Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 14: 21/10/24 – 26/10/24

Introducción

Vimos que:

- ✓ En el análisis discriminante (AD) se obtiene una función (discriminante) que separa varios grupos definidos a priori.
- ✓ La función discriminante es una combinación de las variables X, que minimiza los errores de clasficación.
- ✓ Uno de los enfoques del AD es comprobar si las $X's$ permiten diferenciar los grupos/poblaciones definidas previamente.
- ✓ Otro enfoque del AD clasificar a "nuevas" unidades muestrales, en uno de los grupos de acuerdo con una regla de clasificación/localización.

Introducción

- ✓ Reglas de discriminación para $g = 2$ grupos:
  - Función Discriminante de Fisher ($\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$).
  - Función Discriminante Cuadrática ($\boldsymbol{\Sigma_1} \neq \boldsymbol{\Sigma_2}$).
  - Discriminaación Bayesiana.
  - Discriminante Logística.
- ✓ Reglas de discriminación para g grupos

Introducción

- ✓ Reglas de discriminación para $g = 2$ grupos:
  - Función Discriminante de Fisher ($\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$).
  - Función Discriminante Cuadrática ($\boldsymbol{\Sigma_1} \neq \boldsymbol{\Sigma_2}$).
  - Discriminaación Bayesiana.
  - Discriminante Logística.
- ✓ Reglas de discriminación para g grupos

Sabemos que a mayor intersección entre los grupos/poblaciones, mayor es la probabilidad de clasificaciones incorrectas (erróneas).

Introducción

- ✓ Reglas de discriminación para $g = 2$ grupos:
  - Función Discriminante de Fisher ($\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$).
  - Función Discriminante Cuadrática ($\boldsymbol{\Sigma_1} \neq \boldsymbol{\Sigma_2}$).
  - Discriminaación Bayesiana.
  - Discriminante Logística.
- ✓ Reglas de discriminación para g grupos

Sabemos que a mayor intersección entre los grupos/poblaciones, mayor es la probabilidad de clasificaciones incorrectas (erróneas). A continuación estudiaremos métodos para estimar las tasas de error de discriminación.

Estimación de Tasas de error de Discriminación.

Una vez que se ha obtenido una regla de clasificación, debemos/queremos determinar la tasa de clasificación correcta, esto es, la probabilidad de clasificar una unidad muestral en el grupo/población al que realmente pertenece. Así mismo, se interés en la tasa de error de clasificación incorrecta, que es complementaria a la anterior. Además, se quiere evaluar la capacidad de la regla para predecir el grupo a que pertenece una unidad muestral.

Estimación de tasas de error de discriminación.

La siguiente tabla ilustra la calidad de las posibles decisiones que se podrían tomar, con relación a la clasificación de unidades muestrales en uno de dos grupos.

Grupos	Decisión		Total
	Grupo 1	Grupo 2	
1	n_{11}	n_{12}	n_1
2	n_{21}	n_{22}	n_2

Observe que

- Entre las n_1 observaciones del Grupo 1, n_{11} son clasificadas **correctamente** en el Grupo 1 y n_{12} son clasificadas incorrectamente. Por lo que en este grupo:

$$\text{Prop. correcta} = \frac{n_{11}}{n_1} \qquad \text{Prop. incorrecta} = \frac{n_{12}}{n_1}$$

Estimación de tasas de error de discriminación.

- Entre las n_2 observaciones del Grupo 2, n_{22} son clasificadas **correctamente** en el Grupo 2 y n_{12} son clasificadas incorrectamente. Por lo que en este grupo:

$$\text{Prop. correcta} = \frac{n_{22}}{n_2} \qquad \text{Prop. incorrecta} = \frac{n_{21}}{n_2}$$

De esta forma, la tasa de error (clasificaciones incorrectas) llamada **tasa de error aparente** es:

$$t = \frac{n_{12} + n_{21}}{n_1 + n_2} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$$

Estimación de tasas de error de discriminación.

