گزارش تمرین شماره ۲

Admission Predict dataset

گردآوری : **علی شریفی**

شماره دانشجویی : **۹۵۲۲۲۰۴۶**

۱ مقدمه و مسئله :

این تمرین در راستای آشنایی با مفاهیم مدل های خطی همانند رگرسیون خطی داده شده بود . دیتاست مورد استفاده در این تمرین مربوط به داده های شانس پذیرش در دانشگاه و پارامتر هایی که ممکن است درپذیرش خانه تاثیرگذار هستند ،می باشد . این دیتاست شامل اطلاعات مربوط به ۴۰۰ متقاضی می باشد . هر متقاضی شامل ۹ رکورد است که این رکورد ها به تریتب عبارتند از :

- Serial No.
- GRE Score
- TOEFL Score
- University Rating
- SOP
- LOR
- CGPA
- Research
- Chance of Admit

رکورد Chance of Admit همان رکورد هدف ما می باشد یعنی ما میخواهیم با ساخت بهترین مدل خطی ممکن با ۷ پارامتر دیگر تخمین دقیقی از رکورد Chance of Admit داشته باشیم . توزیع متغیر هدف در دیتاست درشکل (۱) نمایش داده شده است .

۲ راه حل و ایده ها :

در ابتدا با حذف تاثیر نویز در داده ها ، داده ها را با استفاده از روش mean normalize ، نرمال میکنیم . حال به سراغ محاسبه تابع هزینه [1] با استفاده از معادله زیر می رویم :

$$h = (w^T X + b) \qquad (۱)$$

که در معادله (۱) ، w ماتریس وزن ها برای هریک از پارامتر ها و b ، بایاس مسئله می باشد .
حال تابع هزینه را به صورت زیر محاسبه میکنیم و هدف کلی ما کاهش مقدار تابع هزینه است که در عمل مسئله مطرح شده را به یک مسئله بهنیه سازی تبدیل میکند .

$$J = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h^{(i)})^2 \qquad (۲)$$

هم چنین دو فرمول زیر را هم خواهیم داشت :

$$\frac{\partial J}{\partial w} = \frac{-2}{m} X(y - h) \qquad (۳)$$

$$\frac{\partial J}{\partial b} = \frac{-2}{m}\sum_{i=1}^{m}(y^{(i)} - h^{(i)}) \qquad (۴)$$

حال با استفاده از روش گرادیان کاهشی خواهیم داشت :
در هر iteration ، مقادیر w و b را بروز رسانی میکنیم تا تابع هزینه به کمترین حالت خود دست یابد .

$$w = w - \alpha \times \frac{\partial J}{\partial w} \qquad (۵)$$

$$b = b - \alpha \times \frac{\partial J}{\partial b} \qquad (۶)$$

در فرمول (۵) و(۶) ، α ، ضریب یادگیری می باشد . با توجه به همگرا بودن روش گرادیان کاهشی انتظار خواهیم داشت تا حد امکان به نقطه اکسترمم تابع هزینه نزدیک شویم .
برای حل مسئله دو رویکرد متفاوت به کار گرفته میشود :

