perplexity

資料剖析

一、資料概述

- 資料包含特徵（X）與標籤（y），通常為數值型資料。
- 需確認資料完整性，檢查是否有缺失值或異常值。

二、資料前處理

- 缺失值處理：可選擇刪除或填補（均值、中位數、眾數等）。
- 異常值檢測與處理：利用箱型圖（boxplot）或標準差方法。
- 特徵標準化：將特徵縮放至相同尺度（如 Z-score 標準化）。

三、資料視覺化

- 散佈圖（Scatter plot）：觀察特徵與標籤間的線性關係。
- 直方圖（Histogram）：檢視特徵分布情況。
- 箱型圖（Boxplot）：檢視資料分布及異常值。
- 熱力圖（Heatmap）：檢視特徵間相關性。

四、統計描述

- 計算特徵與標籤的平均值、標準差、最大值、最小值等。
- 計算相關係數（如皮爾森相關係數）以評估特徵與標籤的線性相關性。

五、資料分割

- 將資料分為訓練集與測試集，常見比例為 7:3 或 8:2。

六、範例程式碼

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

# 讀取資料
# df = pd.read_csv('data.csv')

# 缺失值檢查
print(df.isnull().sum())

# 異常值視覺化
sns.boxplot(data=df)
```

```
plt.show()

# 特徵與標籤散佈圖
sns.scatterplot(x='feature', y='target', data=df)
plt.show()

# 計算相關係數
corr = df.corr()
print(corr)

# 熱力圖
sns.heatmap(corr, annot=True)
plt.show()

# 資料分割
from sklearn.model_selection import train_test_split
X = df.drop('target', axis=1)
y = df['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

⁂