Temas Selectos de Física Computacional III: Introducción a la ciencia de datos. Clase 4: 25 Febrero 2025

Dr. Leonid Serkin

1 Introducción

En esta clase, estudiaremos el conjunto de datos Wisconsin Diagnostic Breast Cancer (WDBC) usando pandas y veremos que es la correlación lineal de Pearson.

2 Conjunto de datos de cáncer de mama

El conjunto de datos de cáncer de mama, accesible desde Breast Cancer es utilizado para predecir si un tumor es maligno o benigno basándose en 30 características numéricas obtenidas de imágenes digitalizadas de células. Estas características describen propiedades físicas como el radio, la textura, el perímetro, el área, y la suavidad de los núcleos celulares.

El conjunto de datos incluye un total de 569 muestras y está clasificado en dos clases:

- **Maligno**: Tumor canceroso (clase 0).
- **Benigno**: Tumor no canceroso (clase 1).

El conjunto incluye 30 características calculadas para cada imagen de una masa mamaria, derivadas de imágenes digitalizadas de aspiraciones con aguja fina del tejido mamario. Estas características describen propiedades de los núcleos celulares en la imagen.

- **Características medias (3-12)**: Representan el valor promedio de cada característica para todos los núcleos celulares en la imagen.
- **Error estándar (13-22)**: Representa el error estándar de la característica medida, proporcionando una estimación de la variación.
- **Peores características (23-32)**: Se calculan como la media de los tres valores más grandes de cada característica en una imagen.

Entre las principales características numéricas presentes en el conjunto de datos se encuentran:

- **radio medio**: Distancia media desde el centro hasta los puntos del perímetro.
- **textura media**: Desviación estándar de los valores de escala de gris.
- **perímetro medio**: Medida promedio del perímetro de las células.
- **área media**: Área promedio de las células.
- **suavidad media**: Variación local en la longitud de los radios.

Vamos a estudiarlo el conjunto usando pandas. Vamos a ver las primeras filas del DataFrame:

```
import pandas as pd
# Cargar el conjunto de datos desde el archivo CSV (sin encabezados)
df = pd.read_csv('wdbc.csv', header=None)
print(df.head())
print(df.shape)  # Ver número de filas y columnas
```

Puedes obtener un resumen estadístico usando:

```
print(df.describe())
print(df.info())
```

Para saber cuántos casos son benignos (B) y malignos (M), usamos la cantidad de veces que aparece cada valor en la segunda columna (Diagnosis):

```
print(df.columns)
print(df[1].value_counts())
```

Vamos a crear un un diccionario donde asignamos nombres solo a las primeras 6 columnas.

```
nombres_columnas = {
    0: "ID",
    1: "Diagnóstico",
    2: "Radio medio",
    3: "Textura media",
    4: "Perímetro medio",
    5: "Área media"
}
df.rename(columns=nombres_columnas, inplace=True)
print(df.head())
```

Ya para cambiar todos los nombres de las columnas:

```
nombres_columnas =  [ 'ID', .... ]
df.columns = column_names
```

Qué pasa si estamos interesados en cómo difiere la forma de una distribución entre las dos clases? Podemos simplemente agregar la columna `target` al DataFrame:

```
# Convertir la columna 'Diagnosis' en binaria (M = maligno, B = benigno)
df['target'] = df['Diagnóstico'].map({'M': 1, 'B': 0})
```

Y ahora eliminamos las columnas no necesarias (por ejemplo, 'ID' y 'Diagnóstico')?

```
df = df.drop(columns=['ID', 'Diagnóstico'])
print(df.iloc[:10, :4])
```

3 Exploración visual de datos

Vamos a dibujar los histogramas de las primeras 4 columnas:

```
import matplotlib.pyplot as plt
df.iloc[:, :4].hist(figsize=(12, 8), bins=20)
plt.show()
```

Vamos a graficar el conjunto de datos de cáncer de mama utilizando las características *radio medio* y *suavidad media*, coloreando los puntos en función de si el tumor es benigno o maligno.

```
plt.scatter(df['Radio medio'], df['Textura media'], c=df['target'], cmap='bwr')
plt.xlabel('Radio medio')
plt.ylabel('Textura media')
plt.show()
```

Luego, podemos usar la útil función groupby y graficar una estimación de densidad de kernel (kde):

```
df.groupby("target")["Radio medio"].plot(kind='kde', figsize=(10, 10))
plt.legend(['Maligno', 'Benigno'], loc='upper right')
plt.xlabel('Radio medio')
plt.show()
```

También podríamos generar un histograma que compara la distribución del radio medio entre tumores malignos y benignos en el conjunto de datos.

```
maligno = df[df['target'] == 1]  # Clase 'Maligno'
benigno = df[df['target'] == 0]  # Clase 'Benigno'

plt.figure(figsize=(10, 10))
```

```
plt.hist(maligno['Radio medio'], alpha=0.5, label='M', color='red', bins=30)
plt.hist(benigno['Radio medio'], alpha=0.5, label='B', color='blue', bins=30)

plt.xlabel('Radio medio')
plt.ylabel('Frecuencia')
plt.legend(loc='upper right')
plt.show()
```

