Análisis de Discriminantes y Regresión Logística

Tomás Fontecilla

26 de agosto de 2022

Anásis Discriminante - Motivación

- Queremos clasificar diversos países en función de sus datos macroeconómicos:
  Países subdesarrollados, emergentes o desarrollados (grupos). Creamos la función discriminante de manera que podamos calcular la probabilidad de un país de pertenecer a uno u otro grupo.

- Queremos realizar una campaña de marketing y nos interesa saber en qué grupos clasificar a los individuos:
  Así, podremos responder a ciertas preguntas como cuáles serían las características de un cliente ocasional.

- Queremos conocer el nivel de riesgo (grupo) de determinados clientes respecto a la concesión de un crédito:
  Utilizaremos variables relativas a su renta, gastos mensuales, historial o tipo de trabajo. La función discriminante nos aporta información relevante sobre la solvencia.

Anásis Discriminante

El análisis de discriminante lineal, o LDA usualmente es usado como una técnica de reducción de dimensiones. En este sentido, es usado como un paso previo en el preprocesamiento de datos en Machine Learning y en aplicaciones de clasificación de patrones.

El objetivo de LDA es **proyectar las características en un espacio de dimensionalidad alta a uno de menor dimensionalidad**, con el fin de evitar la *maldición de la dimensionalidad* y también reducir los costos y recursos dimensionales.

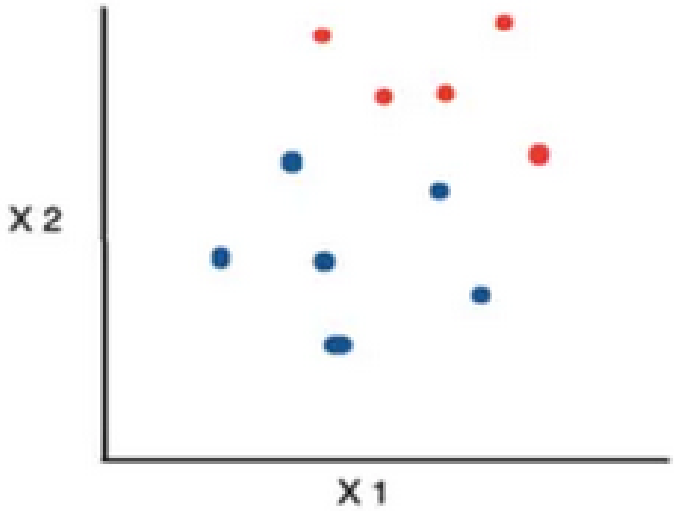
Análisis de Discriminante - Maldición de dimensionalidad

La maldición de la dimensionalidad hace referencia a varios fenómenos que surgen cuando se analiza y organiza espacios de alta dimensionalidad que no ocurren en la configuración de baja dimensioanlidad como sería el caso del espacio tridimensional físico de las experiencias diarias.

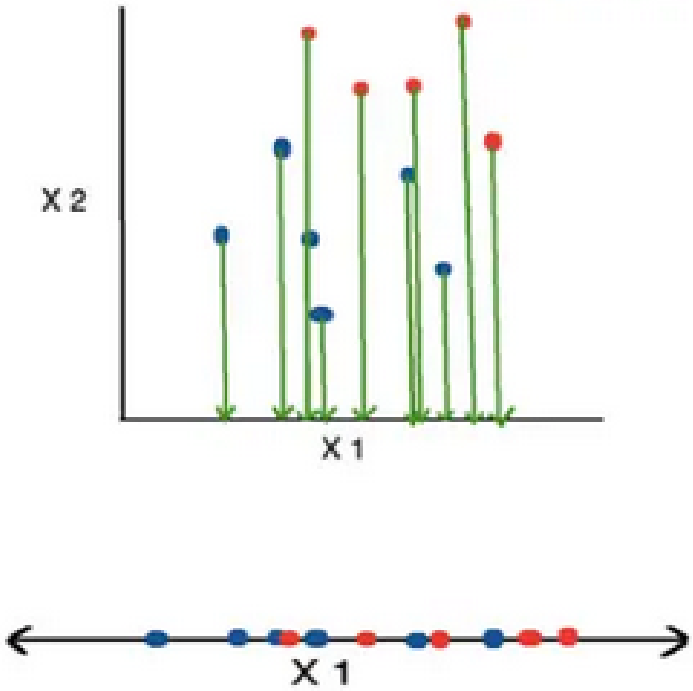
En machine learning, al crecer el número de características o dimensiones el monto de datos necesarios para generalizar certeramente crece exponencialmente.

Análisis de Discriminante - Enfoque Práctico - Motivación

Consideremos la situación donde hemos graficado la relación entre dos variables donde los colores representan su clase.



