مینی‌پروژهٔ شمارهٔ دو

۱ پرسش یک

الف) طبقه‌بند بیز ساده[۱] و بیز بهینه[۲] را به‌صورت مختصر توضیح دهید و تفاوت‌های میان آن‌ها را بررسی نمایید.

ب) با توجه به مجموعه‌داده‌ی پیامک‌های اسپم، مشخص نمایید که کدام نوع از طبقه‌بند بیز ساده[۳] برای این داده مناسب‌تر است و علت انتخاب خود را بیان کنید.

ج) طبقه‌بند انتخاب‌شده را بدون استفاده از هیچ‌گونه کتابخانه‌ی آماده پیاده‌سازی نمایید.

د) در این بخش، دو الگوریتم یادشده را با استفاده از کتابخانه‌ی سایکیت‌لِرن[۴] بر روی همان مجموعه‌داده اجرا نمایید. سپس، خروجی‌های به‌دست‌آمده شامل ماتریس آشفتگی، دقت، دقت مثبت[۵] و بازیابی[۶] را با نتایج پیاده‌سازی دستی مقایسه کرده و مورد تحلیل قرار دهید.

هـ) با ارائه‌ی یک اثبات ریاضی نشان دهید که **قانون تصمیم‌گیری بیز بر اساس بیشینه‌ی پسین**[۷]، احتمال خطا را در مسئله‌ی طبقه‌بندی به‌صورت نظری کمینه می‌نماید.

و) فرض کنید در یک سامانه‌ی فیلتر پیام، اشتباه در تشخیص یک پیام تبلیغاتی به‌عنوان پیام معتبر، دارای هزینه‌ای معادل ۵ برابر بیشتر از حالتی است که یک پیام معتبر به‌اشتباه به‌عنوان اسپم شناسایی شود. با در نظر گرفتن این تفاوت در هزینه‌ها، تابع **ریسک تصمیم‌گیری**[۸] را برای این مسئله تعریف کرده و بر اساس آن، دسته‌بندی بهینه را تعیین نمایید. توضیح دهید که این موضوع چگونه مرز تصمیم‌گیری را تحت تأثیر قرار می‌دهد.

۲ پرسش دو

داده‌های مربوط به تصاویر ارقام دست‌نویس در دیتاست MNIST را در نظر بگیرید. این دیتاست شامل تصاویر خاکستری با اندازه 28×28 پیکسل می‌باشد که هر تصویر به‌صورت یک بردار با ۷۸۴ ویژگی نمایش داده شده است. مراحل زیر را انجام دهید:

آ) از دیتاست MNIST، تعداد 10000 نمونه انتخاب کرده و داده‌ها را به کمک یکی از روش‌های StandardScaler یا MinMaxScaler نرمال‌سازی کنید. سپس دلیل انتخاب روش نرمال‌سازی را با توجه به ماهیت داده‌ها بیان نمایید.

ب) داده‌های نرمال‌سازی‌شده را به دو بخش آموزش و تست با نسبت 70% به 30% تقسیم کنید.

---

[۱] naïve bayes
[۲] optimal bayes
[۳] مانند: gaussian ، multinomial، bernoulli
[۴] scikit-learn
[۵] precision
[۶] recall
[۷] posteriori a maximum
[۸] risk function

ج) با استفاده از الگوریتم K-Nearest Neighbors (KNN)، مدل را برای حداقل سه مقدار مختلف k (برای مثال $k = 3, 5, 9$) آموزش دهید. سپس با تغییر مقدار k در بازه‌ای مانند 1 تا 25 با گام 2، عملکرد مدل را ارزیابی کرده و بهترین مقدار k را تعیین کنید.

د) دقت مدل را برای مقادیر مختلف k در یک نمودار رسم کنید تا تاثیر پارامتر k بر عملکرد مدل به صورت بصری مشخص شود.

د) یک روش بهبود برای افزایش دقت الگوریتم KNN پیشنهاد دهید. برای مثال می‌توانید از کاهش بعد با روش PCA استفاده کنید. توضیح دهید که چرا این روش می‌تواند موثر باشد و سپس عملکرد مدل را برای مقادیر مختلف مؤلفه‌های اصلی (مثلاً 20، 40، 60، ...، 100) ارزیابی کنید. نتیجه را در قالب نمودار Accuracy vs. Number of PCA Components نمایش دهید.

