مشروع نهاية الفصل 4

مقرر Machine Learning — Training Models — مشروع تدريب النماذج

الهدف: تقارن Linear و Ridge و Lasso و Polynomial على نفس الداتا وتشوف أثر الـ regularization.

الداتا: sklearn.datasets.load_diabetes (٤٤٢ مريض، ١٠ خصائص). Efron وآخرون ٢٠٠٤. مدمجة في scikit-learn.

الأسئلة العملية

1 جهّز (Prepare)

- حمّل الداتا وقسّمها ٨٠/٢٠ بـ random_state=0.

2 Linear (Linear)

- درّب LinearRegression واطبع R2 للـ training والـ test.

3 Ridge و Lasso (Ridge and Lasso)

- جرّب Ridge(alpha=10) و Lasso(alpha=1) بعد StandardScaler واطبع R2 للـ test.
- كام وزن بقى صفر في الـ Lasso؟

4 Polynomial (Polynomial)

- جرّب PolynomialFeatures(2) مع Ridge(alpha=10) وقارن الـ train والـ test.

5 اختار (Choose)

- اختار نموذج واكتب ليه، بناءً على الـ test R2 والفجوة بين train و test.

بونس

- بونس: استخدم RidgeCV للبحث عن أحسن alpha.
- بونس: جرّب SGDRegressor وقارنه.

أسئلة تفكير (اكتب إجابة في سطرين)

1. إيه علامة الـ overfitting في الأرقام؟
2. ليه الـ Lasso بيختار features؟

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)
import numpy as np
from sklearn.datasets import load_diabetes
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.linear_model import LinearRegression, Ridge, Lasso

X, y = load_diabetes(return_X_y=True)
X_tr, X_te, y_tr, y_te = train_test_split(X, y, test_size=0.2, random_state=0)
models = {
 'linear': LinearRegression(),
 'ridge': make_pipeline(StandardScaler(), Ridge(alpha=10)),
 'lasso': make_pipeline(StandardScaler(), Lasso(alpha=1)),
 'poly2+ridge': make_pipeline(StandardScaler(), PolynomialFeatures(2), Ridge(alpha=10)),
}
for n, m in models.items():
    m.fit(X_tr, y_tr)
    print(n.ljust(12), 'train R2', round(m.score(X_tr, y_tr), 3), '| test R2',
round(m.score(X_te, y_te), 3))
print('lasso weights equal to 0:', int((models['lasso'][-1].coef_ == 0).sum()))
```

النتيجة

```
linear       train R2 0.554 | test R2 0.332
ridge        train R2 0.552 | test R2 0.331
lasso        train R2 0.55 | test R2 0.334
poly2+ridge  train R2 0.624 | test R2 0.295
lasso weights equal to 0: 2
```

ملاحظات على الإجابات

- الفجوة الكبيرة بين R2 الـ training والـ test علامة overfitting.
- الـ Lasso بيعاقب بمجموع القيم المطلقة فبيدفع بعض الأوزان لصفر، يعني بيشيل features.
- النموذج الأبسط اللي أداؤه قريب هو غالباً الاختيار الأحسن.

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