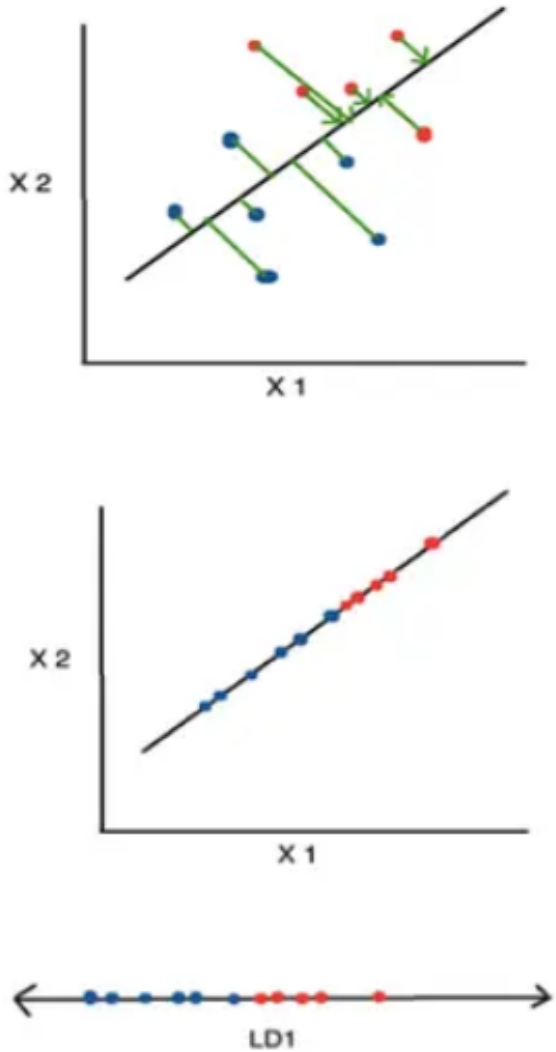
Si uno está dispuestoa reducir el número de dimensiones a 1, se puede proyectar todo sobre un eje de la siguiente forma:



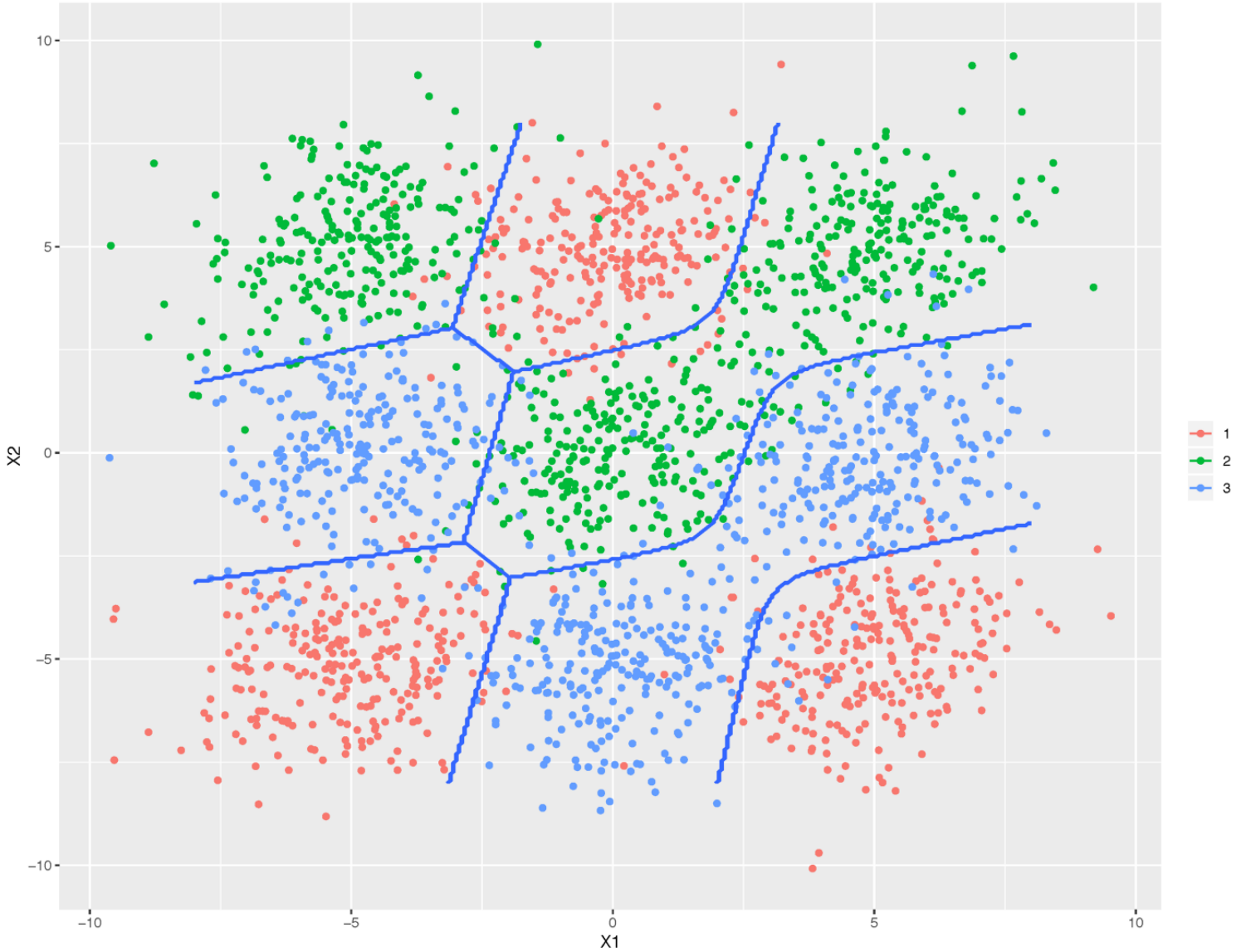
Este enfoque desecha cualquier información útil que pueda entregar el segundo "feature".

