gpa_years_experience_post

February 10, 2024

0.1 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

[ ]:

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

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

```
C:\Users\naumh\AppData\Local\Temp\ipykernel_10948\689908142.py:2:
DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of
pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better
interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

  import pandas as pd
```

[ ]:

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

```
[ ]:    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
```

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

     X = data.drop('salary', axis=1).values  #write in notes
     y = data['salary'].values

     X,y
```

```
[ ]: (array([[70. ,  1. ],
             [80. ,  2. ],
             [65. ,  2. ],
             [70. ,  2.5],
             [65. ,  2.7],
             [80. ,  3. ],
             [90. ,  3. ],
             [92. ,  3.2],
             [60. ,  3.5],
             [70. ,  3.7],
             [76. ,  4. ],
             [85. ,  4.5],
             [80. ,  5. ],
             [60. ,  5.5],
             [64. ,  5.8],
             [60. ,  6. ],
             [87. ,  6. ],
             [90. ,  6.5],
             [75. ,  7. ],
             [80. ,  7. ],
             [75. ,  7.5],
             [70. ,  8. ],
             [80. ,  8.5],
             [90. ,  8.7],
             [85. ,  9. ]]),
      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], dtype=int64))
```

[ ]:
```
# 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
```

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

[ ]:
```
# y should be a numpy array of shape (m, 1), use reshape(m, 1) to reshape y␣
 ↪from (m,) to (m, 1)
#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], dtype=int64)
y=y.reshape(m,1)

y
```

[ ]:
```
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]], dtype=int64)
```

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

```
X = scaler.fit_transform(X)

X
```

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

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

```
[ ]: 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.     ]])
```

[ ]:

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

[ ]:

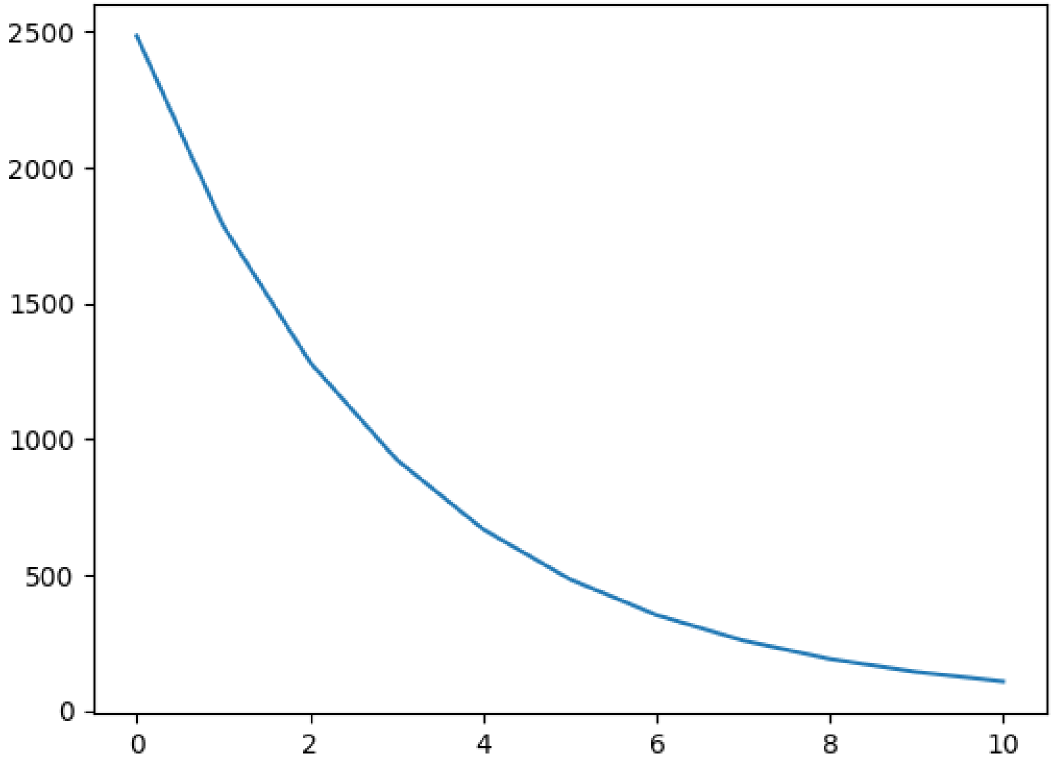
```
((25, 3), (25, 1))
```

[ ]:

```
#def h(theta, x):
 #   return x @ theta
#def J(theta, X, y):
#return 1/(2*m) * np.sum( (h(theta,X) - y)**2 )


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.
 ↪random.randn
theta = np.random.randn(n+1,1)

E = [1/(2*m) * np.sum(((X @ theta)-y)**2)] # list to store errors at each epoch
# compute error for initial theta and append to E

# loop over n_epochs
# for each epoch: compute gradients, update theta, compute error, append error␣
 ↪to E
for epoch in range(n_epochs):
    theta -= eta * (1/m * X.T @ ((X @ theta)-y))
    E.append(1/(2*m) * np.sum( ((X @ theta) - y)**2 ))

# plot error vs epoch
plt.plot(E)
```

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

```
[[37.20536484]
 [18.41593008]
 [21.12027021]]
```



```
[ ]: # 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
     from sklearn.preprocessing import MinMaxScaler

     x = np.array([[70,3]])
     print(x.shape)
     x = scaler.transform(x)
     x = add_dummy_feature(x)
     x
     theta_best = x @ theta
     print(x.shape)
```

```
theta_best
```

```
(1, 2)
(1, 3)
```

[ ]:
```
array([[48.24041055]])
```

[ ]:
```
# 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,y)
```

[ ]:
```
SGDRegressor(eta0=0.1, random_state=42)
```

[ ]:
```
# predict salary of x using sgd
sgd_reg.predict(x)
```

[ ]:
```
array([59.44433847])
```

[ ]:
```
# 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)
X_poly = poly_features.fit_transform(X)
x_poly = poly_features.transform(x)
sgd_reg.fit(X_poly, y)
predicted_salary = sgd_reg.predict(x_poly)

predicted_salary
```

[ ]:
```
array([60.05690321])
```