۳ پرسش سه

یك شرکت تولید پوشاك به دنبال این است که بفهمد کدام ویژگی‌ها یا بخش‌ها باعث افزایش فروش می‌شوند. برای این منظور، از الگوریتم‌های مختلفی برای تحلیل داده‌ها استفاده می‌شود. داده‌هایی که به شما داده شده شامل ۱۰ ویژگی و ۴۰۰ رکورد است. هدف این است که از درخت تصمیم و الگوریتم‌های دیگر برای شناسایی ویژگی‌هایی استفاده کنید که بیشترین تاثیر را بر افزایش فروش دارند.

۱.۳

مجموعهsection دادگان[۱] دیتاستی که در این پروژه استفاده می‌شود در قالب Comma Separated Values یا به اختصار (.CSV) می‌باشد و بایستی ابتدا آنرا به شکل DataFrame بخوانید. (راهنمایی: با استفاده از تابع pandas.read_csv() می‌توانید این کار را انجام دهید.) برای دریافت دیتاست می‌توانید از این پیوند استفاده کنید.

- با استفاده از تابع head(). ده سطر اول دیتاست را بخوانید و نمایش دهید.

۲.۳ پیش‌پردازش دیتاست

پیش‌پردازش [۲] دیتاست یکی از مراحل ابتدایی و بسیار مهم در علم داده، یادگیری ماشین و هوش مصنوعی است. این مرحله شامل مجموعه‌ای از تکنیک‌ها و عملیات‌هایی است که روی داده‌های خام[۳] انجام می‌شود تا آنها را برای تحلیل یا آموزش مدل آماده کند. داده‌های خام معمولاً ناقص، ناپیوسته، نویزی یا دارای مقیاس‌های مختلف هستند. بدون انجام پیش‌پردازش، مدل‌های یادگیری ماشین نمی‌توانند به خوبی آموزش ببینند یا ممکن است نتایج نادرست و غیرقابل اطمینانی تولید کنند. بدین منظور مراحل زیر را دنبال می‌نماییم:

- ابتدا مشخص نمایید که در درون دیتاست داده‌های ناقص یا گمشده[۴] وجود دارد یا خیر؟ درصورت وجود داده‌های ناقص یا گمشده، آنها را از دیتاست حذف نمایید.
- یکی از مشکلات رایج در درون دیتاست‌ها وجود داده‌های تکراری[۵] می‌باشد. ابتدا توضیح دهید که داده‌های تکراری چگونه می‌توانند برای مدل آموزش دیده و تحلیل داده‌ها مشکل ایجاد نمایند. سپس مشخص نمایید که چه تعداد دادهٔ تکراری در درون دیتاست است و در صورت وجود آنها را حذف نمایید.
- ویژگی‌هایی مانند محل فروختن محصول(شهر یا غیرشهر)، در آمریکا یا خارج از آمریکا فروخته شدن محصول و محل قفسه‌های فروش محصول جزو ویژگی‌های دسته‌ای می‌باشند. برای اینکه از چنین ویژگی‌هایی بتوانیم در آموزش مدل استفاده نماییم، می‌بایست آنها را به ویژگی‌های عددی[۶] تبدیل نماییم بنابراین این ویژگی‌ها را Encode نمایید.

---

[۱] dataset
[۲] data preprocessing
[۳] raw data
[۴] missing values
[۵] duplicate data
[۶] numerical

- همانطور که بیان شد، هدف دیتاست این است که ببیند این ویژگی‌ها چطور می‌توانند باعث افزایش فروش شوند، بنابراین متغیر هدف در چنین دیتاستی میزان فروش محصول (sales) می‌باشد، اما همانطور که مشاهده می‌شود جنس خروجی‌ها از جنس متغیرهای عددی می‌باشند و هدف ما نیز آموزش مدل با درخت تصمیم است. برای این منظور نیاز است تا متغیرهای عددی را به متغیرهای دسته‌ای تبدیل نماییم. این کار را برای متغیر هدف پیاده‌سازی نمایید. (تعداد کلاس‌ها را با توجه به برآورد خودتان از متغیر هدف تنظیم نمایید.)

بدین ترتیب مرحلهٔ پیش‌پردازش بر روی دیتاست به پایان می‌رسد.

