

Bài thực hành

Phân loại dữ liệu với máy học vector hỗ trợ (Support Vector Machine - SVM)

1. Ví dụ minh họa sử dụng giải thuật máy học vector hỗ trợ

```
# Nạp các gói thư viện cần thiết
import pandas as pd
from sklearn import svm

# 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("Kiểm tra xem dữ liệu có bị thiếu (NULL) không?")
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)

# Xây dựng mô hình svm sử dụng hàm nhân (kernel) là RBF
# SVC là viết tắt của từ Support Vector Classification
model = svm.SVC(kernel='rbf')
model.fit(X_train, y_train)

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

# Tính độ chính xác
print("Độ chính xác của mô hình với nghi thức kiểm tra hold-out: %.3f" %
      model.score(X_test, y_test))
```

2. Anh/chị hãy hiệu chỉnh ví dụ bên trên để chương trình 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

3. Hãy viết chương trình phân loại các tập dữ liệu sau với giải thuật máy học vector hỗ trợ (Support Vector Machine - SVM): Breast Cancer Wisconsin, Wine, Optical recognition of handwritten digits dataset. Khi chạy cần tuân theo nghi thức kiểm tra chéo k-fold, nhớ thay đổi các kiểu hàm nhân, tham số hàm nhân. Ghi nhận kết quả.