دانشگاه تهران

به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری اول یادگیری ماشین

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
2. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
3. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.
4. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.
5. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW#_StudentNumber داشته باشد.
6. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.
7. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.
8. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.
9. در صورت داشتن سوال، از طریق ایمیل taheriarmin60@gmail.com سوال خود را مطرح کنید.

سوال ۱: (۱۵ نمره)

تابع توزیع کوشی[1] را برای یک مسئله طبقه‌بندی دو کلاسه و یک بعدی در نظر بگیرید:

$$P(x|\omega_i) = \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_i}{b}\right)^2} \qquad i = 1,2 \qquad a_2 > a_1$$

آ) با فرض $P(\omega_1) = P(\omega_2)$، نشان دهید $P(\omega_1|x) = P(\omega_2|x)$ اگر $x = \frac{a_1+a_2}{2}$. به کمک متلب یا پایتون $P(\omega_1|x)$ و $P(\omega_2|x)$ روی یک axis رسم کنید.$(a_1 = 3, a_2 = 5, b = 1)$

ب) نشان دهید که حداقل احتمال خطا برابر است با:

$$P(error) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_2 - a_1}{2b}\right|$$

پ) بیش‌ترین مقدار $p(error)$ چیست و تحت چه شرایطی اتفاق می‌افتد؟

ت) یک طبقه‌بند بهینه بیزی طراحی کنید بر اساس a_i, b اگر $P(\omega_1) = P(\omega_2)$. مرز تصمیم[2] را برای این حالت رسم کنید و میزان احتمال خطا را گزارش کنید.

ث) یک طبقه‌بند بیزی برای کمینه کردن ریسک با وزن‌های زیر طراحی کنید:

$$\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{pmatrix} = \begin{pmatrix}0 & 1\\ 2 & 0\end{pmatrix}$$

مرز تصمیم را رسم کنید و احتمال خطا را گزارش کنید. نتایج حاصل از این بخش را با نتایج بخش قبل مقایسه کنید.

[1] Cauchy distribution
[2] Decision boundary

سوال ۲: (۱۰ نمره)

یک مسئله کمینه‌کردن ریسک در حالت دو کلاسه را در نظر بگیرید. فرض کنید $\lambda_{ii} \neq 0$

آ) ناحیه تصمیم مربوط به کلاس اول را محاسبه کنید.

ب) حد بالای $\frac{f(x|\omega_2)}{f(x|\omega_1)}$ را محاسبه کنید.(راهنمایی: از نامساوی‌ای که در بخش آ به دست می‌آید استفاده کنید.)

سوال ۳: (۱۰ نمره)

یک مسئله دو کلاسه یک بعدی با توزیع رایلی[3] برای هر دو کلاس را در نظر بگیرید:

$$P(x|\omega_i) = \begin{cases} \frac{x}{\sigma_i^2} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

با فرض یکسان بودن توزیع پیشین[4] برای هر دو کلاس، مرز تصمیم را محاسبه نمایید.

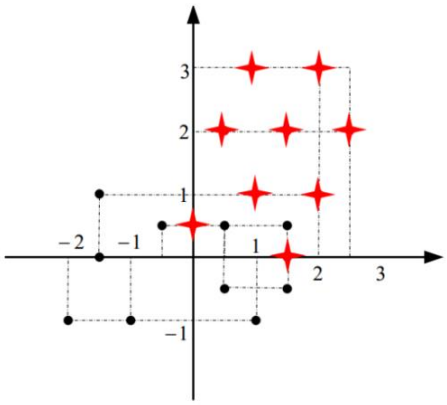
[3] Rayleigh distribution
[4] Prior

سوال ۴: (۱۰ نمره)

در نظر بگیرید دو تابع توزیع نرمال با میانگین متفاوت و واریانس یکسان داریم. با در نظر گرفتن احتمال پیشین $p(\omega_1)$ و $p(\omega_2)$ بیان کنید با چه شرطی مرز تصمیم بیز بین دو میانگین قرار نمی‌گیرد.

سوال ۵: (۱۵ نمره)

شکل زیر را در نظر بگیرید.



1 Figure

آ) میانگین و کواریانس را برای توزیع گوسی هر دو کلاس بیابید.

ب) فرض کنید احتمال پیشین برای هر دو کلاس برابر با 0.5 باشد. مرز تصمیم را بیابید و رسم کنید. خطای آموزش تجربی را روی این داده‌ها محاسبه کنید.(مثلا درصد نقاطی که اشتباه طبقه‌بندی شده‌اند.)

پ) مرز تصمیم را برای یک طبقه‌بند بیزی برای کاهش ریسک با مقادیر زیر بیابید و رسم کنید.

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 2a \\ a & 0 \end{pmatrix} \quad a > 0$$

ت) برای مقادیر $P(\omega_1) = \frac{1}{3}, P(\omega_2) = \frac{2}{3}$، بخش پ را تکرار کنید.

سوال ۶: (۱۵ نمره)

متغیر تصادفی X را با توزیع احتمال پواسون[5] با پارامتر λ در نظر بگیرید:

$$P(X) = \frac{\lambda^x e^{-\lambda}}{x!}$$

فرض کنید یک مجموعه داده شامل n نمونه $D = \{X_1, \dots, X_n\}$ از این متغیر تصادفی در اختیار داریم.

