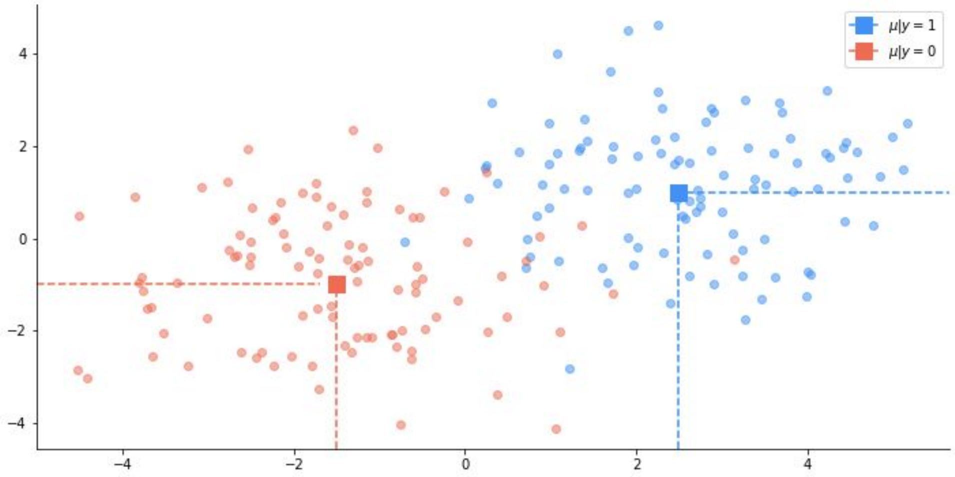
La idea base detrás de LDA

Preliminares

El análisis lineal discriminante se puede representar en tres pasos:

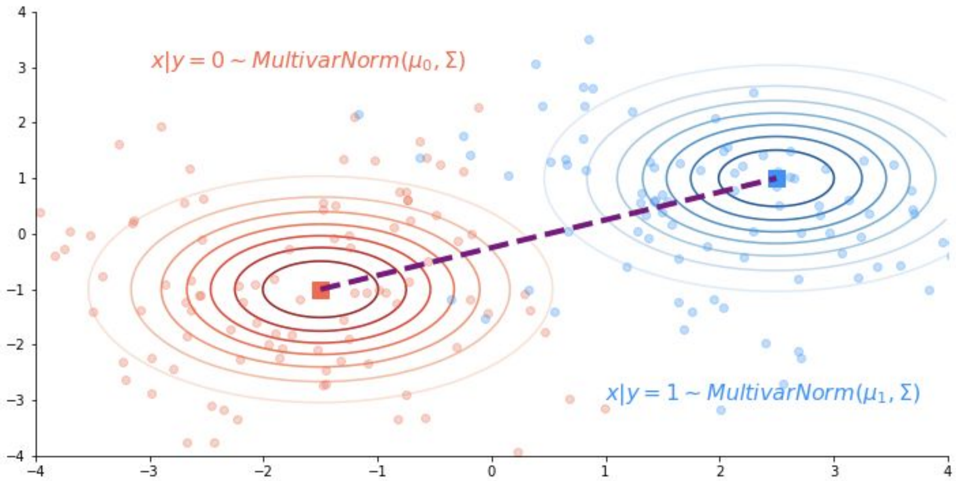
- Extraer parámetros de los datos.
- Estimar las distribuciones en base a los parámetros.
- Reexpresar la clasificación en función a una nueva proyección de datos.

Extracción de parámetros para cada clase



Buscamos las medias de cada atributo $\boldsymbol{\mu} \quad \forall x \in \mathbf{X}$ condicional la clase.
Hasta el momento, asumimos que la matriz de covarianza será idéntica para cada clase.

