

# Análisis de datos

Selección de features

# ¿Por qué es importante reducir la cantidad de features?

Maldición de la dimensión (curse of dimensionality):

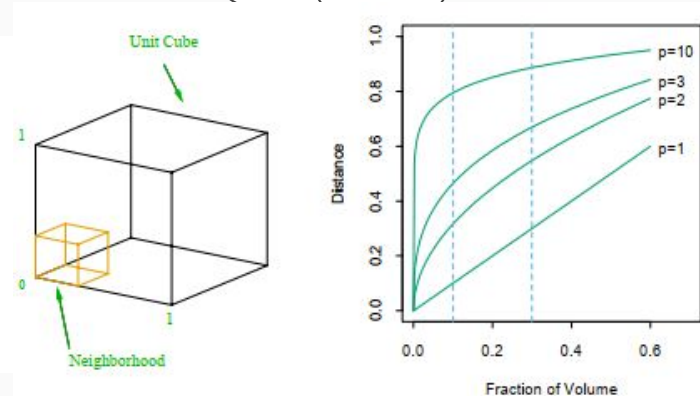
↑ la cantidad de  
dimensiones del  
espacio de features



↑ en volumen del  
espacio



las muestras que  
tenemos se vuelven  
poco representativas  
(muestra pequeña)



# Distintas formas de reducir dimensiones

- Selección de features: elegimos, con cierto criterio, un subconjunto de los variables originales. Existen tres enfoques:
  - **Métodos de filtrado:** se realiza un análisis supervisado de los features para determinar cuáles son los más relevantes, y sólo luego se procede al modelado. Ej. selección basada en test estadísticos.
  - **Métodos embebidos:** la selección de features se encuentra naturalmente incorporada al proceso de modelado. Ej: árboles de decisión, LASSO
  - **Métodos Wrapper :** emplean un método iterativo de búsqueda, donde en cada paso se da al predictor un subconjunto distinto de features, y utiliza la performance del predictor para guiar la selección del siguiente subconjunto de variables. Ej: eliminación recursiva de features (*recursive feature elimination* - RFE)
- Métodos de proyección de variables: busco transformar mis variables para llevarlas a un espacio de menor dimensión. Ejemplo: PCA, ICA, SVD, etc.

# Métodos básicos de selección

1. Eliminar variables constantes: Si existe algún feature que toma siempre el mismo valor para todas las mediciones, debemos quitarlo
2. Eliminar variables cuasi-constantes: una buena idea puede ser eliminar variables cuya varianza sea muy pequeña.
3. Eliminar variables duplicadas
4. Eliminar variables muy correlacionadas. ¿Cómo elegir entre todas las variables correlacionadas?
  - La que tenga menos # de datos faltantes
  - Elegir la más correlacionada con la variable de salida
  - Entrenar algún algoritmo de ML con las variables correlacionadas y elegir la más informativa

**Observación:** Estos métodos son no paramétricos, ya que no dependen de la variable de salida

# Métodos de filtrado

Los métodos de filtrado disponibles dependen de los tipos de las variables de entrada y salida.

Caso	Variable de Entrada	Variable de Salida	Método
1	Númerica	Numérica	Pearson, Spearman's, Información Mutua
2	Númerica	Categórica	ANOVA, Kendall's, Información Mutua
3	Categórica	Numérica	Poco frecuente.
4	Categórica	Categórica	$\chi^2$ , Información Mutua

# Numérica-Numérica

## Coeficiente de correlación de Pearson

- Coeficiente de correlación de Pearson:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- Asume que las variables siguen una distribución normal
- Es un estimador de  $\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \Rightarrow$  mide relación lineal entre variables
- Bajo la hipótesis nula que las variables están descorrelacionadas (independientes)
  - $t = r \sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$
  - $z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) = \text{arctanh}(r) \approx \mathcal{N} \left( 0, \frac{1}{\sqrt{n-3}} \right)$  (transformación de Fisher)
  - $f = \frac{r^2}{1-r^2} (n-1) \sim F_{1,n-2}$  . Representa la proporción de la varianza explicada por una función lineal de la variable X.

