الگوشناسی آماری ۷۲۵-۴۰
تمرین سری اول
موعد تحویل: ۱۳ اسفند ۹۱

نیمسال دوم ۹۲-۹۱
مدرس: سلیمانی

سوال ۱ (۱۰ نمره): روابط آماری زوج متغیرهای تصادفی

دو متغیر تصادفی X و Y را در نظر بگیرید و معیارهای زیر را برای رابطه این دو متغیر تصادفی ملاحظه نمایید:

$$cov(X,Y) = \sigma_{XY} = E_{X,Y}[(X-\mu_X)(Y-\mu_Y)]$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$I(X,Y) = E_{X,Y}\left[\ln \frac{P(X,Y)}{P(X)P(Y)}\right]$$

a) (۲ نمره) با استفاده از نامساوی کوشی-شوارتز ثابت کنید $-1 \le \rho_{XY} \le 1$.

b) (۲ نمره) رابطه $\rho_{XY} = 0 \Rightarrow I(X,Y) = 0$ را در صورت درستی اثبات نموده و در صورت نادرست بودن برای آن مثال نقض بیاورید.

c) (۲ نمره) رابطه $I(X,Y) = 0 \Rightarrow \rho_{XY} = 0$ را در صورت درستی اثبات نموده و در صورت نادرست بودن برای آن مثال نقض بیاورید.

d) (۲ نمره) آیا $\rho_{XY} = 0$ نشان‌دهنده استقلال X و Y است؟ $I(X,Y) = 0$ چطور؟ (توضیح)

e) (۲ نمره) چنان‌چه X و Y مستقل آماری باشند، در مورد مقدار ρ_{XY} و $I(X,Y)$ چه اظهارنظری می‌توان کرد؟

سوال ۲ (۱۰ نمره): ماتریس کوواریانس و فاصله ماهالانبیس

توزیع گاوسی چند متغیره $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ را در فضای ویژگی در نظر بگیرید:

a) (۲ نمره) فاصله ماهالانبیس بردار $\boldsymbol{x}$ را با این توزیع مشخص نموده و رابطه را برای حالتی که ماتریس کواریانس قطری باشد ساده کنید ($\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$).

b) (۱ نمره) در چه شرایطی فاصله اقلیدسی $\boldsymbol{x}$ با $\boldsymbol{\mu}$ برابر با فاصله ماهالانبیس $\boldsymbol{x}$ با توزیع $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ خواهد بود؟

c) (۲ نمره) در چه شرایطی استفاده از فاصله ماهالانبیس بهتر از فاصله اقلیدسی است؟ و بالعکس؟

d) (۳ نمره) راهکاری برای تعمیم فاصله ماهالانبیس برای محاسبه فاصله دو نقطه $\boldsymbol{x}_1$ و $\boldsymbol{x}_2$ (از این توزیع) از همدیگر ارائه کنید؟ در چه حالتی این فاصله برابر با فاصله اقلیدسی $\boldsymbol{x}_1$ و $\boldsymbol{x}_2$ خواهد بود؟

e) (۲ نمره) وجود مقدار ویژه صفر برای ماتریس کوواریانس یک گاوسی به چه معناست؟ این موضوع چه تاثیری بر روی فاصله ماهالانبیس یک نقطه تا گاوسی مربوطه می‌گذارد؟

سوال ۳ (۱۴ نمره): تجزیه ویژه ماتریس کوواریانس و تبدیلات خطی مربوطه

با استفاده از Matlab برای مجموعه داده موجود در فایل "data_3.mat" به پرسش‌های زیر پاسخ دهید:

a) (۲ نمره) ماتریس $\boldsymbol{C} = \frac{1}{N}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}$ را به عنوان تخمینی از کوواریانس ویژگی‌ها برای این مجموعه داده محاسبه نمایید ($\widetilde{\boldsymbol{X}}$ از روی $\boldsymbol{X}$ و با کم کردن میانگین نمونه‌ها از همه‌ی داده‌ها به دست آمده است $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{\mu}$).

b) (۲ نمره) با استفاده از مقادیر درایه‌های ماتریس کواریانس در مورد ارتباط هر کدام از زوج ویژگی‌ها و همچنین پراکندگی مقدار هر یک از ویژگی‌ها بحث کنید.

c) (۲ نمره) ماتریس $\boldsymbol{V} = [\boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_d]$ که ستون‌های آن بردارهای ویژه (متناظر با مقادیر ویژه متمایز) ماتریس $\boldsymbol{C}$ هستند را تشکیل دهید. ضرب داخلی دوبه‌دوی این بردارهای ویژه را مشخص کنید. نتیجه به دست آمده را با استفاده از روابط نیز نشان دهید.