- ماتریس همبستگی[1] بین ویژگی‌های دیتاست و متغیر هدف را ترسیم نمایید.
- پس از ترسیم ماتریس تحلیل نمایید که کدام ویژگی‌ها هبستگی بیشتری با یکدیگر و متغیر هدف دارند.

۳.۳ محاسبه آنتروپی داده‌ها

در یادگیری ماشین، زمانی که از درخت تصمیم برای دسته‌بندی داده‌ها استفاده می‌کنیم، نیاز داریم در هر گره تصمیم بگیریم که کدام ویژگی داده را به بهترین شکل تقسیم می‌کند. برای این کار از معیارهایی برای اندازه‌گیری ناپایداری یا بی‌نظمی[2] استفاده می‌کنیم. آنتروپی یک مفهوم از نظریه اطلاعات است که میزان بی‌نظمی یا عدم قطعیت در یک مجموعه داده را اندازه‌گیری می‌کند.
فرمول آن به شکل زیر است:

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

که در آن:

- S: مجموعه‌ای از نمونه‌هاست
- p_i: احتمال وقوع کلاس i در مجموعه S است.

اگر همه نمونه‌ها از یک کلاس باشند، آنتروپی صفر است (یعنی بی‌نظمی نداریم). اگر نمونه‌ها به صورت مساوی بین کلاس‌ها پخش شده باشند، آنتروپی بیشترین مقدار را خواهد داشت.

- تابع calculate_entropy(y) را به صورت صحیح پیاده سازی کنید.(راهنمایی: y یک آرایه از جنس ndarray می باشد.)

۴.۳ محاسبهٔ information gain

در ساخت درخت تصمیم، انتخاب بهترین ویژگی برای تقسیم داده‌ها نقش بسیار مهمی در دقت و عملکرد مدل دارد. یکی از رایج‌ترین معیارها برای انتخاب ویژگی مناسب، یا کسب اطلاعات است. کسب اطلاعات نشان می‌دهد که با تقسیم داده‌ها بر اساس یک ویژگی خاص، تا چه اندازه از بی‌نظمی (Entropy) مجموعه کاهش پیدا می‌کند. به بیان ساده‌تر، این معیار به ما می‌گوید یک ویژگی چقدر برای پیش‌بینی درست کلاس داده‌ها مفید است.

فرمول آن به شکل زیر است:

$$IG(S, A) = Entropy(S) - \sum_{v \in Values(A)} weight_v . Entropy(S_v)$$

که در آن:

- S: مجموعه داده‌ی اصلی.
- $Values(A)$: مقادیر ممکن برای ویژگی A.
- S_v: زیرمجموعه‌ای از S که در آن ویژگی $A = v$ است.

[1] correlation matrix
[2] impurity

- $Weight_v = \frac{|s_v|}{|s|}$: نسبت تعداد نمونه‌های زیرمجموعه S_v (فرزند)به کل داده‌ها گره پدر(وزن زیرمجموعه).
- تابع info_gain(parent, children) را به صورت صحیح پیاده سازی کنید.(راهنمایی: parent لیستی از برچسب‌های مجموعه اصلی، children لیستی از زیرمجموعه‌های حاصل از تقسیم است که هرکدام شامل برچسب‌های کلاس خود هستند.)

۵.۳ درخت تصمیم

همانطور که قبلا توضیح داده شد، درخت تصمیم یک مدل یادگیری نظارتی است که ساختاری مشابه یک درخت دارد و برای پیش‌بینی مقدار یک متغیر هدف استفاده می‌شود. این ساختار شامل گره‌ها شاخه‌هایی است که در هر گره، داده‌ها بر اساس یک ویژگی و مقدار مشخص تقسیم می‌شوند. این تقسیم‌ها به گونه‌ای انجام می‌شوند که داده‌ها به سمت گره‌های برگ با خلوص بیشتر هدایت شوند. فرآیند رشد درخت تا زمانی ادامه دارد که معیار توقف برآورده شود (مانند رسیدن به عمق مشخص یا تعداد نمونه کم در هر گره)

