به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی و علوم

یادگیری ماشین

دکتر میرزایی

تمرین 2

اسفند ماه ۹۸

سوال اول:

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	False	Yes
Rain	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rain	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rain	Mild	High	True	No

جدول بالا، معروف به جدول بازی تنیس، یکی از دادگان محبوب برای آزمایش مفاهیم یادگیری ماشین بر روی دادگان با اندازه کوچک می‌باشد. در ابتدا با تحقیق در خصوص جدول داده شده مفهوم آن را بیابید.

1- طبقه بند با استفاده از مدل مولد را به طور کامل شرح دهید، سپس روابط استفاده شده برای این طبقه بند را به صورت شهودی تفسیر کنید.

2-مزایا و معایب استفاده از این طبقه بندی را شرح دهید. برای حل یک مسئله یادگیری ماشین، چه زمانی استفاده از طبقه بند مولد را پیشنهاد میکنید؟

3-در جدول دادگان بازی تنیس، همان گونه که مشاهده می‌کنید، ویژگی اول دارای 3 مقدار، ویژگی دوم 3، ویژگی سوم 2 و ویژگی چهارم هم 3 مقدار متفاوت دارند، یعنی در مجموع 54 حالت مختلف برای بازی تنیس بوجود می‌آید، که دادگان آموزش شما تنها شامل 14 حالت می‌باشد. امکان برگزاری بازی تنیس برای 40 حالت دیگر که در جدول ذکر نشده اند را با استفاده از

طبقه بند مولدی که طراحی کرده اید پیش‌بینی کنید. برای مثال یکی ازین40 حالت: هوای بارانی، دمای گرم، رطوبت بالا و باد ضعیف می‌باشد.

4-آیا میتوانید از این طبقه بند برای وقتی که ازهمه ویژگی های دادگان تست به صورت کامل اطلاع نداریم استفاده کنید؟ برای مثال اگر بدانیم وضعیت هوا آفتابی و دما به صورت معتدل میباشد ولی از وضعیت رطوبت و باد اطلاع نداشته باشیم. اگر امکان طبقه بندی به این صورت امکان پذیر می‌باشد 10 تا دوتایی از ویژگی ها، طبقه بندی را انجام دهید.

سوال دوم:

به همراه فایل پروژه دو دسته داده به نام های data_train و data_test آپلود شده است که در داده های data_train اطلاعات 10000 بیمار و در داده های data_test اطلاعات 1000 بیمار داده شده است. این داده ها در 6 کلاس به شرح زیر طبقه بندی شده اند:

Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
چربی خون	قند خون	اوره خون	تعداد گلبول های قرمز	تعداد گلبول های سفید	کلاس داده ها

یک بیماری نادر نیز در بین بیمار ها وجود دارد که در آن کلاس بیمار برابر 1 است.
در این تمرین میخواهیم با استفاده از الگوریتم Naïve Bayes مدل را آموزش دهیم و بیمارها را تشخیص دهیم.

1- اگر بدون اینکه بخواهیم از اطلاعات بیماران استفاده کنیم (میزان چربی وقند و...)و بگوییم یک فرد بیماری را دارد یا خیر به چه احتمالی ممکن است بیماری را داشته باشد؟

2- در این قسمت فرض کنید تمامی متغیر های هر یک از attributeها از یک متغییر تصادفی نرمال با واریانس و میانگین متفاوت برای افراد بیمار و غیر بیمار پیروی کنند. حال با استفاده از روش Naïve Bayes طبقه بند آموزش دهید و دقت را روی داده های تست بیان کنید.

3- حال با توجه به قسمت 2 اگر جواب آزمایش یک فرد مثبت باشد به چه احتمالی آن فرد بیماری را دارد؟

سوال سوم:

به همراه فایل تمرین، دادگان iris آپلود شده است که با تحقیق کردن در خصوص این دادگان در منابع، از جزییات و همچنین کلاس های آن، اطلاعات کافی بدست آورید.

1-برای طبقه بندی به کمک طبقه بند بیزین می‌توان از دو تخمین زن ML و MAP استفاده کرد. این دو تخمین زن را با هم مقایسه کنید و مزایا و معایب هر کدام را شرح دهید.

2-با shuffle کردن مجموعه دادگان، 70 درصد آنها را بعنوان دادگان آموزش و 30 درصد را بعنوان دادگان تست در نظر بگیرید.

3-با پیش پردازش مناسب بر روی دادگان آموزش، سعی کنید با استفاده از تخمین زن ML کلاس دادگان تست را تخمین بزنید(برای نرمال سازی دادگان، از توزیع های معروف به دلخواه خودتان استفاده کنید. یکی از توزیع های معروف که مورد استفاده قرار میگیرد، توزیع نرمال می‌باشد)

4-برای تخمین های خود، ماتریس کانفیوژن را تشکیل دهید و از روی آن دقت پیش بینی های خود را گزارش کنید.

5- اگر مجموعه دادگان آموزش کمتر و کمتر شود مثلا بجای 70 درصد، فقط مجاز به استفاده از 10 درصد دادگان آموزش باشیم، قسمت 3 را مجددا تکرار کنید و نتیجه را با وقتی که از 70 درصد دادگان به عنوان آموزش استفاده کردید مقایسه کنید.

6-اگر نتیجه قسمت 5 درمقایسه با وقتی که از 70 درصد دادگان استفاده کردید بدتر شد، راهی برای جبران و تقویت مدل خود پیشنهاد دهید؟ (امتیازی: مدل مد نظر را پیاده سازی کنید و دقت را با 2 حالت قبل در ماتریس کانفیوژن مقایسه کنید)

نکات:

- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفا تمامی نکات و فرض‌هایی که برای پیاده سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- برای دادگان iris و همچنین دادگان بیماران، با استفاده از توابع مناسب در کتابخانه های پایتون، فایل mat. را بخوانید و لود کنید.
- در صورت مشاهده‌ی تقلب نمرات تمامی افراد شرکت کننده در آن صفر لحاظ می‌شود.
- استفاده از کدهای آماده برای تمرینها مجاز نمی‌باشد. برای تمرین ها فقط برای قسمتهایی از کد برای پیاده‌سازی میتوانید از کدهای آماده راهنمایی بگیرید. بنابراین، کپی کردن ساختار ها و کدهای آماده و حل شده از اینترنت **تقلب** محسوب می‌شود.
- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌ی زیر با دستیار آموزشی مربوطه در تماس باشید:

هاشم پور(hamidreza.hashemp@ut.ac.ir)

- گزارش کار های خود را به آیدی @cactus_995 در تلگرام و همچنین ایمیل استاد ارسال کنید.