Si p_1 y p_2 son las respectivas probabilidades apriori de los grupos 1 y 2, la **tasa actual de error (TAE)** es:

$$TAE = p_1 P[\text{Asignar a G1}|\text{Pert. a}G2] + p_2 P[\text{Asignar a G2}|\text{Pert. a}G1]$$

Podemos estimar (basados en todas las muestras posibles) TAE:

$$\begin{aligned} TAEE = p_1 \, \mathrm{E}\,[P[\text{Asignar a G1}|\text{Pert. a}G2]] \\ + p_2 \, \mathrm{E}\,[P[\text{Asignar a G2}|\text{Pert. a}G1]] \end{aligned}$$

Para calcular TAE o TAEE, necesitamos conocer los parámetros poblacionales, los cuales son desconocidos, por lo que se requiere de algunos estimadores de la tasa de error.

Estimación de tasas de error de discriminación.

Si p_1 y p_2 son las respectivas probabilidades apriori de los grupos 1 y 2, la **tasa actual de error (TAE)** es:

$$TAE = p_1 P[\text{Asignar a G1}|\text{Pert. a}G2] + p_2 P[\text{Asignar a G2}|\text{Pert. a}G1]$$

Podemos estimar (basados en todas las muestras posibles) TAE:

$$\begin{aligned} TAEE = p_1 \, \mathrm{E}\,[P[\text{Asignar a G1}|\text{Pert. a}G2]] \\ + p_2 \, \mathrm{E}\,[P[\text{Asignar a G2}|\text{Pert. a}G1]] \end{aligned}$$

Para calcular TAE o TAEE, necesitamos conocer los parámetros poblacionales, los cuales son desconocidos, por lo que se requiere de algunos estimadores de la tasa de error. Generalmente, para tamaños de muestra grande, la tasa de error actual tienen un sesgo pequeño, mientras que para muestras pequeñas éste suele ser considerable.

Estimación de tasas de error de discriminación.

Existen algunas técnicas que permiten reducir el sesgo en la estimación de la tasa de error aparente:

1. **Partición de la muestra**
   La muestra se divide en dos:
   - Muestra 1 (test): se usa para construir la regla de clasificación.
   - Muestra 2: Se usa para evaluar la regla definida en la muestra 1. La evaluación se hace en cada una de las observaciones de la muestra 2.
     Observación:como las observaciones en la muestra 2 no se usaron en la construcción de la regla de clasificación, la tasa de error resultante es insesgada.

Estimación de tasas de error de discriminación.

La estimación de las tasas de error pueden ser mejoradas intercambiando el papel de las dos muestras, esto es, obtener la regla de clasificación a partir de la muestra de validación, y hacer la validación con la muestra test. Luego, la tasa de error estimada, es el promedio de las dos tasas de error calculadas.

2. **Validación Cruzada**
   Este es un caso particular del anterior, pues se toman $n-1$ observaciones para construir la regla de clasificación y luego con ella se clasifica la observación omitida. El proceso se repite para cada observacón n veces.

Estimación de tasas de error de discriminación.

3. **Bootstrap**
   Escencialmente es una corrección del sesgo para la tasa error aparente. El procedimiento para $g = 2$ grupos consiste en:
   - **Paso 1:** seleccione aleatoriamente una muestra con reemplazo de tamaño n_1 de Grupo 1.
   - **Paso 2:** seleccione aleatoriamente una muestra con reemplazo de tamaño n_2 de Grupo 2.
   - **Paso 3:** Construya la regla de clasificación apartir de las muestras obtenidas en los pasos 1 y 2.
   - **Paso 4:** Clasifique las observaciones de las muestras originales con base en la regla obtenida en el paso 3

Estimación de tasas de error de discriminación.

Las tasas de error de clasificación en cada grupo está dada por

$$TE_i = \frac{e_{i.ori} - e_{i.cla}}{n-1} \qquad i = 1, 2$$

donde:

- $e_{i.ori}$ es el número de observaciones del i-ésimo grupo original incorrectamente clasificada.
- $e_{i.cla}$ es el número de observaciones de la i-ésima muestra nueva que fueron mal clasificada

El procedimiento debe repetirse 100-200 veces y el promedio de TE_i se emplea como corrector del término de sesgo:

$$TE_{boot} = t + \overline{TE}_1 + \overline{TE}_2$$