# Numérica-Numérica

## Coeficiente de Spearman

- **Coeficiente de Spearman**

$$\rho = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}, \text{ si no hay valores repetidos: } \rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}, d_i = (rg(x)_i - rg(y)_i)$$

- Es un método no paramétrico, basado en estadísticas de orden
- Mide la relación monotónica entre variables las variables
- Bajo la hipótesis de variables independientes,  $t = r \sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$
- Menos sensible a outliers

# Numérica - Numérica

## Información Mutua

- Información mutua:

Recordemos que  $I(X, Y) = \int \int f_{X,Y}(x, y) \log \left( \frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)} \right) dx dy = H(X) - H(X|Y)$

- Debemos estimar las funciones de densidad (probabilidad). El algoritmo de Scikit-learn lo hace basándose en el principio de vecinos más cercanos (A. Kraskov, H. Stogbauer and P. Grassberger, “Estimating mutual information”. Phys. Rev. E 69, 2004.).
- No paramétrico (no hace suposiciones de la distribución de las variables)
- Permite identificar relaciones no lineales entre variables. Si  $I(X, Y) = 0 \Rightarrow X, Y$  son independientes



# Numérica-Categorica

## ANOVA

Cálculo del estadístico:

- Calculamos la media total  $\bar{x} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{N}$ , donde  $N = \#$  de muestras y  $n_i = \#$  de muestras de clase  $i$
- Estimamos la varianza entre grupos  $S_e^2 = \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2}{k-1}$
- Estimamos la varianza dentro de los grupos  $S_d^2 = \frac{\sum_{i=1}^k (n_i - 1) S_k}{N - k}$
- Definimos el test  $F = \frac{S_e^2}{S_d^2} \sim F_{k-1, N-k}$

F va a ser grande si la varianza entre clases es mucho mayor que var. dentro de las clases, lo cual es poco probable que ocurra si las medias son todas iguales.

# Numérica - Categórica

## Coeficiente de correlación de Kendall

- Coeficiente de Kendal b (considera empates):

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

$$n_0 = \binom{n}{2} \quad n_1 = \sum_i t_i(t_i - 1)/2 \quad n_2 = \sum_j u_j(u_j - 1)/2$$

$$n_c = \# \text{ pares concordantes} \quad n_d = \# \text{ pares discordantes}$$

$$t_i = \# \text{ valores empatados del grupo } i \text{ de } x \quad u_j = \# \text{ valores empatados del grupo } j \text{ de } y$$

- Test no paramétrico basado en rangos (estadísticas de orden)
- Mide la correlación de rangos
- Asume que la variable categórica tiene ordinalidad
- Costo computacional orden  $n^2$

# Categorica - Categorica

## Test Chi-cuadrado

- Test de Chi-Cuadrado (test de independencia de Pearson):

$$\chi = \sum_{i,j} \frac{O_{ij} - E_{ij}}{E_{ij}}$$

donde  $O_{ij}$  son la cantidad de observaciones pertenecientes a las categorías  $i, j$  de cada variable, y  $E_{ij}$  es el valor esperado observado si las variables fueran independientes.

- Se usa para rechazar la  $H_0$  que las variables son independientes.
  - $\chi \sim \chi^2_{r-1k, -1}$ ,  $r$  y  $k$  son la cantidad de factores de las variables de entrada y salida respectivamente.
- Criterio de Información mutua

# Categórica - Numérica

Es el caso menos frecuente, pero si ocurriera se puede tratar con los mismos criterios que Numérica categórica con los roles intercambiados.

# Comentarios finales

- Ventajas:
  - Son simples y suelen ser rápidos de computar,
- Desventajas:
  - Propensos a la sobre selección de variables,
  - Puede haber desconexión entre lo que el test reconoce como importante y lo que necesita el modelo.

# Bibliografía

- "Python Machine Learning Cookbook, practical solutions from preprocessing to deep learning", Albon, Cris. O'Reilly Media, Inc., 2018.
- "Feature Engineering and Selection, A Practical Approach for Predictive Models", Max Khun and Kjell Johnson. CRC Press, 2020.
- "Measures of Association How to Choose?" Harry Khamis, PhD. Journal of Diagnostic Medical Sonography May/June 2008 VOL. 24, NO. 3  
(<https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006>)
- "The Kendall Rank CorrelationCoefficient", Hervé Abdi  
(<https://personal.utdallas.edu/~herve/Abdi-KendallCorrelation2007-pretty.pdf>)
- W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," Journal of the American Statistical Association, vol. 47, no.260, pp. 583–621, 1952.