Bài thực hành

Phân loại dữ liệu với giải thuật k láng giềng (k nearest neighbors)

1. Ví dụ minh họa sử dụng giải thuật k láng giềng (k nearest neighbors)

```
# Nạp các gói thư viện cần thiết
import pandas as pd
from sklearn.neighbors import KNeighborsClassifier
import numpy as np

# Đọc dữ liệu iris từ UCI (https://archive.ics.uci.edu/ml/datasets/Iris)
# hoặc từ thư viện scikit-learn
# Tham khảo https://scikit-
learn.org/stable/auto_examples/datasets/plot_iris_dataset.html
from sklearn import datasets
from sklearn.model_selection import train_test_split
iris = datasets.load_iris()
columns=["Petal length","Petal Width","Sepal Length","Sepal Width"];
df = pd.DataFrame(iris.data, columns=columns)
y = iris.target
print(df.describe())

print("\n")
print("Kiem tra xem du lieu co bi thieu (NULL) khong?")
print(df.isnull().sum())

# Sử dụng nghi thức kiểm tra hold-out
# Chia dữ liệu ngẫu nhiên thành 2 tập dữ liệu con:
# training set và test set theo tỷ lệ 70/30
X_train, X_test, y_train, y_test = train_test_split(df, y, test_size=0.3)
#print(X_train.shape, y_train.shape)
#print(X_test.shape, y_test.shape)
#print(X_train.head())

# Xây dựng mô hình với k = 3
model = KNeighborsClassifier(n_neighbors=3)
model.fit(X_train, y_train)

# Dự đoán nhãn tập kiểm tra
prediction = model.predict(X_test)
#print(prediction)

# Tính độ chính xác
print("Do chinh xác cua mo hinh voi nghi thuc kiem tra hold-out: %.3f" %
model.score(X_test, y_test))
```

2. Ví dụ minh họa sử dụng giải thuật k láng giềng (k nearest neighbors) tuân theo nghi thức kiểm tra chéo 5-fold. Tài liệu khảo nghi thức kiểm tra chéo k-fold

(i) https://medium.com/datadriveninvestor/k-fold-cross-validation-6b8518070833

(ii) https://scikit-learn.org/stable/modules/cross_validation.html

```
# Sử dụng nghi thức kiểm tra chéo k-fold

from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score

# Thực hiện nghi thức kiểm tra 5 fold
nFold = 5;
model = KNeighborsClassifier(n_neighbors=3)
scores = cross_val_score(model, df, y, cv=nFold)
print("Do chinh xac cua mo hinh voi nghi thuc kiem tra %d-fold %.3f" %
(nFold, np.mean(scores)))
```

3. Hãy viết chương trình phân loại các tập dữ liệu sau với giải thuật k láng giềng (k nearest neighbors): Breast Cancer Wisconsin, Wine, Optical recognition of handwritten digits dataset. Thay đổi k=1, 2, 3, 4, 5. Ghi nhận kết quả. Khi chạy cần tuân theo nghi thức kiểm tra chéo k-fold. Chú ý các tập dữ liệu có thể tìm thấy trong gói thư viện scikit.