به نام خدا

درس تشخیص و شناسایی عیب

تمرین سری دوم

موعد تحویل: ۱۴۰۲/۰۱/۳۱

دانشگاه صنعتی خواجه نصیرالدین طوسی

آزمایشگاه تشخیص و شناسایی خطا

سوال شمارهٔ یک – ۱۰ نمره

۱-الف) برای متغیرهای تصادفی $\{x_1, x_2, \dots, x_n\}$، با استفاده از روش تخمین Maximum Likelihood و در صورت امکان پارامتر مجهول z در هر توزیع آورده‌شده در موارد زیر را به دست آورید و در غیر این صورت دلیل خود را ذکر کنید.

$$f(x \mid z) = \begin{cases} \frac{1}{z}, & \text{for } 0 \le x \le z \\ 0, & \text{otherwise} \end{cases}$$

۱-ب) متغیر تصادفی نرمال X دارای توزیع نرمال $N(\mu, \sigma^2)$ با پارامتر مجهول μ می‌باشد. چنانچه بدانیم μ از توزیع Rayleigh بصورت زیر پیروی می‌کند:

$$p(\mu) = \frac{\mu e^{-\left(\frac{\mu^2}{2\sigma_\mu^2}\right)}}{\sigma_\mu^2}$$

نشان دهید که تخمین MAP پارامتر μ از رابطه زیر بدست می‌آید:

$$\hat{\mu}_{MAP} = \frac{Z}{2R}\left(1 + \sqrt{1 + \frac{4R}{z^2}}\right); \qquad Z = \frac{1}{\sigma^2}\sum_{k=1}^{N} x_k \,, R = \frac{N}{\sigma^2} + \frac{1}{\sigma_\mu^2}$$

۱-ج-امتیازی) توضیح دهید که تحت چه شرایطی ML و MAP با هم کاملاً مشابه هستند.

سوال شمارهٔ دو – ۱۵ نمره

۲-الف) ثابت کنید که روش تخمین احتمال غیرپارامتری ($P \approx k_N/N$)، یک روش تخمین بدون بایاس است.

۲-ب) میزان بایاس روش تخمین پارزن را محاسبه کنید و تعیین کنید که تحت چه شرایطی این روش، یک روش تخمین بدون بایاس خواهد بود.

۲-ج) توضیح دهید که چرا هستهٔ روش پارزن از شمارندهٔ سادهٔ $\phi(x) = \begin{cases} 1 & \text{for } |x| \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$ به توابع گوسی تمایل می‌یابد.

۲-د-امتیازی) میزان واریانس روش تخمین پارزن را محاسبه کنید و تعیین کنید که تحت چه شرایطی این روش، یک روش تخمین بدون واریانس خواهد بود.

۲-هـ-امتیازی) روش‌های غیرپارامتری روش‌های دشواری بوده و از دقت کمی برخوردار هستند. به نظر شما علت استفاده از آن‌ها در مقابل روش‌های پارامتری چیست؟ نقطهٔ ضعف روش تخمین پارزن را بیان کرده و راه‌حلی برای آن ارائه کنید.

سوال شمارهٔ سه – ۸۵ نمره

۳- مجموعه‌داده **DAMADICS** از تمرین سری اول را به‌همراه آموزش مقدماتی‌ای که درخصوص برخی دستورات در محیط پایتون و گوگل‌کولب دیده‌اید به‌خاطر بیاورید.

۳-الف) همان‌طور که می‌دانید مجموعه‌داده از طریق این لینک در دسترس است. هم‌چنین نتیجهٔ پیش‌پردازش‌های تبدیلی مربوط به تمرین سری اول از طریق این لینک در دسترس است (درصورت نیاز). در قسمت اول این سوال با چند فعالیت پیش‌پردازشی دیگر آشنا می‌شوید.

