«Data Mining course»

Project 2

Dr.Farahani , Dr.Kheradpisheh

Ashkan Safavi Sohi

98422096

1. **در ابتدا به بررسی دیتاست با استفاده از پکیج pandas بپردازید .**

به جواب این سوال در فایل پیوست پرداخته شده است .

2. **نمونه های موجود در دیتاست را با نسبت ۸۰ به ۲۰ به دو بخش داده های آموزشی و داده های تست تقسیم بندی کنید . برای این کار میتوانید از پکیج sklearn بهره ببرید.**

به جواب این سوال در فایل پیوست پرداخته شده است .

3. **قضیه بیز را در حداقل یک پاراگراف بیان کنید . سپس دسته بند های Gaussian Bernoulli Naive Bayes ، Multinomial Naive Bayes، Naive Bayesرا با یکدیگر مقایسه کنید و بیان کنید هر کدام از این دسته بندها بیشتر در کجا کاربرد دارند .**

قضیه بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده‌است و در نظریه احتمالات با اهمیت و پرکاربرد است. اگر برای فضای نمونه‌ای مفروضی بتوانیم چنان افرازی انتخاب کنیم که با دانستن اینکه کدامیک از پیشامدهای افراز شده رخ داده‌است، بخش مهمی از عدم قطعیت تقلیل می‌یابد.این قضیه از آن جهت مفید است که می‌توان از طریق آن، احتمال یک پیشامد را با مشروط کردن نسبت به وقوع یا عدم وقوع یک پیشامد دیگر محاسبه کرد. در بسیاری از حالت‌ها، محاسبهٔ احتمال یک پیشامد به صورت مستقیم کاری دشوار است. با استفاده از این قضیه و مشروط کردن پیشامد مورد نظر نسبت به پیشامد دیگر، می‌توان احتمال مورد نظر را محاسبه کرد.

Gaussian Naive Bayes:

اگر داده‌ها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیرهای مربوط به شواهد می‌توان استفاده کرد. در این حالت هر دسته یا گروه دارای توزیع گاوسی است. به این ترتیب اگر k دسته یا کلاس داشته باشیم می‌توانیم برای هر دسته میانگین و واریانس را محاسبه کرده و پارامترهای توزیع نرمال را برای آن‌ها برآورد کنیم. فرض کنید که μk میانگین و σ2k واریانس دسته k ام یعنی Ck باشد. همچنین ν را مشاهدات حاصل از متغیرهای تصادفی X در نظر گرفت. از آنجایی که توزیع X در هر دسته گاوسی (نرمال) فرض شده است، خواهیم داشت:

Multinomial Naive Bayes:

بیز ساده چندجمله‌ای، به عنوان یک دسته‌بند متنی بسیار به کار می‌آید. در این حالت برحسب مدل احتمالی یا توزیع چند جمله‌ای، برداری از n ویژگی برای یک مشاهده به صورت X=(x1,...,xn) با احتمالات (p1,...,pn) در نظر گرفته می‌شود. مشخص است که در این حالت بردار X بیانگر تعداد مشاهداتی است که ویژگی خاصی را دارا هستند. به این ترتیب تابع درستنمایی در چنین مدلی به شکل زیر نوشته می‌شود:

Bernoulli Naive Bayes:

در این قسمت به بررسی توزیع برنولی و دسته‌بندی بیز خواهیم پرداخت. به شکلی این نوع از دسته‌بند بیز بیشترین کاربرد را در دسته‌بندی متن‌های کوتاه داشته، به همین دلیل محبوبیت بیشتری نیز دارد. در این مدل در حالت چند متغیره، فرض بر این است که وجود یا ناموجود بودن یک ویژگی در نظر گرفته شود. برای مثال با توجه به یک لغتنامه مربوط به اصطلاحات ورزشی، متن دلخواهی مورد تجزیه و تحلیل قرار می‌گیرد و بررسی می‌شود که آیا کلمات مربوط به لغتنامه ورزشی در متن وجود دارند یا خیر. به این ترتیب مدل تابع درستنمایی متن براساس کلاس های مختلف Ck به شکل زیر نوشته می‌شود:

4 . **با در نظر گرفتن فیچر ها thalach ، trestbps، cholو لیبل target یک دسته بند Gaussian Naive Bayes را از پایه پیاده سازی کنید .)بدون استفاده از پکیج)برای این کار شما نیاز است که در دیتاست آموزشی خود اعضای مختلف قاعده بیز را محاسبه کنید .**

به جواب این سوال در فایل پیوست پرداخته شده است .

5 . **پس پیاده سازی Gaussian Naive Bayes و آموزش آن بر روی داده های آموزشی)۸۰ درصد دیتاست(. نتایج را برای داده های تست)۲۰ درصد باقی دیتاست(بررسی کنید به عبارت دیگر برای داده ورودی بررسی کنید در بخش تست لیبل را پیش بینی کنید . با توجه به این لیبل های واقعی را نیز دارید معیار های زیر گزارش دهید .**

- F1 score
- Recall
- Precision

به جواب این سوال در فایل پیوست پرداخته شده است .

6 . **با استفاده از پکیج sklearn و GaussianNB یک مدل بسازید و بر روی داده های آموزشی ، ترین کنید سپس بر روی داده های تست همانند سوال (۵) سه معیار را گزارش دهید .**

به جواب این سوال در فایل پیوست پرداخته شده است .

