موضوع هفته:

آمار و ریاضی مقدماتی

تکلیف شماره:

۵

آخرین زمان تحویل:

جمعه، ۷ آذر ۱۴۰۴، ساعت 8:00 صبح

توضیح HW:

این تکلیف شامل دو قسمت ۱) تحلیلی/تحقیقی و ۲) برنامه‌نویسی می‌باشد.

در آخرین صفحه یک مساله تحلیلی برای شما قرار داده شده که جنبه تفریحی دارد و ارائه آن در قالب حل تمرین غیرمجاز هست.

با آرزوی موفقیت شما در حل مسائل

مدت زمان مفید احتمالی حل مسائل بر اساس توانایی شما: بین ۲ تا ۱۲ ساعت (ارزش: بسیار زیاد)

مدت زمان مفید احتمالی حل مسائل بر اساس هوش مصنوعی: بین ۱۰ تا ۲۴۰ ثانیه (ارزش: نمیدانم)

سوالات تحلیلی/تحقیقی

سوال ۱: در حوزه Machine Learning، مبحثی وجود دارد به اسم Curse of Dimensionality (COD)، که به دلیل اثر تعداد فیچرهای زیاد بر روی مدل‌ها، تحلیل داده‌ها توسط مدلها را تحت تاثیر قرار می‌دهد و به خصوص مدلهای مبتنی بر فاصله بیشتر نسبت به این مساله چالش دارند.

- اثر COD را بر روی فاصله اقلیدسی بررسی کنید (به عبارت دیگر، افزایش و کاهش ابعاد/فیچر چه تاثیری بر تعریف فاصله دارد؟).
- چه ابزار(ها)ی برای کاهش این دغدغه در حوزه تحلیل داده وجود دارند؟

سوال ۲:

(این سوال نمره ندارد ولی حل آن اجباریست. سعی کنید با منطق خودتان حل کنید هرچند که اشتباه باشد)

با توجه به اینکه یاد گرفتیم افزودن فیچر جدید به اطلاعات میتواند به ما کمک کند تا داده‌ها را بهتر تحلیل کنیم، فرض کنید به عنوان Machine Learning Engineer وظیفه دارین برای یک شرکت مرتبط با حوزه دامداری، موقعیت و فعالیت‌ها گاوهای دامداری را مانیتور کنید. برای اینکه بتوانید نوع فعالیت یک گاو را نشان دهید میتوانید از سنسورهایی (هر سنسور میشه یک فیچر) استفاده کنید. برای مثال ما میتوانیم از صد سنسور استفاده کنیم و هر کدام از این سنسورها را به قسمت هایی از این گاو وصل کنیم (مثل ۴ سنسور در ۴ پا، سر، گردن، شگم، دم، و ...) تا به صورت کامل فعالیت گاو را رصد کنیم. اما در واقعیت، هر سنسور اضافه هزینه‌ایی مضاعف ایجاد میکند و با توجه به بحث cost-benefit ترجیح شرکت‌ها بر این قاعده استوار هست که با کمترین تعداد سنسور (هزینه) بهترین عملکرد را داشته باشند.

به نظر شما، حداقل چند سنسور نیاز هست که بتوانیم با آن تشخیص دهیم گاو در حال انجام چه کاری هست و محل این سنسورها در چه قسمتی باشد؟

سوال ۳:

(این سوال نمره ندارد ولی حل آن اجباریست. سعی کنید با منطق خودتان حل کنید هرچند که اشتباه باشد)

در حوزه مالی و بخصوص بانکی، تشخیص کلاهبرداری یا Fraud Detection چالش خاص خودش را دارد که نیازمند تحلیل رفتار کاربران بر اساس تراکنش‌های مالی آنها و خارج کردن اطلاعات مهم برای تحلیل می‌باشد.

پایش عملکرد تعداد زیادی از اطلاعات خارج از توانایی بررسی دستی توسط افراد متخصص هست و به همین دلیل این روش نیازمند یک روش هوشمند اتوماتیک توسط ماشین می‌باشد.

اگر سیستم امنیت نتواند بدرستی عیب را تشخیص دهد، در نتیجه نتوانسته جلوی کلاهبرداری را بگیرد و به اصطلاح آماری Missed Real Alarm (False Negative-FN) رخ داده. از طرفی هم اگر سیستم امنیت روشش درست نباشد ممکن هست که تراکنش های طبیعی انجام شده توسط افراد سالم به عنوان تقلب نشان داده شوند که به آن Fake Alarm Generation (False Positive-FP) گفته می‌شود و هر دوی این نوع خطاهای FN, FP نشان از عملکرد ضعیف مدل طراحی شده در تصمیم گیری درست می‌باشد.

