Tut 7: Sử dụng Cross-validation

- **Tác giả: Vũ Đức Duy.**
- **Nhóm nghiên cứu về AI Đại học Bách Khoa Thành phố Hồ Chí Minh.**

1. Cross-validation

Khi thực hiện một bài toán machine learning, ta thường chia tập dữ liệu thành 2 phần, gọi là training set và test set. Tập đầu tiên dùng để xây dựng model cho bài toán, tập thứ hai dùng để kiểm tra độ hiệu quả của model sinh ra từ tập đầu tiên. Tuy nhiên, trong thực tế khi chia như vậy vẫn có thể xảy ra hiện tượng overfitting, để giải quyết vấn đề này thì người ta đề xuất một số cách mới để chia tập dữ liệu, trong đó có cross-validation (gọi tắt là CV).

Trong cách này, gọi là k-fold CV, tập dữ liệu được chia thành k tập dữ liệu nhỏ, mỗi tập nhỏ này gọi là 1 fold. Sau đó thực hiện k vòng lặp, mỗi vòng lặp gồm các bước sau:

- Model được train bằng cách sử dụng k – 1 folds như là training data.
- Model sau khi train sẽ được validate trên fold còn lại trong tập dữ liệu.

Độ chính xác của model được tính bằng cách lấy độ chính xác trung bình của toàn bộ vòng lặp.

2. Cross-validation sử dụng thư viện sklearn

Thư viện sklearn đã có sẵn hàm để ta thực hiện cross-validation, tên là cross_val_score. Các tham số quan trọng trong hàm cross_val_score bao gồm:

- estimator: một trong các estimator của sklearn có hiện thực hàm fit. Vd: Kmeans, KNN, SVM,…
- X: dữ liệu để “fit” vào model

- y: label của tập dữ liệu ứng với X
- cv: số lượng fold cần chia trên tập dữ liệu ban đầu, hoặc có thể đưa vào 1 cross-validation generator object hoặc 1 iterable.

Đoạn code dưới đây thực hiện cross-validation trên tập dữ liệu Iris sử dụng linear SVM:

```
>>> from sklearn.model_selection import cross_val_score
>>> from sklearn import datasets
>>> from sklearn import svm
>>> iris = datasets.load_iris()
>>> clf = svm.SVC(kernel='linear', C=1)
>>> scores = cross_val_score(clf, iris.data, iris.target, cv=5)
>>> scores
array([ 0.96...,  1.  ...,  0.96...,  0.96...,  1.        ])
```

Scores chính là list chứa accuracy của model trong mỗi vòng lặp, với số vòng lặp là 5 (tham số cv). Sau khi có được Scores ta có thể tính được accuracy trung bình và độ lệch chuẩn:

```
>>> print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
Accuracy: 0.98 (+/- 0.03)
```

3. Tài liệu tham khảo

Để tìm hiểu kỹ hơn về cross-validation và các tham số nâng cao, các bạn tham khảo 2 đường link dưới đây:

http://scikit-learn.org/stable/modules/cross_validation.html
http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score