7 . **بررسی کنید که در سه معیار مطرح شده مدلی که با استفاده از پکیج ساخته اید و مدلی که خود پیاده سازی کرده اید به چه صورتی عمل کرده اند.**

به جواب این سوال در فایل پیوست پرداخته شده است.

8 . **کلاسیفایر SVM را با استفاده از پکیج sklearn بر سه فیچر مطرح شده در سوال (۴) با استفاده از داده های آموزشی ترین کنید . سپس بر روی داده های تست سه معیار Recall ، Precision، F1 scoreرا گزارش کنید .**

به جواب این سوال در فایل پیوست پرداخته شده است.

9 . **حداقل دو حالت مختلف را برای کرنل در SVM ساخته شده با پکیج در نظر بگیرید و نتایج آن را گزارش دهید. آیا کرنل های مختلف نتایج مختلفی ارایه دادند ؟ به صورت کلی علت استفاده از کرنل ها در SVM چیست توضیح دهید .**

به جواب این سوال در فایل پیوست پرداخته شده است.

10 . **دسته بند SVM را با استفاده از پکیج sklearn بسازید و با در نظر گرفتن کلیه فیچرهای دیتاست بر روی داده های آموزشی ترین کنید سپس نتایج را بر روی داده های تست ، ارزیابی کنید .**

به جواب این سوال در فایل پیوست پرداخته شده است.

11 . **برای سوال (۱۰) ، یکبار مدل را با 5-fold Cross Validation اجرا کنید و نتایج را گزارش دهید . (در این جا برای فولد کردن داده ها از کل دیتاست استفاده میکنیم).**

به جواب این سوال در فایل پیوست پرداخته شده است.

12 . **با استفاده از پکیج sklearn دسته بند را K-NN [۳] را بسازید . با به کارگیری تمامی فیچرها موجود در دیتاست آموزشی ، مدل را ترین کنید سپس بر روی دیتاست تست ، ارزیابی کنید.**

به جواب این سوال در فایل پیوست پرداخته شده است.

13 . **بررسی کنید در سوال (۱۲)تعداد همسایه ها k چه نقشی ایفا میکند ؟ زیاد شدن همسایه ها خوب است ؟ چگونه میتوان مشخص کرد چه تعداد همسایه برای مسئله ما مناسب است .**

تعداد همسایه ها باعث تغییر در دقت یادگیری می‌شود. در الگوریتم k نزدیکترین همسایگی، classification به میزان بیشترین تعداد مشترک دسته بندی شده همسایگان می‌باشد. برای مثال اگر تعداد نزدیکترین همسایگی را همانند شکل زیر یکبار 3 در نظر بگیریم، نتیجه با توجه به دو همسایه نزدیک قرمز و یک همسایه آبی، قرمز خواهد بود. ولی در صورتی که نزدیکترین همسایگی را 5 در نظر بگیریم، نتیجه از نوع کلاس آبی خواهد بود. بنابرین تعداد میزان همسایگی در هر مساله متقاوت هست و معمولا بیشتر با آزمون و خطا می‌توان به نتایج دقیق‌تری دست یافت. البته باید به این نکته توجه کرد که هر چه تعداد همسایگان را بیشتر در نظر بگیریم، احتمال پراکندگی نتایج ممکن است بیشتر باشد. و اگر تعداد همسایگان نیز خیلی کم باشد، باعث خطا (با توجه به داده‌های استثنا) و خطی بودن نتایج شود.

14. در سوال (۱۲) به جای استفاده از تمامی فیچرها فقط از سه فیچر trestbps ، chol، thalachاستفاده کنید و مدل را بسازید . سپس نتایج ارزیابی را گزارش دهید .

به جواب این سوال در فایل پیوست پرداخته شده است.

15. تفاوت بین روش های کلاس بندی پارامتری و غیرپارامتری را به صورت خلاصه بیان کنید . هر کدام بهتر است در چه مواقعی استفاده شوند ؟

در یک مدل پارامتری ، تعداد پارامترها با توجه به اندازه نمونه ثابت می شود. در یک مدل غیر پارامتری ، تعداد (موثر) پارامترها می توانند با اندازه نمونه رشد کنند.در یک رگرسیون OLS ، تعداد پارامترها همیشه به طول β خواهد بود، به علاوه یک واریانس.یک شبکه عصبی با معماری ثابت و بدون پوسیدگی وزن یک مدل پارامتریک است.اما اگر دچار فروپاشی وزن هستید ، مقدار پارامتر پوسیدگی که با اعتبار سنجی متقابل انتخاب می شود ، با داده های بیشتر ، به طور کلی کوچکتر می شود. این می تواند به عنوان افزایش تعداد موثر پارامترها با افزایش اندازه نمونه تفسیر شود.

16. (بخش امتیازی) معیار Matthews Correlation Coefficienچیست و در چه جاهایی استفاده میشود .

Matthews Correlation Coefficien : پارامتری است که برای ارزیابی کارایی الگوریتم‌های یادگیری ماشین از آن استفاده می‌شود. این پارامتر بیان‌گر کیفیت کلاس‌بندی برای یک مجموعه باینری می‌باشد. بنابرین مواقعی از این معیار استفاده می‌گردد که classification ما همیشه دو بخشی باشد.