07/28 DM 下午場

洪子軒

Sent: Thursday, July 28, 2016 3:56 PM
To: 洪子軒

【分類】

converge 收斂

＞SVM 支撐向量機 support vector machine
起源：線性分割，一筆劃定出楚河漢界（羅吉斯）
statistical learning 透過升維，轉換到高維度特微空間（Feature Space）的超平面（hyperplane），可以用一個平面分開資料成兩群

http://www.statsoft.com/Textbook/Support-Vector-Machines
https://cg2010studio.com/2012/05/20/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%A9%9F%E5%99%A8-support-vector-machine/

http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

常見用來升維的三種函數 Kernel function：
0. Linear：線性（原始情形）
1. Polynomial：多項式
2. Radial Basis Function（RBF，常用）：高斯函數，參數包含「gamma γ」、「cost」成本項懲罰 penalty（cost 越大越 hard）
3. Sigmoid（Hyberbolic）：sign(w^T*Xi - b) ，w（法向量） 和 X 內積 dot product 得到純量，b 為截距

※elegant property：高維度內積結果，相等於原本的點低維度內積後，代入函數

◎如何定義上述函數最佳解？
→ margin （d+ / d-） 越寬越好（最小化 W 和 x 內積）

・Hard margin：過度學習
・Soft margin：包含少量錯誤（誤差成本項），但分類平整

◎少數樣本對邊界才有決定性（Support Vectors，點在邊界上 0<=α<=C 或緩衝區 α=C 內，才可以決定 w 值）
邊界：w・Xi - b = -1 左 / +1 右
中線：w・Xi - b = 0
求解（w，b）→之後用來分類新的樣本，可算出 Y>1（右） 或 Y<-1（左）

※屬性選取
http://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection

```
anova_filter = SelectKBest(f_regression, k=5)
```

※ 參數優化
http://scikit-learn.org/stable/modules/grid_search.html

```
clf.best_params_
```

```
clf.best_estimator_
```

http://scikit-learn.org/stable/auto_examples/model_selection/grid_search_digits.html

＞總整理
KNN 數值
NBC（數值類別適用）
LR（屬性重要性，數值類別適用）
C5.0（屬性重要性）
ANN（數值）
CART（屬性重要性，數值類別適用）

RF bagging boosting（集成式，數值類別適用，屬性重要性）
SVM（數值，執行時間久，大多數時候執行比較好）

--
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
--