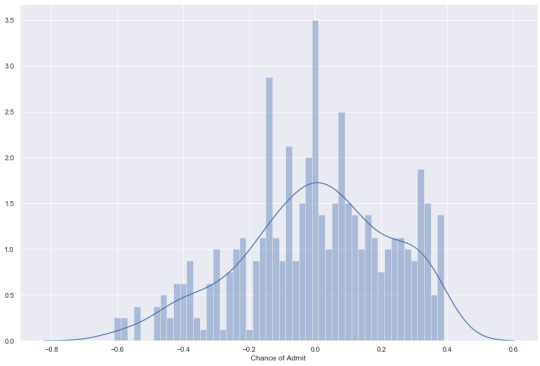
[1] Cost Function

۱. رویکرد ساخت مدل با پارامتر های کمتر با استفاده از ۳۲۰ رکورد

۲. رویکرد ساخت مدل با تمامی پارامتر های مطرح شده با استفاده از ۳۲۰ رکورد

۳. رویکرد ساخت مدل با تمامی پارامتر های مطرح شده با استفاده از ۳۵۰ رکورد

در رویکرد شماره ۱ با محاسبه ، رابطه خطی بین متغیر بیان شده در مسئله ، تنها متغیر هایی را وارد مسئله میکنیم که رابطه خطی نسبتا قابل توجه ای را دارا می باشند . که انتظار داریم در اکثر مسائل این رویکرد نتایج ضعیف تری نسبت به رویکرد ۲ ارایه کند . همان طور که در شکل (۲) مشاهده میشود دو متغیر SOP و CGPA به ترتیب با 0.68 و 0.87 بیشترین رابطه خطی را با متغیر هدف یعنی Chance of Admit را دارا می باشند .



شکل ۱: توزیع متغیر هدف در دیتاست

۳ ارزیابی نتیجه ها :

برای محاسبه دقت مدل ها از معادله زیر استفاده شده است :

(۷)
$$100-(\frac{\sum_{i=1}^{n}|Y_{prediction-train}-Y_{train}|}{n})*100$$

پس از ساخت مدل به رویکرد شماره (۱) به پارامتر های زیر برای مدل خطی خود رسیدیم .

```
-1.0454209676360753
[0.01096358 0.20116581]
```

که به ترتیب بایاس مسئله و ضرایب متغیر های SOP ، CGPA می باشند . مدل خطی ما به صورت زیر قابل نوشتن است :

(۸)
$$\begin{aligned} &MEDV_{Predict} = \\ &-1.0454209676360753+ \\ &(0.01096358)\times SOP+ \\ &(0.20116581)\times CGPA \end{aligned}$$

که دقت این مدل بر روی داده های test و train به ترتیب برابر است با

شکل ۲: ماتریس کورولیشن بین متغیرها

```
train accuracy:
70.76529809176706 %
test accuracy:
77.54853486490056 %
```

حال به بررسی رویکرد ۲ یعنی استفاده از تمامی متغیر ها برای ساخت مدل می پردازیم و خواهیم داشت :
ضرایب متغیر ها :

```
[[0.        ],
[0.        ],
[0.        ],
[0.0143454 ],
[0.171344  ],
[0.89233764],
[0.        ]]
```

بایاس :

```
[-0.00636733]
```

و طبق پیش بینی مشاهده خواهیم کرد این مدل ساخته شده دارای دقت بالاتری نسبت به مدل رویکرد شماره ۱ می باشد . دقت این مدل بر روی داده های test و train به ترتیب برابر است با

```
train accuracy:
92.21080464950819 %
test accuracy:
91.68267918598642 %
```

با افزایش متغیر ها از ۳۲۰ به ۳۵۰ در رویکرد شماره ۳ و با استفاده از تمامی متغیرها مشاهده میکنیم که مدل به دقت بهتری دست خواهد یافت و داریم :
ضرایب متغیر ها :

```
[[0.        ],
[0.        ],
[0.        ],
[0.02524058],
[0.15722352],
[0.87853869],
[0.        ]]
```

بایاس :

```
[-0.00252642]
```

```
train accuracy:
92.27503286561416 %
test accuracy:
91.47705086764267 %
```

۴ جمع بندی و نتیجه گیری

همان طور که مشاهده شد با افزایش متغیر ها در مدل خطی توانستیم به صورت قابل ملاحظه ای مدل خود را ارتقا دهیم .هم چنین نیز تاثیر افزایش تعداد رکورد های مورد استفاده در مدل نیز قابل توجه است . برخلاف انتظار افزایش رکورد ها کمکی به مدل ما نکرد . به دلیل over fit شدن مدل به دلایل مختلف از قبیل کم بودن داده ها شاید مدل کارآرایی خود را در مواجه با داده های جدید از دست دهد که برای جلوگیری از این کار توصیه میشود که از روش هایی از قبیل cross validation استفاده شود . هم چنین انتظار میرود که با افزایش تعداد داده ها، مدل به طور قابل توجه ای ارتقا یابد .