Análisis de Discriminante - Enfoque Práctico

La ventaja de LDA es que usa la información de las dos características y crea un nuevo eje que a su vez minimiza la varianza y maximiza la distancia de clase de las dos variables. Esto se puede ver en el siguiente gráfico:



Análisis de Discriminante - Ejemplo gráfico 3 dimensiones



Análisis de Discriminante - Cómo Funciona

El LDA se enfoca primariamente en proyectar las características en un espacio de dimensiones superiores a uno de dimensiones inferiores. Esto se puede hacer en tres pasos:

- Primero, se calcula qué tan separadas están las clases usando la distancia entre las *medias* de las diferentes clases. Esto es lo que se conoce como la *varianza entre clases* (S_b).
- Segundo, calcula la distancia entre las medias y la muestra de cada clase. Esta es de especial relevancia y se conoce como *varianza intra clases* (S_w).
- Finalmente, construimos un espacio de dimensionalidad más baja que maximice la varianza entre clases y minimice la varianza intra clase. Llamaremos P a la proyección en el espacio de menor dimensiones, también llamado el criterio de Fisher (P_{lda}).

- $S_b = \sum_{i=\mathbf{1}}^{g} (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T$
- $S_w = \sum_{i=\mathbf{1}}^{g} (N_i - 1) S_i = \sum_{i=\mathbf{1}}^{g} \sum_{j=\mathbf{1}}^{N_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$
- $P_{lda} = \underset{\mathbf{P}}{\text{argmax}} \ \frac{P^T S_b P}{P^T S_w P}$

Análisis de Discriminante - Representación

El modelo consiste en una serie de propiedades estadísticas de los datos que han sido calculados para cada clase.

Las mismas propiedades son calculadas sobre la gaussiana multivariada en el caso de variables múltiples. Esto se traduce en medias multivariadas y una matriz de covarianzas.

Se hacen entonces predicciones proveyendo las propiedades estadísticas en las ecuaciones del LDA. Las propiedades son estimadas a los datos.

Finalmente, los valores del modelo son guardados para crear el modelo de LDA.

Análisis de Discriminante - Preparando los datos

- El LDA es principalmente usado en problemas de clasificación con un output categórico, *no necesariamente binario*.
- El modelo estándar usa la distribución normal o Gaussiana de los datos predictores. Sería apropiado verificar que las distribuciones univariadas de cada atributo pertenezca a esta distribución o alternativamente *transformarla* a ella. Un ejemplo, para la distribución exponencial, transformar la variable en un logaritmo.
- Outliers pueden perjudicar los estadísticos básicos usados para separar las clases, así que es buena práctica removerlos.
- Como el LDA asume que cada variable input tiene la misma varianza, siempre es mejor estandarizar los datos antes de usar este modelo.

Análisis de Discriminante - Cómo aprende el modelo

Supuestos

- cada variable en los datos tiene forma gaussiana.
- Los valores de cada variable varían en torno a la media por la misma cantidad en el promedio, i.e., cada atributo tiene la misma varianza.

Con esto, el modelo LDA es capaz de estimar la media y varianza desde los datos para cada clase.

Análisis de Discriminante - Cómo aprende el modelo

La media (μ) de cada valor de entrada (x) para cada clase (k) puede ser estimada en la forma usual al dividir la suma de los valores por el número total de valores.

$\mu_k = \frac{1}{n_k} \cdot \sum(x)$
Donde μ_k es el valor medio de x para la clase k y n_k es el número de instancias en la clase k

La varianza es calculada a traés de todas las clases como el promedio del cuadrado de la diferencia entre los valores y la media.

$\sigma^2 = \frac{1}{(n-k)} \cdot \sum((x - \mu))^2)$
Donde σ^2 es la varianza a través de todos los inputs (x), n es el número de instancias, k es el número de clases y μ es la media de los input x.

