در عصر اطلاعات، دسترسی به داده‌ها و کلان‌داده‌ها می‌تواند به ما قدرت زیادی برای پیش‌بینی برخی اتفاقات آینده بدهد. امروزه دانشمندان داده، با استفاده از الگوریتم‌های کامپیوتری و مفاهیم ریاضی، الگوهایی را در داده‌ها کشف می‌کنند و از آنها برای پیش‌بینی درمورد آینده استفاده می‌کنند.

همان‌طور که احتمالا حدس می‌زنید، در جبر خطی ما با یکی از ابتدایی‌ترین روش‌های الگویابی در داده‌ها آشنا شدیم و این روش چیزی نیست جز **مسئله‌ی کمترین مربعات**. در این پروژه می‌خواهیم راه حل این مسئله را در زبان پایتون پیاده‌سازی کرده و با یکی از مهم‌ترین کاربردهای جبر خطی آشنا شویم. ☺ توضیحات تئوری این مسئله و رگرسیون خطی در درس توضیح داده شده و پیشنهاد می‌کنیم پیش از شروع پروژه مروری بر آن داشته باشید.

یکی از کاربردهای علوم داده، کاربرد آن در علم پزشکی و تشخیص بیماری‌هاست. دانشمندان داده با استفاده از پرونده‌های پزشکی پیشین، الگوهایی برای تشخیص بیماری‌ها طراحی می‌کنند و با توجه به علائم بیماران جدید، احتمال ابتلای هر یک را به یک نوع بیماری خاص پیش‌بینی می‌کنند. نمونه دیتاستی که در فایل data.csv در اختیار شما قرار گرفته، شامل ۱۰۰ پرونده مربوط به بیماری سرطان خون است. ستون اول این داده غلظت گلبول‌های خون و ستون دوم آن احتمال ابتلا به سرطان خون را نشان می‌دهند. در این پروژه فرض می‌کنیم رابطه‌ای خطی میان این دو ستون از داده‌ها وجود دارد.

شما می‌توانید با استفاده از تابع read_csv که در کتابخانه‌ی pandas وجود دارد، محتویات این دیتاست را بخوانید.

شما باید به کمک این دیتاست و الگوریتم کمترین مربعات مدل رگرسیونی طراحی کرده و به پیش‌بینی داده‌های جدید بپردازید. با استفاده از ۹۵ درصد داده‌های موجود در دیتاست، یک مدل رگرسیون درجه ۱ آموزش دهید. سپس ۵ درصد باقی‌مانده را داده‌ی آزمون در نظر بگیرید و خطای مدل خود را محاسبه کنید. به ازای این ۵ درصد داده‌ی آزمون، مقادیر اندازه‌گیری شده را به صورت زیر چاپ کنید:

```
Read value: ?
Estimated value: ?
Error: ?
```

در نهایت خروجی مدل رگرسیونی و داده‌های نقطه‌ای را با استفاده از کتاب‌خانه‌ی matplotlib رسم کنید.

Scanned with CamScanner

نکات

- برای انجام این تکلیف فایل کد خود را به همراه یک فایل pdf. با گزارش مختصری از آنچه انجام داده‌اید را به صورت zip شده در سامانه‌ی courses آپلود کنید.
- برای پیاده سازی این پروژه تنها مجاز به استفاده از زبان پایتون و کتابخانه , Numpy pandas, matplotlibدر کنار توابع و کتابخانه‌های پیش فرض پایتون هستید .استفاده از هر زبان برنامه نویسی یا کتابخانه‌ای دیگر قابل قبول نبوده و در صورت استفاده، نمره‌ای به شما تعلق نخواهد گرفت.
- از رعایت تمیزی کد، استفاده از توابع مختلف برای پیاده سازی پروژه به شدت استقبال می شود.

موفق باشید

Scanned with CamScanner