- مفهوم prune کردن در درخت های تصمیم گیری چیست؟ مزایا و معایب استفاده از این روش را ذکر کنید.
- به کمک تابع GridSearchCV یک مدل درخت تصمیم را آموزش داده و مقادیر بهینه برای پارامترها را بدست آورید. نحوه عملکرد این تابع را نیز به طور خلاصه توضیح دهید.
- درخت تصمیم نهایی خود را رسم کنید. برای این کار میتوانید از plot_tree استفاده کنید.)
- آیا در مدل شما underfitting یا overfitting رخ داده است؟ به طورکلی چه زمانی این پدیده رخ می دهد؟ هر یک را توضیح دهید. همچنین چه راهکارهایی برای کنترل آن برای مدل‌های درخت تصمیم وجود دارد؟
- برای مدل، خروجی معیارهای ارزیابی را با استفاده از داده‌های تست گزارش کنید و تحلیل کنید.
- ماتریس درهم‌ریختگی را برای مدل پیاده‌سازی‌شده برای داده‌های تست[۱] رسم نموده و تحلیل کنید.

[۱] test

در انجام این مینی‌پروژه حتماً به نکات زیر توجه کنید:

- موعد تحویل این تمرین، ساعت ۱۸:۰۰ روز پنجشنبه ۱۱ اردیبهشت‌ماه ۱۴۰۴ است.
- برای گزارش لازم است که پاسخ هر سوال و زیربخش‌هایش به‌ترتیب و به‌صورت مشخص نوشته شده باشند. بخش زیادی از نمره به توضیحات دقیق و تحلیل‌های کافی شما روی نتایج بستگی خواهد داشت.
- لازم است که در صفحهٔ اول گزارش خود لینک مخزن گیت‌هاب را و گوگل‌کولب مربوط به مینی‌پروژهٔ خود را درج کنید. درخصوص گیت‌هاب، یک مخزن خصوصی درست کنید و آی‌دی‌های `MJAHMADEE` و `AliBagheriNejad` را به‌عنوان Collaborator به مخزن اضافه کنید. پروژه‌های گیت‌هاب می‌بایست در انتهای ترم پابلیک شوند. درمقابل، لینک گوگل‌کولب را در حالتی که دسترسی عمومی دارد به اشتراک بگذارید. دفترچه‌کد گوگل‌کولب باید به‌صورت منظم و با بخش‌بندی مشخص تنظیم شده باشد و خروجی سلول‌های اجراشده قابل مشاهده باشد. در گیت‌هاب نیز یک مخزن برای درس و یک پوشهٔ مجزا برای هر مینی‌پروژه ایجاد کنید.
(آموزش پرایوت‌کردن مخزن گیتهاب و آموزش افزودن Collaborator به مخزن گیت‌هاب)
- هرجا از دفترچه‌کد گوگل‌کولب شما نیاز به فراخوانی فایلی خارج از محیط داشت، مطابق آموزش‌های ارائه‌شده ملزم هستید از دستور gdown استفاده کنید و مسیرهای فایل‌ها را طوری تنظیم کنید که صرفاً با اجرای سلول‌های کد، امکان فراخوانی و خواندن فایل‌ها توسط هر کاربری وجود داشته باشد.
- در تمامی مراحل تعریف داده و مدل و هرجای دیگری که مطابق آموزش‌های ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقم آخر شمارهٔ دانشجویی خود در نظر بگیرید.
- استفاده از ابزارهای هوشمند (مانند ChatGPT) در کمک‌گرفتن برای بهبود کدها مجاز است؛ اما لازم است تمام جزئیات مواردی که در خروجی‌های مختلف گزارش خود عنوان می‌کنید را به خوبی خوانده، درک و تحلیل کرده باشید. استفاده از این ابزارهای هوشمند در نوشتن گزارش و تحلیل‌ها ممنوع است.
- در جاهایی که با توجه به دو رقم آخر شمارهٔ دانشجویی خود محدود به انتخاب عدد، متغیر و یا داده‌ای خاص شده‌اید، برای تست‌های اضافه‌تر و نمایش بهبود در نتایج خود، مجاز هستید از مقادیر دیگر هم استفاده کنید.
- رعایت نکات بالا به حرفه‌ای‌ترشدن شما کمک خواهد کرد و اهمیتی معادل مطالب درسی فراگرفته‌شده دارد؛ بنابراین، درصورت عدم رعایت هریک از این نکات، از نمره تمرین شما **کسر خواهد شد**.
- آی‌دی پرسش هرگونه سوال درخصوص پرسش شمارهٔ 1
- آی‌دی پرسش هرگونه سوال درخصوص پرسش شمارهٔ 2
- آی‌دی پرسش هرگونه سوال درخصوص پرسش شمارهٔ 3