سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.

۲. 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.

۳. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.

۴. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

۵. نمره تمرین از ۱۰۰ نمره می‌باشد

۶. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین ۱۰۰- خواهد شد.

۷. در صورتی که تشخیص داده شود از چت بات ها به صورت مستقیم برای پاسخ سوال های تئوری و شبیه سازی استفاده شده است، نمره ۱۰۰- در نظر گرفته خواهد شد.

۸. 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.

۹. در صورت داشتن سوال، از طریق گروه درس یا ایمیل‌های زیر با تدریسیار مربوطه سوال‌های خود را مطرح کنید.

سوال های ۱ و ۲ و ۳ و ۴ و ۸: mahdavijoosaba@gmail.com

سوال های ۵ و ۶ و ۷: smousavichashmi@ut.ac.ir

سوال ۱: (۴ نمره)

انواع مختلف مسائل یادگیری ماشین را در نظر بگیرید. مشخص کنید که وظایف زیر شامل یادگیری نظارت‌شده[1] است یا بدون نظارت[2]. برای مسائل یادگیری نظارت‌شده، تعیین کنید که آیا این وظایف در دسته رگرسیون[3]، طبقه‌بندی[4] یا طبقه‌بندی احتمالاتی[5] قرار می‌گیرند.

الف) پیش‌بینی ریسک تصادف در یک تقاطع با توجه به ویژگی‌هایی مانند زمان روز و آب‌وهوا.

ب) شناسایی خودروها، دوچرخه‌سوارها و عابرین پیاده در ویدیویی که توسط دوربین‌های یک خودروی خودران گرفته شده است.

ج) تعیین احتمال وجود تابلوی ایست در یک تصویر.

د) تولید سناریوهای جدید جاده‌ای (تولید خیابان‌ها، قرار دادن تابلوی ایست و تقاطع‌ها) برای آزمایش خودروهای خودران در یک شبیه‌ساز.

[1] supervised learning
[2] unsupervised learning
[3] regression
[4] classification
[5] probabilistic classification

سوال ۲: (۱۰ نمره)

در بسیاری از مسائل دسته‌بندی الگو، گزینه‌ای وجود دارد که الگو را به یکی از c کلاس تخصیص دهیم یا آن را **رد** کنیم، به این معنی که غیرقابل تشخیص در نظر گرفته شود. اگر هزینه رد کردن خیلی زیاد نباشد، رد کردن می‌تواند یک اقدام مطلوب باشد. تابع زیان را به‌صورت زیر تعریف می‌کنیم:

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \quad i,j = 1,\dots,c \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise,} \end{cases}$$

که در آن، λ_r هزینه‌ای است که در اثر انتخاب عمل $c+1$ ام، یعنی رد کردن، متحمل می‌شویم، و λ_s هزینه‌ای است که در اثر یک خطای جایگزینی رخ می‌دهد. نشان دهید که حداقل ریسک[6] زمانی به دست می‌آید که تصمیم بگیریم ω_i را انتخاب کنیم اگر $P(w_i|x) \geq P(w_j|x)$ برای تمام j ها برقرار باشد و اگر $P(\omega_i \mid \mathrm{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$ باشد. در غیر این صورت، باید الگو را رد کنیم.

[6] minimum risk

سوال ۳: (۱۰ نمره)

یک مسئله طبقه‌بندی دو کلاسه را در نظر بگیرید که در آن ورودی دو بعد دارد:

$$y \in \{0,1\}, \mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

برای جداسازی دو کلاس می‌خواهیم از طبقه‌بند بیز ساده[7] استفاده کنیم. اگر توزیع احتمال ویژگی‌ها به ازای هر کلاس به صورت زیر باشد، رابطه مرز جداکننده دو کلاس را بدست آورید. این مرز معادل کدامیک از نمودارهای شناخته شده است؟ نمودار مرز جداساز را در صفحه (x_1, x_2)ترسیم نمایید و نشان دهید کدام ناحیه از فضا مربوط به کلاس $y = 0$ و کدام ناحیه مربوط به کلاس $y = 1$ است. (احتمال پیشین دو کلاس را مساوی در نظر بگیرید.)

