

مقدمه:

در این تمرین به بررسی دیتاست بیماری های قلبی UCL پرداختیم. این دیتاست شامل فیچرهایی از قبیل age, sex, ... بوده و لیبل مربوطه به صورت صفر در صورت عدم وجود بیماری قلبی و یک در صورت وجود بیماری قلبی میباشد.

سوال ها:

1. در یادگیری ماشین معمولاً در فضای فرضیه h بدنبال بهترین فرضیه ای هستیم که در مورد داده های آموزشی D صدق کند. یک راه تعیین بهترین فرضیه، این است که بدنبال محتمل ترین فرضیه ای باشیم که با داشتن داده های آموزشی D و احتمال قبلی در مورد فرضیه های مختلف می توان انتظار داشت تئوری بیز چنین راه حلی را ارائه می دهد. تئوری بیز امکان محاسبه احتمال ثانویه را بر مبنای احتمالات اولیه می دهد:

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)}$$

تفاوت میان مدل های مختلف این مدل عبارتند از

در gaussian naive bayes فرض میکنم که تمام فیچرها توزیع نرمال داشته و continuous می باشد.

در multinomial naive bayes فرض میکنیم که تمام فیچرها به صورت گسسته میباشند.

در bernoulli naive bayes فرض بر این است که تمام فیچرها به صورت صفر و یکی می باشند.

2. این بخش پیاده سازی شده و کدهای مربوط ضمیمه شده.

3. پس از پیاده سازی مدل و انجام آموزش بر روی سه فیچر انتخاب شده به نتایج زیر رسیدیم:

```
f1_score: 0.763157894736842
precision: 0.6590909090909091
recall: 0.90625
```

4. نتایج زیر به کمک مدل sklearn بر روی سه فیچر انتخاب شده حاصل شدند:

```
f1_score: 0.7761194029850748
precision: 0.7428571428571429
recall: 0.8125
```

5. پس از اعمال مدل خود بر روی تمام دیتا نتایج زیر حاصل شد:

```
f1_score: 0.676056338028169
precision: 0.6153846153846154
recall: 0.75
```

و به کمک مدل sklearn نتایج زیر بدست آمد:

```
f1_score: 0.870967741935484
precision: 0.9
recall: 0.84375
```

با توجه به نتایج حاصل شده میتوان نتیجه گرفت مدل sklearn دارای پیاده سازی کاملتری میباشد.