07/29 DM 上午場

洪子軒

Sent: Friday, July 29, 2016 12:04 PM
To: 洪子軒

【分類】

http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

◎ RF 隨機森林集成分類器
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

n_estimators：The number of trees in the forest.
max_features：The number of features to consider when looking for the best split

◎ 完全 random 的集成分類器（在屬性選取時，用 random）
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

◎ 離散屬性編碼
http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
http://stackoverflow.com/questions/15021521/how-to-encode-a-categorical-variable-in-sklearn

```
>>> from sklearn.preprocessing import OneHotEncoder
>>> enc = OneHotEncoder()
>>> enc.fit([[0, 0, 3], [1, 1, 0], [0, 2, 1], [1, 0, 2]])
OneHotEncoder(categorical_features='all', dtype=<... 'float'>,
       handle_unknown='error', n_values='auto', sparse=True)
>>> enc.n_values_
array([2, 3, 4])
>>> enc.feature_indices_
array([0, 2, 5, 9])
>>> enc.transform([[0, 1, 1]]).toarray()
array([[ 1.,  0.,  0.,  1.,  0.,  0.,  1.,  0.,  0.]])
>>> enc.transform([[0, 1, 1]])
```

◎ 屬性選取

方法：移除值域分佈太窄的屬性（事前）
http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html

```
>>> from sklearn.feature_selection import VarianceThreshold
>>> X = [[0, 0, 1], [0, 1, 0], [1, 0, 0], [0, 1, 1], [0, 1, 0], [0, 1, 1]]
>>> sel = VarianceThreshold(threshold=(.8 * (1 - .8)))
>>> sel.fit_transform(X)
array([[0, 1],
       [1, 0],
       [0, 0],
       [1, 1],
       [1, 0],
       [1, 1]])
```

```
get_support(indices=False)
```

Get a mask, or integer index, of the features selected

方法： 連續屬性 / F檢定（變異數相除）→ T檢定（平均數相減）

http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html
http://www.math.nsysu.edu.tw/~lomn/homepage/R/R_testing.htm

```
>>> from sklearn import svm
>>> from sklearn.datasets import samples_generator
>>> from sklearn.feature_selection import SelectKBest
>>> from sklearn.feature_selection import f_regression
>>> from sklearn.pipeline import Pipeline
>>> # generate some data to play with
>>> X, y = samples_generator.make_classification(
...     n_informative=5, n_redundant=0, random_state=42)
>>> # ANOVA SVM-C
>>> anova_filter = SelectKBest(f_regression, k=5)
>>> clf = svm.SVC(kernel='linear')
>>> anova_svm = Pipeline([('anova', anova_filter), ('svc', clf)])
>>> # You can set the parameters using the names issued
>>> # For instance, fit using a k of 10 in the SelectKBest
>>> # and a parameter 'C' of the svm
>>> anova_svm.set_params(anova__k=10, svc__C=.1).fit(X, y)
...
Pipeline(steps=[...])
>>> prediction = anova_svm.predict(X)
>>> anova_svm.score(X, y)
0.77...
>>> # getting the selected features chosen by anova_filter
>>> anova_svm.named_steps['anova'].get_support()
...
array([ True,  True,  True, False, False,  True, False,  True,  True, True,
       False, False,  True, False,  True, False, False, False, False,
       True], dtype=bool)
```

方法：離散屬性 / chi2 卡方檢定 → 進一步看 Proportion 檢定
http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.feature_selection import SelectKBest
>>> from sklearn.feature_selection import chi2
>>> iris = load_iris()
>>> X, y = iris.data, iris.target
>>> X.shape
(150, 4)
>>> X_new = SelectKBest(chi2, k=2).fit_transform(X, y)
>>> X_new.shape
(150, 2)
```

＞Methods for Comparing Classifiers

單一分類器的正確率（作超過30次）p 分佈的信賴區間要有平均值 p 和標準差 p(1-p)/N 和 1-α 的信心水準
兩兩比較分類器的正確率：實驗次數>30，d=|e1-e2| 比較是否 Interval contains 0 ※投影片 DMBI5

＞非線性迴歸

・迴歸樹 RT- Regression Tree：決策樹變體 ※
・隨機森林 RF- Random Forest：每棵樹加總取平均 ※
・最近鄰居法 KNN- K-nearest neighbor：最近鄰居預測值加總取平均
・支援向量法 SVR- Support Vector Regression
・類神經網路 ANN

・多元適應雲形迴歸 MARS- Multivariate Adaptive Regression Splines ※

※可以判定變數重要性

◎ 迴歸績效指標：
・MSE 誤差平方和 Σ(Actual - Forecast)^2 / (n-1)
・MAD 絕對值偏差 Σ|Actual - Forecast| / n
・MAPE 誤差百分比 (Σ|(Actual - Forecast) / Actual | / n)* 100%

ps 也要和傳統線性迴歸比較誤差

◎ MARS ：片斷線性基底函數 basis

--
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
--