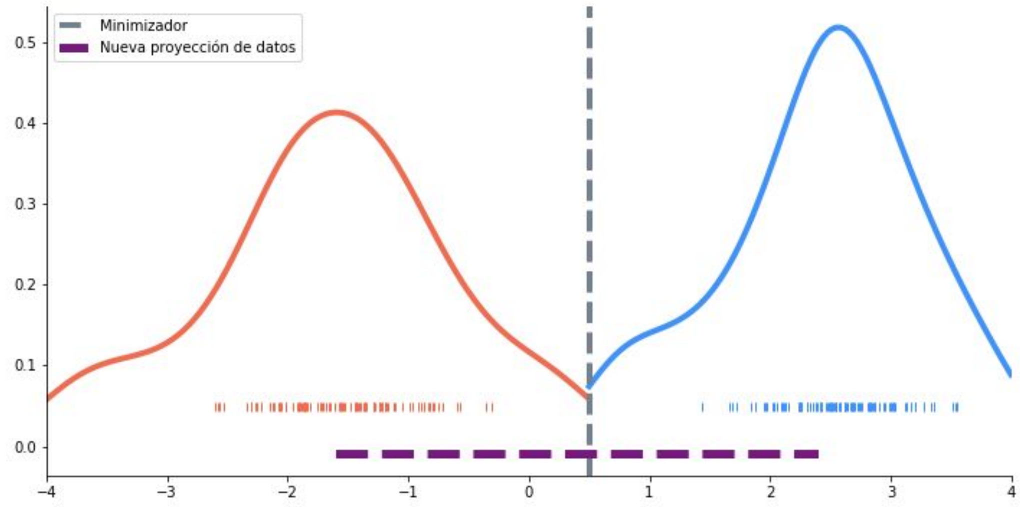
Estimar las distribuciones



Estimamos las elipses de cada clase mediante $x|y \in \mathbb{Y} \sim \mathsf{MultivarNorm}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

En base a las elipses evaluamos la varianza interna y externa.

Preliminares



Mediante la maximización de la varianza externa, evaluamos un discriminante lineal.
Este discriminante buscará reducir la superposición de las observaciones de cada clase.

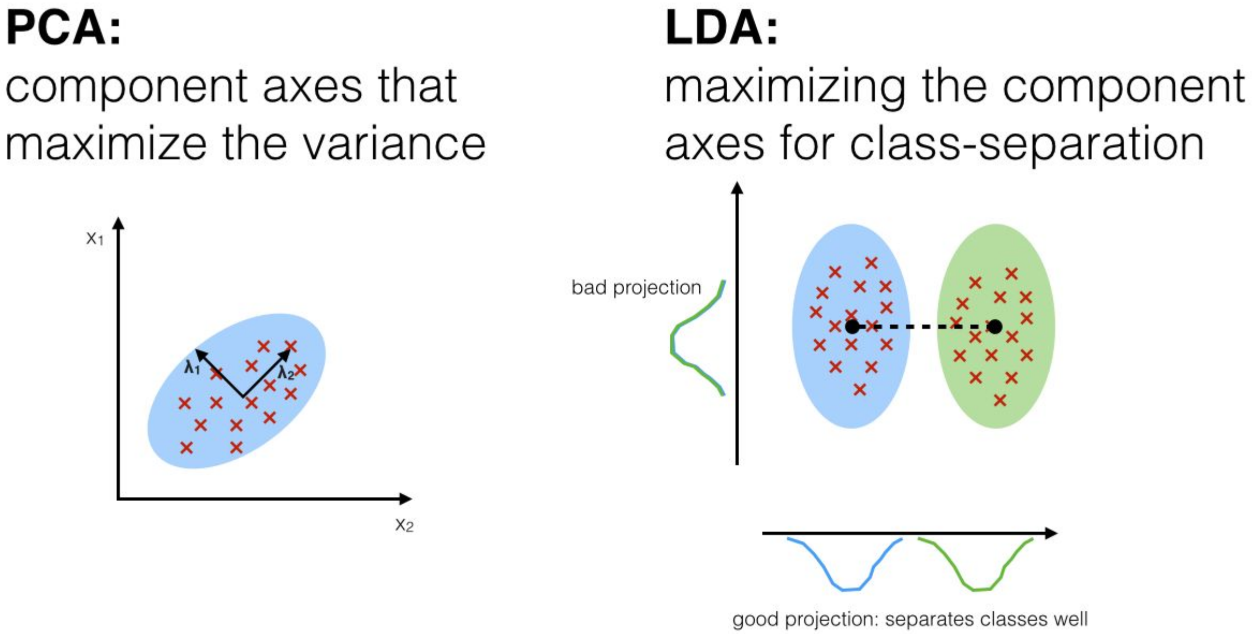
Quadratic Discriminant Analysis

Consideraciones

- **Principal diferencia con ALD clásico**: Los atributos siguen una forma cuadrática.
- **Otra virtud**: la matriz de covarianza es condicional a cada una de las clases.
- De esta manera, ambos elementos apuntan a flexibilizar el comportamiento de las funciones de decisión.

Apreciaciones

Análisis Lineal Discriminante vs PCA



- Ambos métodos buscan generar una nueva proyección entre los datos.
- ALD maximiza la separación de clases.
- PCA maximiza la varianza en los datos, agnóstica a las clases existentes.

Análisis Lineal Discriminante y Regresión Logística

- Ambos métodos generan resultados similares.
- La principal diferencia es cómo se genera el fit en cada uno.
- ALD será asintóticamente eficiente si es que se cumple el supuesto de distribución gaussiana a nivel de clase.

{desafío} latam_

Academia de talentos digitales

www.desafiolatam.com