آ) تابع لگاریتم درست‌نمایی[6] را تشکیل داده و تخمین‌گر بیشینه درست‌نمایی[7] را برای پارامتر λ به دست آورید.

ب) توزیع احتمال پیشین زیر را برای پارامتر λ در نظر بگیرید:

$$P(\lambda) = Gamma(\lambda|\alpha, \beta) = c\lambda^{\alpha-1}e^{-\beta\lambda}$$

که در رابطه بالا، c یک ضریب ثابت است و α و β پارامترهای توزیع گاما هستند. توزیع احتمال پسین را برای پارامتر λ به دست آورید.

$$P(\lambda|D) = ?$$

پ) آیا توزیع احتمال پیشین فوق، برای پارامتر λ یک *conjugate prior* است؟ توضیح دهید.

ت) با استفاده از توزیع احتمال پیشین فوق، تخمین‌گر *MAP* برای پارامتر λ چیست؟(راهنمایی: مقدار بیشینه توزیع گاما در نقطه $\lambda = \frac{\alpha-1}{\beta}$ رخ می‌دهد).

ث) آیا اگر تعداد داده‌ها به بی‌نهایت میل کند، تخمین‌گر *MAP* به *MLE* میل می‌کند؟ توضیح دهید.

ج) توضیح دهید در چه شرایطی استفاده از هر کدام از این دو روش تخمین بر دیگری برتری دارد.

---

[5] Poisson
[6] Log likelihood
[7] Maximum likelihood

سوال ۷: (**شبیه سازی**، ۲۵ نمره)

هدف از این سوال آشنایی و پیاده سازی طبقه‌بند naïve bayes است.

آ) در ابتدا در مورد طبقه‌بند naïve bayes توضیح دهید و تفاوت ساختاری آن‌را با یک طبقه‌بند بیزی بیان کنید. توضیح دهید که چرا به جای طبقه‌بند بیز از این طبقه‌بند استفاده می‌کنیم، هزینه‌ای که می‌دهیم چیست و در چه زمان‌هایی استفاده از این طبقه‌بند کاری منطقی است.

مجموعه داده lung cancer به پیوست ارسال شده. توضیحات مربوط به این مجموعه داده در لینک زیر وجود دارد. در ابتدا لینک زیر را مطالعه کنید.

https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

در مورد پیش‌پردازش‌های معمول قبل از استفاده از داده‌های خام تحقیق کنید. با ذکر دلیل بیان کنید انجام چه پیش‌پردازش‌هایی روی داده‌های این سوال به مسئله کمک می‌کند و این پیش‌پردازش‌ها را اعمال کنید.(حتما از داده‌های ارسالی استفاده کنید و داده‌های مشابه را از لینک بالا استفاده نکنید.)

ب) یک طبقه‌بند naïve bayes را از پایه و بدون استفاده از کتاب‌خانه پیاده‌سازی کنید. از طبقه‌بندی که طراحی کردید استفاده کنید. دقت، precision، Recall و ماتریس آشفتگی[8] را بررسی و تحلیل نمایید.

پ) مورد ب را به کمک کتاب‌خانه SKLEARN انجام دهید. نتایج دو بخش را مقایسه کنید.

در صورتی که عملکرد مدل naïve bayes مناسب نبود علت را شرح دهید.

ت) موارد بالا را روی مجموعه داده web page phishing که به پیوست ارسال شده تکرار نمایید. توضیحات مربوط به این مجموعه داده در لینک زیر قرار داده شده.

https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset

عملکرد مدل را روی دو مجموعه داده مقایسه کنید. در صورت تفاوت به تحلیل علت آن بپردازید.

[8] Confusion Matrix

سوال ۸: (**شبیه سازی**، ۱۰ نمره)

تصاویری از آسمان ابری و آسمان در زمان غروب خورشید در نواحی مختلف گرفته شده. در این سوال می‌خواهیم تصاویر آسمان در این دوحالت را جداسازی کنیم. تصاویر در فایل Images قرار گرفته. تصاویر مربوط به غروب خورشید با برچسب S و تصاویر مربوط به آسمان ابری با برچسب C مشخص شده‌اند. معیاری برای جداسازی تصاویر تعریف کنید و آن‌را پیاده‌سازی کنید. الگوریتم پیاده‌سازی را روی داده‌ها تست نمایید و دقت، ماتریس آشفتگی، Precision و Recall را گزارش کنید.(برای طبقه‌بندی از طبقه‌بند معروفی استفاده نکنید صرفا از ویژگی‌های داده مانند رنگ برای جداسازی استفاده نمایید).

سعی کنید معیاری تعریف کنید که به دقت بهتر برسید. داده‌هایی که به اشتباه جداسازی شدند را معرفی کنید و علت اشتباه مدل را بر اساس معیاری که تعریف کردید تحلیل کنید. اگر چند معیار مختلف را تست کردید دقت آن‌ها و همچنین تصاویر غلط در هر کدام را مقایسه نمایید.