با علم به موارد گفته شده، قرار هست بر اساس آنچه که در دو جلسه قبلی از آمار مقدماتی یاد گرفته‌این و بر اساس اینکه در تیم امنیت پلیس باشین یا در تیم پروفشنال برای رسیدن به اهداف خودتون تلاش کنید. به عبارت دیگه ابتدا در تیم امنیت هستین و مدلی برای تشخیص تقلب طراحی می‌کنید. بعد در تیم پروفشنال هستین و میخواهید از دام‌های قبلی تیم امنیت عبور کنید. بر همین اساس، به خواسته های سوالات زیر بر اساس اینکه در هر گام در چه تیمی هستین پاسخ بدین (فرض بر این هست که گزینه رانت/پارتی وجود ندارد):

۱- **امنیت**: با داشتن میانگین بالانس حساب، چطور میشه از تقلب (کلاهبرداری) در بانک جلوگیری کرد؟
۲- **پروفشنال**: چطور میشه از تشخیص سیستم امنیت قبلی دور ماند و به پولشویی ادامه داد؟
۳- **امنیت**: چطور میشه افراد مخفی گام قبلی رو پیدا کرد؟
۴- **پروفشنال**: راهکار گمنام بودن در سیستم امنیتی گام‌های قبلی چیست تا به صورت موفقیت آمیزی جلو رفت؟
۵- اگر امنیت تنها از Range داده‌ها برای تشخیص تقلب استفاده کند چه مشکلاتی ایجاد میشود؟

سوالات برنامه‌نویسی

سوال ۴: تحلیل آماری کلاس‌های دیتاست iris

۱- بعد از فرآخوانی داده‌ها از iris_dataset.csv، برای هر کدام از کلاسها موارد زیر را محاسبه و در سه جدول مجزا نمایش دهید:

- میانگین
- انحراف معیار
- دامنه (رنج)

۲- ماتریس همبستگی 4x4 را برای چهار ویژگی در هر کلاس بدست بیاورید

۳- نمودار Heatmap این ماتریس همبستگی را رسم کنید

۴- کدام جفت ها همبستگی بیشتری دارند؟

۵- این همبستگی چه معنایی بین ویژگی ها دارد؟ (با در نظر گرفتن کم یا زیاد شدن مقدار یک ویژگی بر بقیه ویژگی‌ها)

سوال ۵: طبقه بندی ساده مبتنی بر نزدیکترین همسایه بر اساس مفهوم فاصله

۱- دیتاست iris_dataset.csv را لود کنید

۲- فقط دو ویژگی *Sepal Length* و *Sepal Width* را انتخاب کنید.

۳- **آخرین نمونه** از مجموعه داده (نمونه شماره ۱۴۹) را به عنوان **نقطه تست** $\mathrm{P_{test}}$ کنار بگذارید و بقیه داده‌ها را به عنوان **مجموعه آموزشی** در نظر بگیرید (کلاس واقعی را یادداشت کنید تا بعدا در آزمایش-ها آنرا چک کنید).

۴- برای یافتن نزدیک ترین همسایه در شرایط ($k = 1$)، فاصله اقلیدسی بین $\mathrm{P_{test}}$ و تک تک نقاط مجموعه آموزشی را محاسبه کنید. نقطه‌ای با **کمترین فاصله** را پیدا کنید و کلاس آن را به عنوان **کلاس پیش‌بینی شده** برای $\mathrm{P_{test}}$ در نظر بگیرید. آیا پیشبینی درست بود؟

۵- در این گام، سه نقطه ($k = 3$) با کمترین فاصله نسبت به $\mathrm{P_{test}}$ سوال ۴ را پیدا کنید. کلاس پیش بینی شده را برای $\mathrm{P_{test}}$ بر اساس رای اکثریت این سه همسایه تعیین کنید.

۶- توضیح دهید که چگونه افزایش k (تعداد همسایه) میتواند حساسیت مدل به نویز/اوتلایر را کاهش دهد؟

سوال ۶: طبقه بندی سوال قبلی با اثرگذاری مقیاس

۱- سوال شماره ۵ (یافتن نزدیکترین همسایه برای آخرین نمونه) را اینبار با در نظر گرفتن **هر ۴ ویژگی** را بدون هیچگونه نرمالسازی برای حالت $k = 3$ انجام دهید

۲- **تمام ۴ ویژگی** دیتاست iris را با استفاده از نرمالسازی z-score که به صورت زیر انجام میشود نرمالیزه کنید:

$$Z = \frac{x - \mu}{\sigma}$$

در اینجا، μ و σ برای هر ستون بر اساس کل داده های آموزش استفاده شود (و نه بر اساس داده‌های هر کلاس).

در این حالت، مجددا نزدیک ترین همسایه را بر اساس k=3 انجام دهید.

۳- دقت پیش‌بینی (درست یا غلط بودن) در حالت **نرمال‌نشده** را با حالت **نرمال‌شده** مقایسه کنید.