قبل از هرچیز، فایل csv مربوط به داده‌های تاریخ ۱۷ نوامبر ۲۰۰۱ را روی گوگل‌درایو خود بارگذاری کنید و یک لینک مستقیم با دسترسی عمومی به‌صورتی که در زیر نشان داده شده است ایجاد کنید و لینک آن را کپی کنید. با انجام این‌کار فایل مربوطه از طریق لینکی شبیه به لینک زیر دردسترس خواهد بود:

https://drive.google.com/file/d/1-GFtU-DPbWDsAJ-hckTe0dA2KikfwubJ/view?usp=sharing

قسمتی که در لینک بالا به رنگ قرمز درآمده است ID یا شناسهٔ فایل شماست. در ادامه دفترچه‌کدی در محیط گوگل کولب ایجاد کنید(New notebook). با استفاده از دستور زیر فایل csv مربوطه را در محیط گوگل‌کولب فراخوانی کنید:

```
!pip install --upgrade --no-cache-dir gdown
!gdown ID
```

قسمت ID موجود در کد بالا را با ID فایل خود جایگزین کنید.

۳-الف-۱) فایل csv فراخوانی‌شده را بخوانید و سپس تابع Info. را از Pandas فراخوانی کنید و اطلاعاتی که می‌توانید از آن کسب کنید را بنویسید. اشاره کنید که اگر فایل داده‌ها دارای header بود، کدام قسمت از نتیجه تغییر پیدا می‌کرد.

۳-الف-۲) با استفاده از دستوراتی مانند isnull در پایتون، تعداد داده‌هایی که در هر ستون Nan هستند را نمایش دهید. یکی از راه‌های تغییر و بهبود ستون‌های دارای دادهٔ Nan را بنویسید.

۳-الف-۲) Correlation Matrix را رسم کنید و بررسی کنید که کدام ویژگی (ستون) با ویژگی ششم که سرستون آن عدد 33.1 می‌باشد، بیش‌ترین Correlation را دارد.

۳-الف-۳) برای نرمال‌سازی داده‌ها می‌توان از منطق مقیاس‌کردن داده‌ها بین ۰و۱، ۱-و۱، و یا نرمال‌سازی به‌صورت توزیع نرمال استفاده کرد. یکی از این منطق‌ها را انتخاب کنید و ضمن توضیح دلیل انتخاب آن، داده‌های هر ستون را نرمال کنید.

۳-الف-۴-امتیازی) مطابق تمرین سری اول، داده‌های روز ۹ نوامبر ۲۰۰۱ را به‌عنوان داده‌های test در نظر بگیرید و داده‌های train و test را با کمک دستوراتی مانند MinMaxScaler، مقیاس (Scale) کنید. تحقیق کنید که اگر از داده‌های test برای scaleکردن مجموعه‌داده استفاده کنید چه مشکلی پیش می‌آید.

۳-ب) داده‌های تاریخ ۱۷ نوامبر ۲۰۰۱ را در نظر بگیرید. عیوبی در گام‌های زمانی مشخص و مطابق جدول زیر به سیستم وارد می‌شوند:

Item	Fault	Sample	Date
1	f18	54600 - 54700	November 17,2001
2	f16	56670 - 56770	November 17,2001
3	f17	53780 - 53794	November 17,2001
4	f17	54193 - 54215	November 17,2001
5	f19	55482 - 55517	November 17,2001
6	f19	55977 - 56015	November 17, 2001
7	f19	57030 - 57072	November 17, 2001
8	f16	57475 - 57530	November 17, 2001
9	f16	57675 - 57800	November 17, 2001
10	f19	58150 - 58325	November 17, 2001

مشابه تمرین سری اول، داده‌ها را در پنج دسته عادی (سالم)، عیب۱ (f16)، عیب۲ (f17)، عیب۳ (f18) و عیب۴ (f19) در نظر بگیرید (ترتیب و شماره دسته‌ها به انتخاب شماست). برنامه‌ای بنویسید که با استفاده از روش غیرپارامتری پارزن و هسته‌های مربعی، مثلثی و گوسی داده‌ها را دسته‌بندی کند. با تغییر طول پنجرهٔ پارزن در محدوده‌ای مشخص و با گامی مشخص، مقدار بهینهٔ h رامحاسبه کنید و نتایج را به‌صورت ماتریس درهم‌ریختگی، نمودار ROC و حداقل سه شاخص فراگیری‌شده در درس نشان دهید. نتایج مربوط به هر نوع هسته را مقایسه و تحلیل کنید.

