07/25 DM 下午場

洪子軒

Sent: Monday, July 25, 2016 4:19 PM
To: 洪子軒

【分類】

antecedent 條件
consequent 結果
Coverage 規則覆蓋率（support）
Strength 規則強度（confidence）

KNN分類器
http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

決策樹分類器顯示
http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
http://chrisstrelioff.ws/sandbox/2015/06/08/decision_trees_in_python_with_scikit_learn_and_pandas.html

分類範例，含 score （平均正確率）
http://scikit-learn.org/stable/auto_examples/exercises/digits_classification_exercise.html

```
from sklearn import datasets, neighbors, linear_model

digits = datasets.load_digits()
X_digits = digits.data
y_digits = digits.target

n_samples = len(X_digits)

X_train = X_digits[:.9 * n_samples]
y_train = y_digits[:.9 * n_samples]
X_test = X_digits[.9 * n_samples:]
y_test = y_digits[.9 * n_samples:]

knn = neighbors.KNeighborsClassifier()
logistic = linear_model.LogisticRegression()

print('KNN score: %f' % knn.fit(X_train, y_train).score(X_test, y_test))
print('LogisticRegression score: %f'
      % logistic.fit(X_train, y_train).score(X_test, y_test))
```

分訓練/測試資料工具 cross validation
http://scikit-learn.org/stable/modules/cross_validation.html

```
>>> import numpy as np
>>> from sklearn import cross_validation
>>> from sklearn import datasets
>>> from sklearn import svm

>>> iris = datasets.load_iris()
>>> iris.data.shape, iris.target.shape
((150, 4), (150,))
```

```
>>> X_train, X_test, y_train, y_test = cross_validation.train_test_split(
...     iris.data, iris.target, test_size=0.4, random_state=0)

>>> X_train.shape, y_train.shape
((90, 4), (90,))
>>> X_test.shape, y_test.shape
((60, 4), (60,))

>>> clf = svm.SVC(kernel='linear', C=1).fit(X_train, y_train)
>>> clf.score(X_test, y_test)
0.96...
```

k-fold CV

```
>>> clf = svm.SVC(kernel='linear', C=1)
>>> scores = cross_validation.cross_val_score(
...    clf, iris.data, iris.target, cv=5)
...
>>> scores
array([ 0.96...,  1.  ...,  0.96...,  0.96...,  1.        ])
```

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

```
>>> predicted = cross_validation.cross_val_predict(clf, iris.data,
...                                                iris.target, cv=10)
>>> metrics.accuracy_score(iris.target, predicted)
0.966...
```

分類器混淆矩陣
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
https://en.wikipedia.org/wiki/Confusion_matrix

評估指標
http://scikit-learn.org/stable/modules/classes.html

		Predicted		
		Cat	**Dog**	**Rabbit**
Actual class	**Cat**	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

F-measure 越大越好：把 Precision（a/(a+c) 精確率哪些是TP ） 和 Recall（a/(a+b) 回復率） 合在一起

TP 正陽性，TN 正陰性
FP 偽陽性，FN 偽陰性

※ROC Curve （TP-FP rate 線）
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
http://scikit-learn.org/stable/auto_examples/ensemble/plot_feature_transformation.htm

--
洪子軒 Tzu-Hsuan Hung

中華電信研究院 巨量資料所
TEL: (03)-4245128
Email: Lucas@cht.com.tw
32661桃園市楊梅區電研路99號
--