سوال ۷: دسته بندی اطلاعات iris با روش های مبتنی بر میانگین و واریانس

۱- ابتدا داده های iris_dataset.csv رو فراخوانی کنید

۲- ۴ فیچر اول را به عنوان ورودی (X) در نظر بگیرید و فیچر خروجی را Y (لیبل هر کلاس) در نظر بگیرید

۳- مقدار میانگین، واریانس، range، max و min برا برای هر فیچر در کلاسها بدست بیاورید

۴- با در نظر گرفتن تنها اطلاعات (میانگین و واریانس) مرتبط به فیچر سوم ((petal length (cm)) برای هر کلاس، داده های petal length (cm) از فایل iris_dataset.csv را با توجه به مقدار احتمالشان نسبت به کلاس‌های تشکیل شده به عضویت هر کلاس در بیاورید (میتوانید از احتمال ساخته شده از میانگین و واریانس استفاده کنید که در پایین سوال آورده شده است). کلاس ها را بر اساس رنگشان به صورت زیر نمایش دهید:

A=blue | B=black | C=red

۵- کلاس‌هایی که در گام قبلی درست کردین، دارای چه دقتی هستند؟ برای محاسبه دقت باید از اطلاعات فیچر خروجی Y استفاده کنید.

$$\text{Precision} = \frac{\text{Correct Answers}}{\text{Correct Answers} + \text{ Wrong Answers}} \times 100$$

۶- با در نظر گرفتن تنها اطلاعات (میانگین و واریانس) مرتبط به فیچر سوم ((petal length (cm)) برای هر کلاس، داده های تست را از فایل iris_test_samples.csv بگیرید و به عضویت هر کلاس در بیاورید (میتوانید از احتمال ساخته شده از میانگین و واریانس استفاده کنید که در پایین سوال آورده شده است).

۷- دقت تصمیم گیری چقدر هست؟

۸- علاوه بر فیچر شماره ۳، از فیچر شماره ۴ ((petal width (cm)) هم استفاده کنید. در اینجا، برای هر فیچر یک احتمال وجود دارد (پس برای دو فیچر، دو احتمال). میانگین این دو احتمال را به عنوان احتمال محاسبه شده برای هر کلاس استفاده کنید و دسته بندی جدیدی را انجام دهید.

۹- دقت تصمیم گیری چقدر هست؟

۱۰- از تمام ۴ فیچر ورودی استفاده کنید و سپس، دو احتمال برتر از بین چهار احتمال را در نظر بگیرید برای تصمیم گیری. بعد از انتخاب دو احتمال برتر، از آنها میانگین گرفته، و این میانگین را در بین کلاس ها مقایسه کنید تا تشخیص دهید داده های تست به چه کلاسی مرتبط می‌شوند.

(راهنمایی: برای محاسبه احتمال داده x برای برای هر کلاس، از تعریف تابع زیر برای هر کلاس با داشتن مقادر میانگین μ و مقدار واریانس σ^2 آن کلاس به صورت زیر استفاده کنید (اگر از روش مبتنی بر پکیج برای بدست آوردن مقدار احتمال استفاده کرده اید، اشکالی ندارد، اما توصیه میشود برای یادگیری بهتر از ساختن تابع احتمالی زیر با محاسبه خودتون استفاده کنید تا توانایی نوشتن کد برای شما تقویت شود):

$$P(x|Class) = \frac{1}{\sqrt{2\pi} \times \sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$P(x|Class_A)$ یعنی مقدار احتمال نقطه x برای کلاس A. به صورت ریاضی میشه: مقدار احتمالی داده x به شرطی که در کلاس A باشد. شما میتوانید با نوشتن این فرمول به صورت تابع (که در مقدماتی پایتون یادگرفتین) برای بدست آوردن این احتمال استفاده کنید. به عبارت دیگر، میانگین و واریانس رو دارین، با گرفتن داده میتوانید مقدار احتمال را برای آن دسته بدست بیاورید. هر کلاس میانگین و واریانس خودش را دارد.)

از نمونه سوالات تحلیلی در زمان مصاحبه یک شرکت با نیروی جدید

- با داشتن دو ظرف ۵ لیتری و ۳ لیتری چطور میتوان ۴ لیتر آب را بطور دقیق بدست آورد (با فرض نداشتن سایر وسایل اندازه گیری یا کمکی).

{نکته: شیر آب باز هست و میتوان ظرف‌ها هر زمان که نیاز بود خالی و پر کرد. همینطور، ۴ لیتر باید در ظرفها قرار بگیرد. با فرض داشتن استرس در زمان مصاحبه و زمان محدود ۵ دقیقه‌ایی در پاسخ‌گویی به این مساله فکر کنید.}