d) (۳ نمره) داده‌ها را با استفاده از تبدیل $\boldsymbol{x}' = \boldsymbol{V}^T\boldsymbol{x}$ به فضای جدید ببرید. سپس یک بار داده‌ها را به ازای هر یک از زوج ابعاد (x_1, x_2)، (x_1, x_3) و (x_2, x_3) در فضای دوبعدی مربوطه و یک‌بار در فضاهای جدید (x'_1, x'_2)، (x'_1, x'_3) و (x'_2, x'_3) رسم نمایید و نتایج به‌دست آمده را توصیف کنید.

e) (۳ نمره) برای داده‌های تبدیل یافته ماتریس کواریانس را محاسبه نمایید. نتیجه به دست آمده را با استفاده از روابط نیز نشان دهید.

f) (۲ نمره) چنان‌چه از تبدیل $\boldsymbol{x}' = \left(\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2}\right)^T\boldsymbol{x}$ استفاده شود ($\boldsymbol{\Lambda}$ ماتریسی قطری است که مقادیر ویژه متناظر با بردارهای ویژه موجود در ماتریس $\boldsymbol{V}$ به ترتیب روی قطر آن واقع شده‌اند)، با استفاده از روابط نشان دهید که ماتریس کواریانس داده‌ها در فضای جدید همانی خواهد شد.

سوال ۴: (۲۲ نمره) خطا و ریسک بیز

یک مساله دسته‌بندی به دو دسته را در نظر بگیرید. برای یک دسته‌بند که فضا را به نواحی $\mathcal{R}_1$ و $\mathcal{R}_2$ افراز می‌کند، احتمال خطا برای دسته‌های اول و دوم به ترتیب به صورت $\epsilon_1 = \int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1)d\boldsymbol{x}$ و $\epsilon_2 = \int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2)d\boldsymbol{x}$ محاسبه می‌شود و احتمال خطای کلی دسته‌بند برابر است با $\epsilon_1 \times P(\omega_1) + \epsilon_2 \times P(\omega_2)$.

۱.۴. (۴ نمره) نشان دهید برای دسته‌بند بهینه ($c = 2$) احتمال خطا بیشتر از 0.5 نیست. تعمیم این گزاره برای حالتی‌که تعداد دسته‌ها محدود به دو دسته نباشد چیست؟ درستی گزاره حاصل را نشان دهید.

۲.۴. (۳ نمره) نشان دهید ریسک دسته‌بندی به نواحی $\mathcal{R}_1$ و $\mathcal{R}_2$ به صورت زیر است:

$$R = P(\omega_1)\lambda_{11} + P(\omega_2)\lambda_{22} + P(\omega_1)(\lambda_{12} - \lambda_{11})\epsilon_1 + P(\omega_2)(\lambda_{21} - \lambda_{22})\epsilon_2$$

۳.۴. (۵ نمره) اگر احتمال پیشین دسته‌ها برابر و چگالی شرطی دسته‌ها به صورت زیر باشد:

$$p(x|\omega_1) \sim N(\mu, \sigma^2)$$
$$p(x|\omega_2) \sim U(a, b)$$

نشان دهید که احتمال خطای دسته‌بند بیز کوچکتر مساوی $G((b-\mu)/\sigma) - G((a-\mu)/\sigma)$ است.

$G(x) \equiv Pr(y \leq x)$ و $y \sim N(0,1)$.

۴.۴. (۵ نمره) فرض کنید توزیع بردارهای ویژگی در هر دو دسته، گاوسی با ماتریس کوواریانس $\boldsymbol{\Sigma}$ و بردارهای میانگین دو دسته $\boldsymbol{\mu}_1$ و $\boldsymbol{\mu}_2$ باشند. همچنین احتمال پیشین دو دسته مساوی باشد. نشان دهید احتمال خطای دسته‌بند بیز از رابطه زیر به دست می‌آید:

$$P_B(error) = \int_{d_M/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

که d_M فاصله ماهالانبیس $\boldsymbol{\mu}_1$ و $\boldsymbol{\mu}_2$ را نشان می‌دهد.

(راهنمایی: برای محاسبه احتمال خطای بیز یک متغیر تصادفی $r = \ln\frac{p(x|\omega_1)}{p(x|\omega_2)}$ تعریف کنید و نشان دهید توزیع r برای $\boldsymbol{x} \in \mathcal{R}_1$، $N(d_M^2/2, d_M^2)$ و برای $\boldsymbol{x} \in \mathcal{R}_2$، $N(-d_M^2/2, d_M^2)$ است)

۵.۴. (۵ نمره) برای مساله دسته‌بندی به دو دسته با استفاده از یک ویژگی که توزیع مقادیر ویژگی در هر دو دسته گاوسی فرض می‌شود، تابعی در Matlab بنویسید که با گرفتن میانگین و انحراف استاندارد گاوسی‌های مربوطه و همچنین احتمال پیشین دو دسته، مقدار خطای بیز را برگرداند. نتیجه به‌دست آمده را به ازای ورودی $\mu_1 = 0.5$، $\mu_2 = 2$، $\sigma_1 = 1$، $\sigma_1 = 0.75$، $P(\omega_1) = P(\omega_2)$ گزارش کنید.