Podemos analizar la matriz completa de dispersión para las primeras 4 características. Aqui cada celda en la matriz de dispersión representa un gráfico de dispersión (scatter plot) entre dos variables (una en el eje X y otra en el eje Y). La diagonal muestra histogramas (o gráficos de densidad) de cada variable individual. Las unidades de los ejes de cada gráfico son simplemente los valores de las características seleccionadas.

```
from pandas.plotting import scatter_matrix
scatter_matrix(df.iloc[:, 0:4], c=df['target'], alpha=0.8, figsize=(20, 20), s=20)
plt.show()
```

4 Correlaciones de Pearson y Spearman

La correlación utilizada por defecto en el método `df.corr()` de pandas es la correlación lineal de Pearson. Este coeficiente de correlación mide la relación lineal entre dos variables. El coeficiente de correlación de Pearson, r, tiene los siguientes valores:

- $r = 1$: Correlación lineal positiva perfecta.
- $r = 0$: No hay correlación lineal entre las variables.
- $r = -1$: Correlación lineal negativa perfecta.

La fórmula para calcular el coeficiente de correlación de Pearson entre dos variables X e Y es:

$$r_{xy} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}} \tag{1}$$

Ahora en caso de llamar el método `df.corr(method="spearman")`, se obtiene la correlación de rangos de Spearman. Este coeficiente mide la relación monótona entre dos variables, sin asumir que la relación sea estrictamente lineal. El coeficiente de correlación de Spearman, ρ, tiene los siguientes valores:

- $\rho = 1$: Relación monótona creciente perfecta.
- $\rho = 0$: No hay relación monótona entre las variables.

- $\rho = -1$: Relación monótona decreciente perfecta.

La fórmula para calcular el coeficiente de correlación de Spearman entre dos variables X e Y es:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

donde:

- d_i es la diferencia entre los rangos de cada par de valores X_i e Y_i.
- n es el número total de observaciones.

A diferencia de la correlación de Pearson, la correlación de Spearman es útil cuando la relación entre las variables es no lineal pero sigue una tendencia creciente o decreciente.

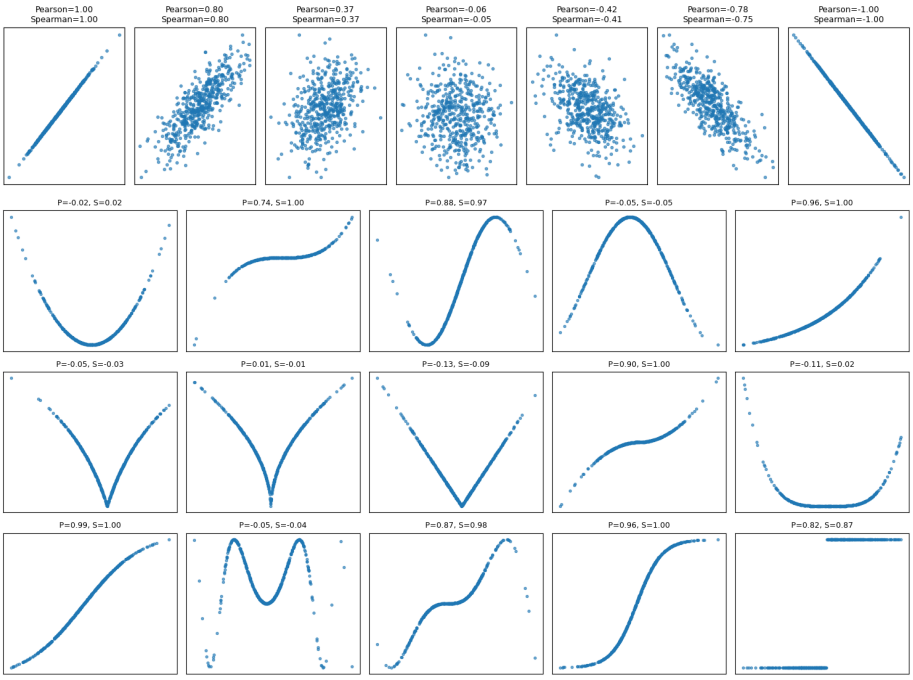

Figure 1: Comparación de valores de correlación.

Ahora usaremos seaborn: visualización estadística de datos para obtener las correlaciones lineales entre las características de entrada.

```
import seaborn as sns
plt.figure(figsize=(10, 10))
sns.heatmap(df.iloc[:, 0:4].corr(), annot=True, square=True, cmap='coolwarm')
plt.show()
```

Ahora analiza la matriz de correlación de Pearson y Spearman para todas las columnas, y extrae tus conclusiones:

1. Reducción de dimensionalidad: Si tuvieras que eliminar una de las variables entre radio, perímetro y área, ¿cuál elegirías y por qué?
2. Relación con el diagnóstico: Si quisieras construir un modelo de clasificación basado solo en dos variables, ¿cuáles escogerías y por qué?
3. Variables con baja correlación: Si una variable tiene una correlación baja con el target, eso significa que es irrelevante para el modelo de clasificación?
4. Comparación con Pearson: en qué casos podríamos encontrar diferencias significativas entre la correlación de Pearson y la de Spearman?