$$p(x_1 \mid y = 0) = 2e^{-2x_1}, \quad x_1 \geq 0$$

$$p(x_2 \mid y = 0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_2-1)^2}, \quad x_2 \in \mathbb{R}$$

$$p(x_1 \mid y = 1) = e^{-x_1}, \quad x_1 \geq 0$$

$$p(x_2 \mid y = 1) = \frac{1}{\sqrt{\pi/2}} e^{-2(x_2-\frac{1}{2})^2}, \quad x_2 \in \mathbb{R}$$

[7] naïve bayes

سوال ۴: (۱۶ نمره)

توزیع پواسون یک توزیع گسسته مفید است که می‌تواند برای مدل‌سازی تعداد وقوعات یک رویداد در واحد زمان استفاده شود. برای مثال، در شبکه‌ها ، تعداد ورود بسته‌ها در یک پنجره زمانی مشخص اغلب با توزیع پواسون مدل‌سازی می‌شود. اگر X دارای توزیع پواسون باشد، یعنی $X \sim Poisson(\lambda)$ ، تابع جرم احتمال آن به صورت زیر است:

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

هم چنین می‌توان نشان داد که امید ریاضی X برابر است با:

$$\mathbb{E}(X) = \lambda$$

فرض کنید اکنون n مشاهده مستقل و هم‌توزیع (i.i.d.) از توزیع $Poisson(\lambda)$ داریم:

$$\mathcal{D} = \{X_1, X_2, \dots, X_n\}$$

(برای حل این مسئله، فقط می‌توانید از اطلاعات مربوط به توزیع‌های پواسون و گاما که در این مسئله ارائه شده است، استفاده کنید.)

الف) نشان دهید که میانگین نمونه‌ای[8] ($\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$) برابر با تخمین بیشینه درست نمایی[9] برای λ است و و همچنین این تخمین ناریب[10] است ($\mathbb{E}(\hat{\lambda}) = \lambda$).

ب) حالا فرض کنید در یک چارچوب بیزی قرار داریم و یک توزیع پیشین[11] برای λ در نظر می‌گیریم. فرض کنید که λ از توزیع گاما با پارامترهای (α,β) پیروی می‌کند که تابع چگالی احتمال آن به صورت زیر داده شده است:

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

[8] sample mean
[9] maximum likelihood estimate (MLE)
[10] unbiased
[11] prior distribution

که در آن $\Gamma(\alpha) = (\alpha - 1)!$ (در اینجا فرض می‌کنیم α یک عدد صحیح مثبت است). توزیع پسین[12] را برای λ محاسبه کنید.

ج) یک عبارت تحلیلی برای تخمین گر بیشینه گر احتمال پسین[13] از λ تحت توزیع پیشین $Gamma(\alpha,\beta)$ به دست آورید.

د) توضیح دهید در چه شرایطی استفاده از هر کدام از این دو روش تخمین (MAP و MLE) بر دیگری برتری دارد.

[12] posterior distribution
[13] maximum a posterior (MAP)

سوال ۵: (۱۵ نمره)

یک مسئله دسته بندی c کلاسه را در نظر بگیرد، در صورتی که احتمال پیشین کلاس ها با هم برابر باشد، اثبات کنید که حد بالا خطای دسته بند نزدیک ترین همسایه به صورت زیر است. (خطای دسته بند بیز با p^* نمایش داده شده است)

$$P \leq P^* \left(2 - \frac{c}{c-1}\right) P^*$$

سوال ۶: (۱۵ نمره)

توزیع نرمال $p(x) = N(\mu, \sigma^2)$ با تابع کرنل $\varphi(x) = N(x, 1)$ را در نظر بگیرید توزیع تخمینی بر اساس N داده که مستقل از هم و هم توزیع و از توزیع $p(x)$ با استفاده از تخمین پارزن به صورت زیر است:

$$p_n(x) = \frac{1}{n\, h_n} \sum_{i=1}^{N} \varphi\left(\frac{x - x_i}{h_n}\right)$$

در مورد تابع خطا MISE[۱۴] تحقیق کنید و بر اساس این خطا، خطای بایاس و خطای واریانس توزیع تخمینی $p_n(x)$ را بدست بیاورید. (جواب تقریبی برای هر دو مورد خطای بایاس و خطای واریانس کفایت می‌کند.)

