به نام خدا

جبر خطی کاربردی

گزارش پروژه‌ی سوم

پاییز ۱۴۰۰ - ۱۴۰۱

چمران معینی
۹۹۳۱۰۵۳

در این پروژه، قصد داریم یک رگرسیون خطی یک متغیره را، بر اساس داده‌های یک فایل CSV و به کمک زبان پایتون، پیدا بکنیم.

اولین کاری که می‌کنیم، خواندن داده‌های فایل است:

```
df = pd.read_csv('data.csv', header=None, names=['x', 'y'])
```

سپس مشخص می‌کنیم که چه نسبتی از دیتا را برای یادگیری می‌خواهیم استفاده کنیم:

```
train_percent = 0.95
train_size = int(train_percent * len(df))

train_df = df.iloc[:train_size]
test_df = df.iloc[train_size:]
```

باقی دیتا نیز برای تست استفاده خواهد شد.

در این جا، متغیر مستقل‌مان را x نامیده‌ایم و می‌خواهیم وابستگی y به آن را در فرمول $ax + b = y$ پیدا کنیم، پس مجهول‌های مسئله a و b هستند که با این تابع آن‌ها را به دست می‌آوریم:

```
   a, b = find_a_b_based_on_least_squares(train_df)


def find_a_b_based_on_least_squares(df):
   x = df['x']
   y = df['y']
   n = len(x)
   sum_x = sum(x)
   sum_y = sum(y)
   sum_xy = sum(x * y)
   sum_x_squared = sum(x ** 2)
   a = (n * sum_xy - sum_x * sum_y) / (n * sum_x_squared - sum_x ** 2)
   b = (sum_y - a * sum_x) / n
   return a, b
```

در این فانکشن، ساده‌ی شده‌ی روش مرسومی که برای یافتن least square وجود دارد را برای حالت تک‌متغیره می‌بینیم.

حال با داشتن a و b تنها کاری که می‌ماند، تست کردن این دو است که مطابق گزارش نیازهای پروژه، به این شکل دیتاهای تست را بررسی می‌کنیم:

```
    a, b = find_a_b_based_on_least_squares(train_df)
    test_result = test_a_b(a, b, test_df)

    for i in range(len(test_result)):
        print(f'Read value: {test_result[i][0]}'
              f'\nEstimated value: {test_result[i][1]}'
              f'\nError: {test_result[i][2]}\n')



def test_a_b(a, b, df):
    x = df['x'].values
    y = df['y'].values
    test_result = []
    for i in range(len(x)):
        read_value = y[i]
        estimated_value = a * x[i] + b
        error = abs(read_value - estimated_value)
        test_result.append([read_value, estimated_value, error])

    return test_result
```