سوال ۵: (۲۰ نمره) مرزهای تصمیم‌گیری با فرض توزیع‌های گاوسی $p(\boldsymbol{x}|\omega_i)$

یک مساله دسته‌بندی با دو دسته را درنظر بگیرید. فرض کنید نمونه‌های هر دسته یک توزیع گاوسی در فضای دو بعدی ویژگی‌ها دارد.

۱.۵. چنان‌چه پارامترهای توزیع بردارهای ویژگی در دو دسته به صورت زیر باشد:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

و احتمال پیشین دو دسته یکسان فرض شود:

a. (۲ نمره) مرز به دست آمده توسط دسته‌بند بیز را مشخص کنید.

b. (۲ نمره) چنان‌چه از ماتریس ضرر زیر استفاده شود مرز تصمیم حاصل چگونه خواهد بود؟

$$\begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$$

c. (۲ نمره) اگر به جای فرض احتمال پیشین برابر دسته‌ها داشته باشیم $P(\omega_1) = 3P(\omega_2)$ مرز تصمیم در این حالت و با لحاظ کردن ماتریس ضرر معرفی شده در قسمت b چگونه خواهد بود؟

۲.۵. نمایش توزیع و مرز دسته‌ها در Matlab

a. (۸ نمره) یک تابع بنویسید که با دریافت بردار میانگین و ماتریس کواریانس هر دسته و با فرض احتمال پیشین برابر برای دسته‌ها، توزیع داده‌های دسته‌ها و مرز بین دو دسته را برای دسته‌بند بیز نمایش دهد.

b. (۶ نمره) به ازای مسائل موجود در فایل "Problems.mat" تصویر خروجی مشخص کننده توزیع داده‌ها و مرزها را در گزارش خود بیاورید.

سوال ۶ (۹ نمره): ریسک بیز و کنش عدم تشخیص

در یک مساله دسته‌بندی به c دسته فرض کنید برای تصمیم‌گیری $c+1$ کنش $\mathcal{A}=\{\alpha_1,\ldots,\alpha_c,\alpha_{c+1}\}$ در نظر گرفته شده که c کنش اول معادل با انتساب به دسته‌ی اول تا c–ام هستند و کنش α_{c+1} برای رد کردن (عدم تشخیص) لحاظ شده است. فرض کنید تابع ضرر برای این مساله به صورت زیر است:

$$\lambda(\alpha_i|\omega_j)=\begin{cases}0 & i=j\\ \lambda_r & i=c+1\\ \lambda_s & o.w.\end{cases}$$

a) (۳ نمره) نشان دهید که قانون تصمیم زیر ریسک را کمینه می‌کند:

If $\forall j\neq i, P(\omega_i|\boldsymbol{x})\geq P(\omega_j|\boldsymbol{x})$ and $P(\omega_i|\boldsymbol{x})\geq 1-\lambda_r/\lambda_s$ decide α_i

Otherwise decide α_{c+1}

b) (۲ نمره) در مورد تاثیر مقدار λ_r/λ_s بر نحوه‌ی تصمیم‌گیری بحث کنید.

c) (۴ نمره) نشان دهید که توابع جداساز زیر برای این مساله بهینه هستند:

$$g_i(x)=\begin{cases}p(x|\omega_i)P(\omega_i) & i=1,\ldots,c\\ \dfrac{\lambda_s-\lambda_r}{\lambda_s}\displaystyle\sum_{j=1}^{c}p(x|\omega_j)P(\omega_j) & i=c+1\end{cases}$$

سوال ۷ (۱۵ نمره): انواع معیارهای تصمیم

۱.۷. مساله‌ی دسته‌بندی به دو دسته با احتمال پیشین یکسان و $p(x|\omega_1)\sim U(0,1)$ و $p(x|\omega_2)\sim U(-0.5,1.5)$ را در نظر بگیرید:

a. (۱ نمره) نواحی تصمیم بیز را مشخص نمایید.

b. (۲ نمره) نمودار ROC را با در نظر گرفتن نقطه مرزی به عنوان پارامتر متغیر رسم نمایید.

c. (۳ نمره) نواحی تصمیم Neyman-Pearson را برای $\epsilon_2=0.2$ مشخص کنید.

d. (۴ نمره) نواحی تصمیم Minimax را تعیین نمایید.

۲.۷. (۵ نمره) نشان دهید در معیار Neyman-Pearson اگر ϵ_1 را برابر ثابت α در نظر بگیریم، حداقل کردن ϵ_2 معادل با تصمیم‌گیری به صورت زیر است:

اگر $\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})}>k_\alpha$ آنگاه ω_1 را انتخاب کن وگرنه ω_2.

که k_α به گونه‌ای انتخاب شده که $\epsilon_1=\alpha$ برقرار شود.

(راهنمایی: از یک ضریب لاگرانژ استفاده کنید و نشان دهید که حل مساله‌ی موردنظر معادل با حداقل کردن $k_\alpha(\epsilon_1-\epsilon_2)+\epsilon_2$ است.)