۳-ج) سوال «۳-ب» را با روش غیرپارامتری KNN و حداقل برای سه مقدار متفاوت k که تفاوت معناداری ایجاد کند تکرار کنید و سعی کنید با تغییر k در محدوده‌ای مشخص و با گامی مشخص، مقدار بهینهٔ k را به‌دست آورید. در صورت امکان (امتیازی) نتیجهٔ خود را به‌صورت یک شاخص ارزیابی که با تغییر k، عوض می‌شود در قالب یک نمودار نشان دهید.

۳- د) یک طبقه‌بند خطی چنددسته‌ای را به این صورت که هر دسته را از دسته‌ای غیرمعادل با خود جدا می‌کند در نظر بگیرید:

۳- د-۱) با فرض WZ>0، W را طوری آموزش دهید که خطای آموزش حداقل شود. ماتریس درهم‌ریختگی را هم به‌دست آورده و نتایج را تحلیل کنید.

۳- د-۲) با فرض WZ=b ($b \in [0,1]$ معلوم)، در ابتدا W را با استفاده از روش LS محاسبه کنید و سپس ماتریس درهم‌ریختگی را نمایش دهید. آیا تغییر b تأثیری در جواب به‌دست‌آمده دارد؟

۳- د-۳) W را با استفاده از روش Sequential Widrowf-Hoff و با فرض $b \in [0,1]$ معلوم، محاسبه کنید و ماتریس درهم‌ریختگی را نمایش دهید. تأثیر تغییر گام آموزش و پارامتر b را بررسی کنید.

۳-د-۴) حال اگر در قسمت قبل به‌جای ثابت فرض کردن b، آن را با استفاده از گرادیان نزولی آموزش دهیم نتایج چگونه خواهد بود؟

۳-هـ-امتیازی) یکی از روش‌های خواسته‌شده در سوالات قبل را انتخاب کنید و نتایج را در دو حالت که داده‌ها نرمال شده باشند (بر اساس سوال «۳-الف-۳») و نرمال نشده باشند، را مقایسه و تحلیل کنید.

۳-و-امتیازی) ساختار پیاده‌سازی‌شده در سوال «۳-د» را در نظر بگیرید. با استفاده از داده‌های روز ۹ نوامبر ۲۰۰۱ به‌عنوان داده تست، عمل‌کرد ساختارهای طراحی‌شدهٔ در سوالات قبلی را آزمایش و نتیجه را گزارش و تحلیل کنید.

Item	Fault	Sample	Date
1	f16	57275 - 57550	November 9, 2001
2	f18	58830 - 58930	November 9, 2001
3	f18	58520 - 58625	November 9, 2001
4	f16	60650 - 60700	November 9, 2001
5	f16	60870 - 60960	November 9, 2001

۳-ز-امتیازی) سوال «۳-ب» را با روش SVM تکرار کنید و سعی کنید با تغییر σ^2 در محدوده‌ای مشخص و با گامی مشخص، مقدار بهینهٔ آن را به‌دست آورید.

ضمن عرض سلام و خداقوت، لطفاً برای ارسال تمرین‌ها به نکات زیر توجه فرمایید:

نتایج خود را به‌صورت کامل توضیح دهید و شکل یا نتیجه‌ای را بدون ارائهٔ توضیح و تحلیل رها نکنید.

نوشتن تمرین به‌صورت مرتب و در قالب LaTeX مدنظر، درصدی متغیر به نمرهٔ شما اضافه خواهد کرد.

هر مطلبی درخصوص تمرین‌ها و نمرات را از طریق رایانامهٔ faultcourse@gmail.com، در میان بگذارید.

در سوالاتی که از شما تحلیل خواسته شده، عمق تحلیل‌ها و بحث روی حالات و علل مختلف اهمیت بالایی دارد.

هر تمرین ۱۰۰ نمره دارد و در صورت تأخیر، 2^n نمره (n تعداد روز تأخیر) از مجموع نمرهٔ تمرین شما کاسته خواهد شد.

«در مسیر حق موفق و پیروز باشید.»