REGRESJA LINIOWA

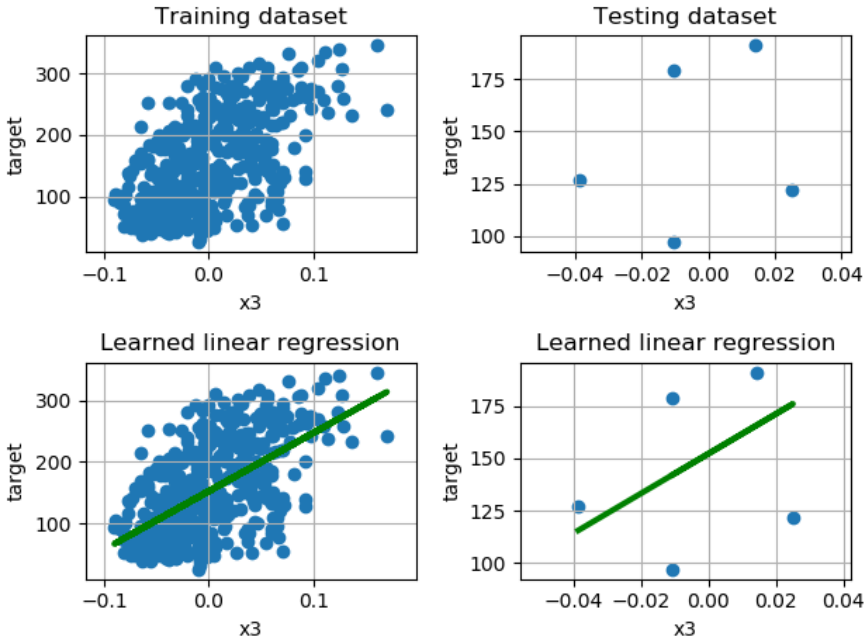
REGRESJA LINIOWA I LOGISTYCZNA

ZADANIA

Korzystając z biblioteki **sklearn** (`http://scikit-learn.org`), wykonaj następujące zadania:

1. Załaduj dane „`diabetes`” i wyłuskaj z nich tylko trzecią cechę ($x3$)
   ```
   from sklearn import datasets
   diabetes = datasets.load_diabetes()
   ```
2. Zwizualizuj dane z wykorzystaniem wykresu punktowego $f(x)$: $x3 \rightarrow y \in \Re$
3. Podziel dane na zbiór uczący i testowy
4. Wytrenuj model regresji liniowej $h_\theta(x)$ na danych uczących
   ```
   from sklearn import linear_model

   regr = linear_model.LinearRegression()
   regr.fit(diabetes_X_train, diabetes_y_train)
   diabetes_y_pred = regr.predict(diabetes_X_test)
   ```
5. Zwizualizuj wytrenowaną hipotezę na zbiorze uczącym i testowym



6. Wytrenuj wieloraką regresją liniową (wiele zmiennych objaśniających) na pełnych danych uczących.

7. Jak można ocenić „jakość” wytrenowanej hipotezy na danych wielowymiarowych?

   a. błąd średniokwadratowy na zbiorze testującym
   b. współczynnik determinacji[1] R^2:

$$R^2 = \frac{\sum_{i=1}^{test_{size}} (\hat{y}_i - \overline{y})}{\sum_{i=1}^{test_{size}} (y_i - \overline{y})} \qquad \overline{y} = \frac{\sum_{i=1}^{test_{size}} y_i}{test_{size}} \qquad \hat{y}_i = h_\theta(x_i)$$

```
from sklearn.metrics import mean_squared_error, r2_score

mean_squared_error(real_values, predicts)
r2_score(diabetes_all_features_target_test, predicts)
```

8. Wytrenuj binarny[2] klasyfikator zbudowany na modelu regresji logistycznej $h_\theta(x)$
   a. załaduj dane opisujące kliniczne przypadki rozpoznania raka łagodny/złośliwy
      `load_breast_cancer()`
   b. w razie potrzeby ustandaryzuj dane
      `StandardScaler()`
   c. podziel dane na zbiór uczący i testujący
   d. przetestuj działanie klasyfikatora na danych testowych

```
classifier = linear_model.SGDClassifier(loss='log')

classifier.fit(data_train, target_train)
predict_probabilites=classifier.predict_proba(data_test)
```

9. Jak można ocenić jakoś klasyfikacji binarnej?
   a. Macierz konfuzji (ang. confusion matrix)

```
from sklearn.metrics import confusion_matrix
confusion_matrix(target_test,
classifier.predict(data_test))
```

   b. Dokładność

```
from sklearn.metrics import accuracy_score
accuracy_score(target_test, classifier.predict(data_test))
```

   c. Współczynnik f1

```
from sklearn.metrics import f1_score
f1_score(target_test, classifier.predict(data_test))
```

[1] https://pl.wikipedia.org/wiki/Wsp%C3%B3%C5%82czynnik_determinacji
[2] klasyfikacja pomiędzy dwoma klasami, $y \in \{0,1\}$