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: (۲۰ نمره)

در یک مسئله طبقه‌بندی چند کلاسه:

الف) نشان دهید که تصمیم‌گیری به کمک روش Bayes احتمال خطا را کمینه می‌کند.

ب) ثابت کنید اگر M کلاس داشته باشیم، حد بالای خطا به صورت $p_e \leq \frac{M-1}{M}$ خواهد بود.

ج) راهی برای رسم نمودار ROC در حالت چند کلاسه پیشنهاد کنید.

د) توضیح دهید که در چه مجموعه داده‌هایی naïve Bayes عملکرد بهینه خواهد داشت. علت را به تفصیل شرح دهید.

سوال ۲: (۱۰ نمره)

یک طبقه‌بند دو کلاسه با احتمال پیشین مساوی را در نظر بگیرید. فرض کنید داده‌های دو کلاس بر اساس توزیع‌های زیر تولید می‌شوند:

$$p(x|y=1) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right) \quad x \geq 0$$
$$p(x|y=2) = \theta x exp(-\theta x) \quad x \leq 0$$

که $\sigma > 0$ و $\theta > 0$ پارامترهای مدل هستند. ناحیه مربوط به دو کلاس را در طبقه‌بند بیز به دست آورید.

سوال ۳: (۱۰ نمره)

ماتریس ریسک زیر را در نظر بگیرید.

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

الف) نشان دهید با این شرایط مرز تصمیم شرط زیر را ارضا می‌کند.

$$\int_{R_2} p(x|\omega_1)\, dx = \int_{R_1} p(x|\omega_2)\, dx$$

ب)آیا این پاسخ همواره یکتاست؟ در غیر این صورت یک مثال نقض بزنید.

سوال ۴: (۲۰ نمره)

متغیر تصادفی x از توزیع $N(\mu, \sigma^2)$ است که pdf مربوط به پارامتر μ به شکل زیر است:

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

نشان دهید که تخمین MAP پارامتر μ برابر است با:

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

سوال ۵ : (۱۰ نمره)

فرض کنید تابع چگالی احتمال متغیر تصادفی Y به شکل زیر است.

$$f_Y(y|\theta) = \begin{cases} \frac{1}{\theta} r y^{r-1} e^{-\frac{y^r}{\theta}}, & \theta > 0, y > 0 \\ 0, & elsewhere \end{cases}$$

که r یک ثابت مثبت است.

الف) تابع log likelihood را به دست بیاورید.

ب) توضیح دهید در چه حالتی و چرا تخمین‌گر MAP به ML میل می‌کند.

سوال ۶ : (**شبیه سازی**، ۲۰ 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 Breast Cancer Wisconsin را از لینک زیر دانلود کنید.

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

توضیحات مربوط به این مجموعه داده به طور کامل در صفحه فوق وجود دارد؛ لطفا قبل از شروع به انجام تمرین توضیحات را مطالعه نمایید.

در ابتدا در صورت نیاز روی داده‌ها پیش‌پردازش انجام دهید.(هر پیش‌پردازشی که روی داده‌ها انجام می‌دهید را باید با ذکر دلیل توضیح دهید).

ب) این مجموعه داده شامل دو کلاس است. یک طبقه‌بند naïve bayes را از پایه و بدون استفاده از کتاب‌خانه پیاده‌سازی کنید. و طبقه‌بندی که طراحی کردید استفاده کنید. دقت، precision، Recall و ماتریس آشفتگی[1] را بررسی و تحلیل نمایید.

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

[1] Confusion Matrix

سوال ۷ : (**شبیه سازی**، ۲۰ نمره)

یک طبقه‌بند دو کلاسه برای تشخیص تصاویر مربوط به دریا و جنگل در مجموعه داده image طراحی کنید. الگوریتم پیاده‌سازی را روی داده‌ها تست نمایید و دقت، ماتریس آشفتگی، Precision و Recall را گزارش کنید.(راهنمایی: برای طبقه‌بندی نیازی به استفاده از طبقه‌بند معروفی نیست صرفا از ویژگی‌های داده مانند رنگ برای جداسازی استفاده نمایید).

داده‌هایی که به اشتباه جداسازی شدند را معرفی کنید و بیان کنید با توجه به ویژگی‌ای که بر طبق آن جداسازی انجام شده آیا این اشتباهات منطقی است یا خیر.