Análisis de Discriminante - Predicciones

El análisis discriminante hace predicciones estimando la probabilidad que un nuevo conjunto de predictores pertenezcan a cada clase. La clase con mayor probabilidad es la clase de salida en la predicción.

El modelo usa el teorema bayesiano para estimar las probabilidades. Esto es:

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}_k \cdot f_k(x)}{\sum \mathbb{P}_i(x) f_i(x)}$$

Donde $\mathbb{P}_k$ es la probabilidad base de pertenecer a la clase (k) observada en los *datos de entrenamiento*. $f_k(x)$ es la verosimilitud de pertenecer a la clase k, representado por una distribución gaussiana.

Todo esto resulta en la función discriminante:

$$D_k(x) = x \cdot (\frac{\mu_k}{\sigma^2}) - \frac{\mu_k^2}{2 \cdot \sigma^2} + ln(P_k)$$

Donde D_k es la función discriminante para la clase k dado el input x, y μ_k,σ^2 y P_k son estimados de los datos.

Regresión Logística

¿ Qué ocurre si queremos *explicar* los pesos de cada una de las variables?
El Análisis por discriminante lineal busca la frontera que minimiza la varianza dentro de los grupos maximizando la distancia entre los grupos, a través de proyecciones y algebra lineal, utilizando los valores propios y vectores propios.
En este caso, utilizaremos como respuesta una variable dicotómica donde podremos ver qué peso tiene cada variable en el modelo.

Regresión Logística - Definición

La regresión logística analiza datos distribuidos de forma binomial, es decir, cada observación corresponde a una distribución bernoulli, por la cantidad de observaciones que se haga el experimento. Así

$$Y_i \sim Ber(p_i) \text{ para } i = 1, ..., n$$

El modelo se basa en que cada ensayo toma uno de dos valores *explicado* por las variables independientes X_i. De esta forma, $p_i = \mathbb{E}(\frac{Y_i}{n_i}|X_i)$
Dado que Y_i es una variable dicotómica, no es modelable como tal, por lo que modelaremos la razón de chances $\frac{p_i}{1-pi}$ que sigue una distribución similar a la exponencial.

Regresión Logística - Modelo

Luego, aunque las chances no tiene una forma lineal, sí lo tiene su logaritmo, que es la función de *enlace* de esta regresión. Por tanto, el modelo es:

Definición

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + ... + \beta_k \cdot X_{k,i}$$

Es de nuestro interés ajustar el modelo para estimar los parámetros β_k, que se hace utilizando Máxima Verosimilitud.

Regresión Logística - Resultados

Luego, para efectos de clasificación, el modelo entregará el logaritmo del odds que podemos despejar y obtener la estimación asociada a p_i

$$log(\frac{p_i}{1-p_i}) = \sum_{k=0}^{k} \beta_k \cdot X_{k,i}$$
$$\frac{p_i}{1-p_i} = e^{\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}$$
$$p_i = e^{\sum_{k=0}^{k} \beta_k \cdot X_{k,i}} - p_i \cdot e^{\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}$$
$$p_i = \frac{e^{\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}}{1+e^{\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}}$$
$$p_i = \frac{1}{1+e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}}$$

Luego, las chances que la observación pertenezca al "éxito" está dado por :

$$\frac{p_i}{1-p_i} = \frac{\frac{1}{1+e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}}}{1-\frac{1}{1+e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}}} = \frac{1}{e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}}$$

Esta última se conoce como sigmoide y es muy utilizada en redes neuronales e inteligencia artifical.

Regresión Logística - Resultados interesantes

Lo curioso de estos resultados anteriores es que permite revisar comparaciones de cambio entre actividades.
Ejemplo de esto, es, suponiendo que todas las variables permanecen en el tiempo iguales, excepto X_3, entonces el incremento de la razón de chances es:

$$\begin{aligned}
\frac{\frac{p_i}{1-p_i}}{\frac{p_j j}{1-p_j}} &= \frac{\frac{1}{e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}}}{\frac{1}{e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,j}}}} \\
&= \frac{e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,j}}}{e^{-\sum_{k=0}^{k} \beta_k \cdot X_{k,i}}} \\
&= e^{\beta_3 \cdot (X_{3i} - X_{3j})}
\end{aligned}$$

A efectos de modelos computacionales, se debe tener cuidado, ya que éstos entregarán el valor de β_k, pero el resultado interesante es su exponencial.

Regresión Logística - Validación

- Para determinar el desempeño general, es comúnmente usada el área bajo la curva de ROC o "Reciever operating characteristic" como indicador de qué tan bien predice el modelo.
- Precisión, Recall y Certeza según vimos la clase pasada.
- Por otra parte, el test de Wald es usado para comparar entre modelos, determinando la significancia de las variables predictoras.