[14] Mean Integrated Squared Error

سوال ۷: (۱۰ نمره)

به سوالات زیر با بیان توضیح کامل پاسخ دهید.

۱- در دسته بند KNN[15]با افزایش K باياس مدل افزایش و واریانس آن کاهش می یابد یا بر عکس؟

۲- دو مورد از مزیت های استفاده از فاصله mahalanobis نسبت به فاصله اقلیدسی در دسته بند KNN را بیان کنید.

۳- در تخمین چگالی احتمال پارازن، تحت چه شرایطی بهتر است از تابع کرنل تطبیق پذیر به جای تابع کرنل ثابت استفاده کنیم؟

۴- در مورد ساختمان داده LSH[16] تحقیق کنید و نحوه کارکرد آن را توضیح دهید و همچنین بیان کنید چگونه این ساختمان داده می‌تواند باعث افزایش سرعت دسته بند KNN شود و هم چنین مزیت آن را نسب به ساختمال داده K-d tree بیان کنید.

[15] K-nearest neighbors
[16] Locality-Sensitive Hashing

سوال ۸: (**شبیه سازی**، ۲۰ نمره)

در این تمرین طبقه‌بند بیز، بدون استفاده از کتابخانه پیاده‌سازی شده و روی مجموعه‌داده "Iris" اعمال خواهد شد. این مجموعه‌داده شامل اندازه‌گیری‌های ۴ ویژگی مختلف برای ۳ گونه گل زنبق است.

الف) ابتدا توضیح مختصری راجع به طبقه‌بندهای optimal bayes و naïve bayes دهید و توضیح دهید که چرا بجای طبقه‌بند بیز از naïve bayes استفاده می‌کنیم، هزینه ای که می‌دهیم چیست و در چه زمان هایی استفاده از این طبقه‌بند کاری منطقی است.

ب) مجموعه‌داده داخلی Iris را بارگیری کنید. دو ویژگی طول و عرض گلبرگ[17] گونه‌های "virginica"و "versicolor"را استخراج کنید (ما فقط با این دو گونه کار خواهیم کرد). ویژگی‌های هر دو کلاس را در یک نمودار دو بعدی نمایش دهید.

ج) naïve bayes را با فرض گوسی بودن داده‌های ورودی پیاده‌سازی کنید (در این قسمت مجاز به استفاده از کتابخانه های آماده مثل sklearn.naive_bayes.GaussianNB نیستید و باید الگوریتم خودتان را پیاده سازی کنید).

د) الگوریتم‌های پیاده‌سازی شده را برای مجموعه‌داده Iris تست کنید و مرز تصمیم‌گیری حاصل را نمایش دهید. برای ترسیم مرز تصمیم می‌توانید از "meshgrid"(برای ایجاد یک شبکه دوبعدی) و "contour"یا "contourf" (برای تجسم مرز بین کلاس‌ها) استفاده کنید.

ه) ماتریس در هم ریختگی[18] را گزارش کنید و مقادیر صحت[19] و دقت[20] را محاسبه کرده و نتایج هر کدام را توضیح دهید.

و) مورد ج را به کمک کتابخانه scikit-learn انجام دهید و نتایج دو بخش را مقایسه کنید.

توجه:

برای بارگذاری مجموعه داده می‌توانید از sklearn.datasets استفاده کنید.

برای قسمت «ه» و «و» می‌توانید از کتابخانه sklearn استفاده کنید.

[17] petal length and petal width
[18] Confusion Matrix
[19] accuracy
[20] precision