Linear Regression from scratch

The goal of this exercise is to implement the linear regression algorithm. The dataset is about predicting salary given gpa and years of experience. The steps to implement are as follows.

1. Read the data from a file (gpa_year_experience.csv)
2. Scale the attributes
3. Compute the error at each iteration and save the error values in vector
4. Plot the error vector as a curve in the end
5. Predict a new instance.
6. Compare with SGDRegressor
7. Create polynomial features and predict new instance

In [ ]:

```
# import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# ignore warnings
import warnings
warnings.filterwarnings('ignore')
```

In [ ]:

```
# load data and show first 5 rows
data = pd.read_csv('https://raw.githubusercontent.com/thomouvic/SENG474/main/data/g
data.head()
```

Out[ ]:

	gpa	years_of_experience	salary
0	70	1.0	50
1	80	2.0	55
2	65	2.0	45
3	70	2.5	60
4	65	2.7	58

In [ ]:

```
# prepare data, split columns into X and y
# X should be a numpy array of shape (m, n), use .values to convert from dataframe
# y should be a numpy array of shape (m,), use .values to convert from dataframe to

X = data[['gpa', 'years_of_experience']].values
y = data['salary'].values
```

In [ ]:

```
# extract m and n from X using X.shape[0] to get m and X.shape[1] to get n
m = X.shape[0]
```

```
n = X.shape[1]
m,n
```

Out[ ]:
```
(25, 2)
```

In [ ]:
```
# y should be a numpy array of shape (m, 1), use reshape(m, 1) to reshape y from (m
y = y.reshape(m, 1)
y
```

Out[ ]:
```
array([[50],
       [55],
       [45],
       [60],
       [58],
       [60],
       [65],
       [67],
       [55],
       [60],
       [65],
       [70],
       [78],
       [75],
       [78],
       [70],
       [80],
       [82],
       [75],
       [85],
       [80],
       [82],
       [85],
       [90],
       [85]])
```

In [ ]:
```
# normalize X using min-max scaler (sklearn.preprocessing.MinMaxScaler)
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)
X_normalized.shape
```

Out[ ]:
```
(25, 2)
```

In [ ]:
```
# add dummy feature to X using scikit-learn dummy feature (sklearn.preprocessing.ad
from sklearn.preprocessing import add_dummy_feature
#X = add_dummy_feature(X)
X_b = add_dummy_feature(X_normalized)
X_b
```

Out[ ]:

```
array([[1.     , 0.3125 , 0.     ],
       [1.     , 0.625  , 0.125  ],
       [1.     , 0.15625, 0.125  ],
       [1.     , 0.3125 , 0.1875 ],
       [1.     , 0.15625, 0.2125 ],
       [1.     , 0.625  , 0.25   ],
       [1.     , 0.9375 , 0.25   ],
       [1.     , 1.     , 0.275  ],
       [1.     , 0.     , 0.3125 ],
       [1.     , 0.3125 , 0.3375 ],
       [1.     , 0.5    , 0.375  ],
       [1.     , 0.78125, 0.4375 ],
       [1.     , 0.625  , 0.5    ],
       [1.     , 0.     , 0.5625 ],
       [1.     , 0.125  , 0.6    ],
       [1.     , 0.     , 0.625  ],
       [1.     , 0.84375, 0.625  ],
       [1.     , 0.9375 , 0.6875 ],
       [1.     , 0.46875, 0.75   ],
       [1.     , 0.625  , 0.75   ],
       [1.     , 0.46875, 0.8125 ],
       [1.     , 0.3125 , 0.875  ],
       [1.     , 0.625  , 0.9375 ],
       [1.     , 0.9375 , 0.9625 ],
       [1.     , 0.78125, 1.     ]])
```

In [ ]:

