Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2433 Minería de Datos
Guía de Normalización

Normalización

La normalización generalmente se requiere cuando estamos utilizando atributos en diferentes escalas. Si no se normaliza, puede conducir a error en la efectividad de cálculo de distancia de atributos importantes.

Entre los diferentes método de normalización podemos mencionar los siguientes:

- Min-Max
- Z-score
- Z-score media desviación absoluta
- Decimal de escala

Normalización Min-Max

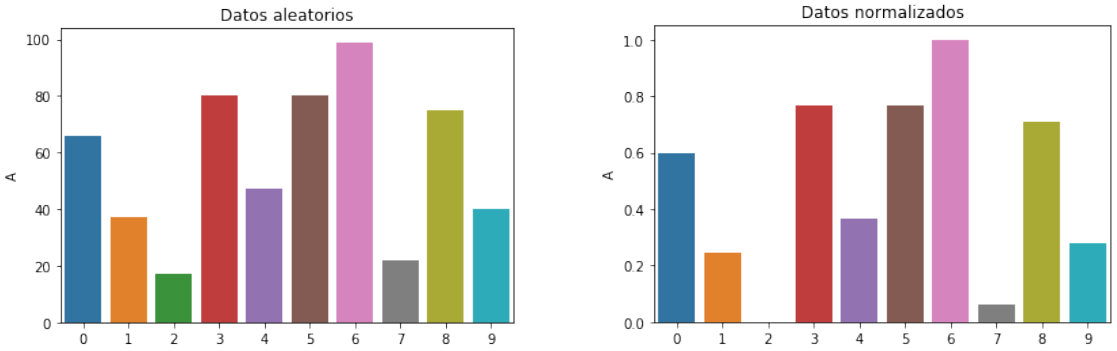
$$Z_i = \frac{X_i - min(X)}{max(X) - min(X)} \tag{1}$$

donde $X = (X_1, X_2, ..., X_n)$ y además X_i es el i^{esimo} término normalizado

Ejemplo

```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt

df = pd.DataFrame(np.random.randint(0,100,size=(10, 1)), columns=list('A'))

print(df)

norm = (df["A"]-min(df["A"]))/(max(df["A"])-min(df["A"]))

print(norm)
```

```
df_plot    = sns.barplot(data=df, y=df["A"], x=df.index.tolist())
plt.title("Datos aleatorios")
plt.show(df_plot)

plt.title("Datos normalizados")
norm_plot = sns.barplot(data=df, y=norm, x=df.index.tolist())
plt.show(norm_plot)
```



(a) Datos aleatorios (b) Datos normalizados

Figura 1: Comparación de magnitud de datos luego de normalizar

Normalización Z-score

La formula de normalización Z-score es la siguiente:

$$Z_i = \frac{X_i - \bar{X}}{S} \tag{2}$$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} \tag{3}$$

donde S es la desviación estándar, $\bar{X}$ el promedio