مشروع نهاية الفصل 5

مقرر Machine Learning — Support Vector Machines — مشروع الـ SVM

الهدف: تبني SVM كامل: kernel، scaling، وضبط C وgamma بالـ cross-validation.

الداتا: sklearn.datasets.load_breast_cancer (٥٦٩ عينة، ٣٠ خاصية). مصدرها UCI. مدمجة في scikit-learn.

الأسئلة العملية

1 جهّز (Prepare)

- حمّل الداتا وقسّمها ٧٠/٣٠ بـ stratify و random_state=0.

2 Baseline (Baseline)

- دقة DummyClassifier.

3 بدون scaling (No scaling)

- درّب SVC() بالقيم الافتراضية بدون scaling واطبع الدقة.

4 مع scaling (With scaling)

- حط StandardScaler في Pipeline وقارن.

5 اضبط (Tune)

- GridSearchCV على C = 1, 10, 100 و gamma = 0.001, 0.01 بـ cv=5.
- اطبع أحسن القيم ودقة الـ test.

بونس

- بونس: جرّب 'kernel='linear وقارن عدد الـ support vectors.
- بونس: اطبع confusion matrix للنموذج النهائي.

أسئلة تفكير (اكتب إجابة في سطرين)

1. ليه الـ scaling مهم للـ SVM؟
2. إيه الخطر لو gamma كبير جداً؟

ماذا تسلّم؟

- ملف الكود (أو notebook) بيشتغل من الأول للآخر.
- فقرة قصيرة (٥ أسطر) بتلخّص النتيجة وإيه اللي اتعلمته.

الحل النموذجي

جرّب بنفسك الأول، وبعدين قارن. الأرقام الحقيقية في النتيجة تحت.

الكود

```
# dataset ships inside scikit-learn (no download needed)
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.dummy import DummyClassifier

X, y = load_breast_cancer(return_X_y=True)
X_tr, X_te, y_tr, y_te = train_test_split(X, y, test_size=0.3, random_state=0,
stratify=y)
print('baseline:', round(DummyClassifier(strategy='most_frequent').fit(X_tr,
y_tr).score(X_te, y_te), 3))
print('no scaling:', round(SVC().fit(X_tr, y_tr).score(X_te, y_te), 3))
pipe = make_pipeline(StandardScaler(), SVC())
print('with scaling:', round(pipe.fit(X_tr, y_tr).score(X_te, y_te), 3))
grid = GridSearchCV(pipe, {'svc__C': [1, 10, 100], 'svc__gamma': [0.001, 0.01]},
cv=5).fit(X_tr, y_tr)
print('best:', grid.best_params_)
print('tuned test accuracy:', round(grid.score(X_te, y_te), 3))
```

النتيجة

```
baseline: 0.626
no scaling: 0.906
with scaling: 0.953
best: {'svc__C': 100, 'svc__gamma': 0.001}
tuned test accuracy: 0.953
```

ملاحظات على الإجابات

- الـ SVM بيعتمد على المسافات فالـ features اللي أرقامها كبيرة بتسيطر من غير scaling.
- gamma كبير جداً يعمل حدود متعرّجة ومحلية يعني overfitting.
- اضبط الـ hyperparameters على الـ training بالـ cross-validation، والـ test مرة واحدة في الآخر.

Credits: topic follows the matching chapter of Aurélien Géron, Hands-On Machine Learning (3rd ed.). Questions, code and explanations are original.