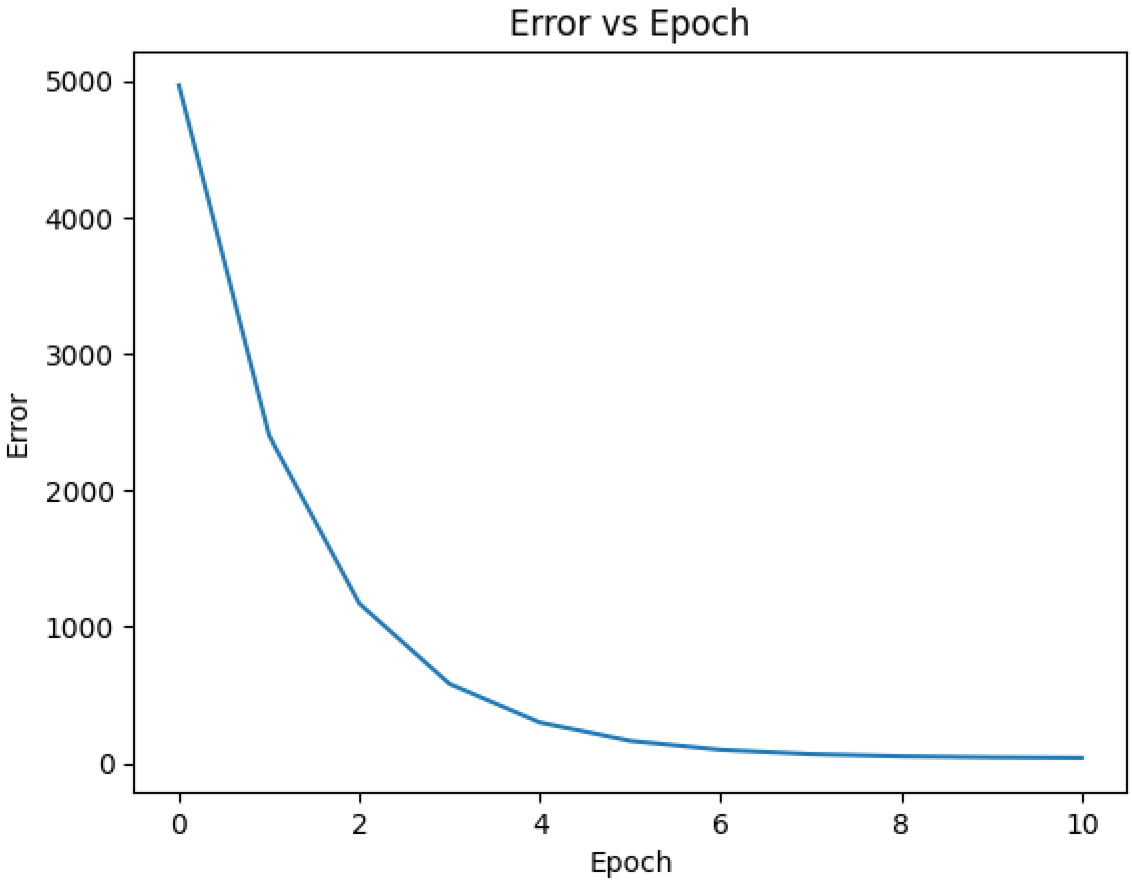
```python
# print shapes of X and y
# X should be (m, n+1) and y should be (m, 1)
print(X_b.shape, y.shape)
```

```
(25, 3) (25, 1)
```

In [ ]:

```python
eta = 0.1 # learning rate
n_epochs = 10
np.random.seed(42) # set random seed to 42 for reproducibility

# create theta, of shape (n+1, 1) and initialize it to random values using np.rando
theta = np.random.randn(n+1, 1)

E = [] # list to store errors at each epoch

# compute error for initial theta and append to E
E.append(np.mean((X_b.dot(theta) - y) ** 2))
# loop over n_epochs
# for each epoch: compute gradients, update theta, compute error, append error to E
for epoch in range(n_epochs):
    gradients = 2/m * X_b.T.dot(X_b.dot(theta) - y)
    theta = theta - eta * gradients
    E.append(np.mean((X_b.dot(theta) - y) ** 2))

# plot error vs epoch
plt.plot(E)
plt.xlabel('Epoch')
plt.ylabel('Error')
plt.title('Error vs Epoch')
plt.show()
```

```
# print final theta
print(theta)
```



```
[[44.69694899]
 [21.26803414]
 [25.80208124]]
```

In [ ]:

```
# let's predict the salary for a person who has gpa=70 and years_of_experience=3.
# create a numpy array x of shape (1, 2) with these values
# scale features using the same scaler we used earlier
# insert dummy feature using dummy feature function
# Predict salary of x
x = np.array([[70,3]])
x = scaler.transform(x)
x_b = add_dummy_feature(x)
x_pred = x_b.dot(theta)
x_pred
```

Out[ ]:

```
array([[57.79372996]])
```

In [ ]:

```
# Let's compare with scikit-learn's SGDRegressor
# use SGDRegressor from scikit-learn to fit the data
# use max_iter=1000, eta0=0.1, random_state=42
from sklearn.linear_model import SGDRegressor
sgd_reg = SGDRegressor(max_iter=1000, eta0=0.1, random_state=42)
sgd_reg.fit(X_normalized, y)
```

Out[ ]:

```
▾ SGDRegressor
SGDRegressor(eta0=0.1, random_state=42)
```

In [ ]:

```python
# predict salary of x using sgd
predicted_salary = sgd_reg.predict(x)
predicted_salary
```

Out[ ]:

```
array([59.34627065])
```

In [ ]:

```python
# create polynomial features of degree 2 using scikit-learn PolynomialFeatures
# create X_poly using fit_transform
# create x_poly using transform
# fit the data using SGDRegressor
# predict salary of x using sgd
from sklearn.preprocessing import PolynomialFeatures
poly_features = PolynomialFeatures(degree=2, include_bias=False)
X_poly = poly_features.fit_transform(X_normalized)
x_poly = poly_features.transform(x)
sgd_reg.fit(X_poly, y)
predicted_salary = sgd_reg.predict(x_poly)
predicted_salary
```

Out[ ]:

```
array([59.512354])
```