07/28 DM 上午場

洪子軒

Sent: Thursday, July 28, 2016 1:27 PM
To: 洪子軒

【分類】

＞Ensemble 集成學習法（三個臭皮匠，勝過一個諸葛亮）：如 Random Forest、AdaBoost

模型效果從好到差的排序通常依次為：隨機森林>Boosting > Bagging > 單棵決策樹

多數決（原始訓練資料集 → 產生出多個不同的資料集 ※ → 個別建分類器 → 結合分類器）使錯誤率大幅下降

※生成方法：Bagging（bootstrap aggregating 根據 uniform distribution 均勻分配重複產生）/ Boosting（動態調整每個樣本出現機率，越容易分錯的sample，出現機率高，使其容易被學習到）

https://read01.com/kRzEkQ.html
https://read01.com/2zBoA3.html
https://read01.com/QAneJm.html

Bagging 每個樣本都有 1-(1 - 1/n)^n 被選中的機率（當實驗回合 n 夠大，代表每個樣本都有機會被選中）
Boosting 每個樣本會調整不同的權重

→要選哪個，要看樣本的結構

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html

◎RandomForest
・ntree 樹的數目（看問題複雜度）
・mtry / max_features 要考慮用來分割的 X 變數數目（一般是建議屬性數開根號）

MDS 降維方式
http://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html
http://scikit-learn.org/stable/auto_examples/ensemble/plot_feature_transformation.html

◎樣本平均度
通常情況：讓不同 class 下的樣本數盡可能平均
特殊情況：有些業務不能這樣看（醫療診斷 or 正例數據難以搜集…等等）

＞類神經網路 ANN：可以分類 / 迴歸

http://scikit-learn.org/dev/modules/neural_networks_supervised.html

Class **`MLPClassifier`** implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation.

Error Back propagation algorithm：誤差修正項會往回傳（誤差對權重微分=0），使權重改變調整

```
>>> from sklearn.neural_network import MLPClassifier
>>> X = [[0., 0.], [1., 1.]]
>>> y = [[0, 1], [1, 1]]
>>> clf = MLPClassifier(algorithm='l-bfgs', alpha=1e-5, hidden_layer_sizes=(15,), rando
m_state=1)
>>> clf.fit(X, y)
MLPClassifier(activation='relu', algorithm='l-bfgs', alpha=1e-05,
       batch_size='auto', beta_1=0.9, beta_2=0.999, early_stopping=False,
       epsilon=1e-08, hidden_layer_sizes=(15,), learning_rate='constant',
       learning_rate_init=0.001, max_iter=200, momentum=0.9,
       nesterovs_momentum=True, power_t=0.5, random_state=1, shuffle=True,
       tol=0.0001, validation_fraction=0.1, verbose=False,
       warm_start=False)
>>> clf.predict([1., 2.])
array([[1, 1]])
>>> clf.predict([0., 0.])
array([[0, 1]])
```

--
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
--