تمرین سری 13

بکتاش انصاری

99521082

بخش عملی :

از آنجایی که داده های بصورت پیوسته هستند ما از توزیع گوسین استفاده میکنیم.

Naïve Bayes Classifier را ابتدا با چندین تابع پیاده سازی میکنیم.

که در هر تابع به صورت خلاصه این کار ها را انجام میدهیم :

تابع sep_class :

در این تابع دیتا های مورد نظر را بر اساس نوع (class) آنها در یک دیکشنری ذخیره میکنیم

تابع avg :

میانگین فیچر دیتا را محاسبه میکنیم.

تابع dev :

انحراف معیار را محاسبه میکنیم

تابع final_dataset :

در این تابع هر دیتا را با مقادیر میانگین و انحراف معیار ذخیره میکنیم.

تابع final_by_class :

در این تابع دیتا را با نوع هر داده (class) ذخیره میکنیم

تابع prob :

مقدار احتمال گوسین را محاسبه میکنیم.

تابع classes_prob :

احتمال هر کلاس را برای دیتا محاسبه و ذخیره میکنیم.

تابع predict :

در این تابع مقدار پیشبینی کلاس هر دیتا را محاسبه میکنیم.

تابع naïve :

در این تابع دیتاهای train و test را بر روی classifier پیاده سازی میکنیم.

حال دیتا را به NB میدهیم (80 درصد دیتا به صورت train و 20 درصد به صورت test)

و نتیجه را مشاهده میکنیم. (accuracy محاسبه شده را خروجی میدهیم.)

```
full Accuracy of my Naive Bayes :  86.20689655172413
```

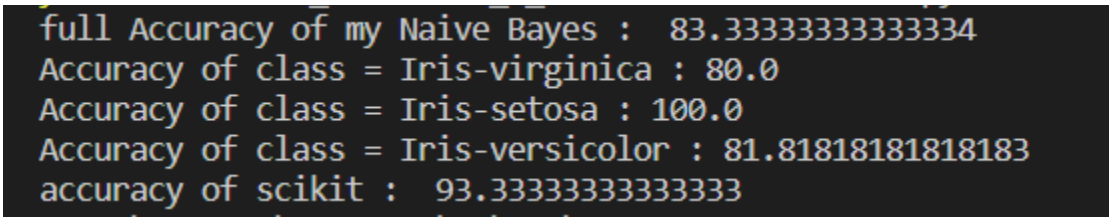
حال برای بهبود عملکرد میتوانیم میزان دیتای train و test را تغییر دهیم.

برای مثال اگر میزان دیتای train را به 90 درصد برسانیم داریم :

```
full Accuracy of my Naive Bayes :  92.85714285714286
```

این مقادیر ثابت نیست و با ران کردن های متعدد مقادیر متفاوتی خواهید گرفت.

حال دقت را برای هر کلاس نیز محاسبه میکنیم. خروجی مورد نظر به شکل زیر خواهد بود :



در نهایت نیز این دیتا ها را بر روی NB آماده در کتابخانه scikit اجرا میکنیم و نتیجه را مشاهده میکنیم.

بخش تئوری :

سوال یک :

Logistic Regression در اوایل قرن 20 ام در علوم بیولوژیکی مورد استفاده قرار میگرفت.

از این روش وقتی استفاده میکنیم که متغیر های ما (target) به صورت طبقه بندی شده باشند.

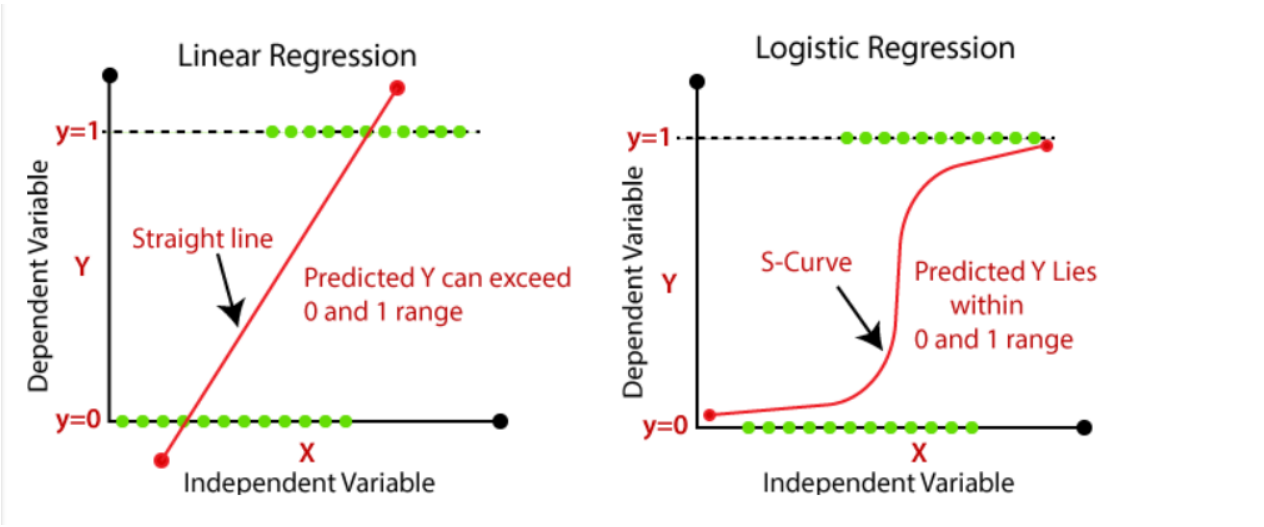
برای مثال :

- پیشبینی کنیم که آیا یک ایمیل spam هست یا نه (0 or 1)
- پیشبینی کنیم که آیا یک تومور بدخیم هست یا نه (0 or 1)

این روش Regression برای توصیف داده ها و توضیح رابطه بین یک متغیر باینری وابسته و یک یا چند متغیر مستقل که به صورت ترتیبی، بازه ای ، یا نسبتی استفاده میشود.

مقایسه روش linear Regression و Logistic Regression :

تفاوت اصلی بین این دو روش این است که Linear Regression برای حل مسائل Regression استفاده میشود در حالی که Logistic Regression برای حل مسائل Classification استفاده میشود.



Linear Regression برای پیشبینی متغیر های وابسته ی پیوسته با کمک متغیر های مستقل استفاده میشود ولی Logistic Regression برای پیشبینی متغیر های وابسته ی طبقه بندی شده به کمک متغیر های مستقل استفاده میشود.

خروجی Linear Regression باید به صورت پیوسته باشد مانند قیمت سن و ....

خروجی Logistic Resgression باید به صورت طبقه بندی شده مانند 0 یا 1 و yes یا no.

سوال دو :

احتمالات مورد نظر :

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

P(c) = ¾

P(j) = ¼

P(chinese|*c*) = 6 / 14 = 3/7

P(tokyo|*c*) = 1/14

P(japan|*c*) = 1/14

P(chinese|*j*) = 2/9

P(tokyo|*j*) = 2/9

P(japan|*j*) = 2/9

احتمالات مورد نظر :

P(c|d5) = 3/4 * (3/7)^3 * 1/14 * 1/14 = 0.0003

P(j|d5) = 1/4 * (2/9)^3 * 2/9 * 2/9 = 0.0001