

بناام خدا


الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

تمرینات سری دوم (مفاهیم کلاسه‌بندی، روش‌های کلاسه‌بندی بر پایه احتمال)

بهار ۱۳۹۱

به نکات زیر توجه فرمائید:

۱. زمان تحویل تمرینات در سایت درس مشخص شده است. دقت نمائید که زمانبندی‌های تعیین شده قابل تغییر نیستند.
۲. تمرینات را با عنوان SPR-HWx-8xxxxxxx (مثلا SPR-HW1-88300785) و در یک فایل فشرده با همین نام به آدرس Muhammadi@dml.ir ایمیل زده و در اولین جلسه بعد از زمان تحویل، بصورت پرینت شده تحویل استاد درس دهید.
۳. گزارش شما باید مختصر و مفید باشد. برای تمرینات پیاده‌سازی که با لوگوی  مشخص شده‌اند باید کد متلب نوشته شده ضمیمه گزارش شده و تمامی خروجی‌های برنامه‌ها در گزارش شما ذکر شوند.

سوال ۱) نمودار ROC را برای حالتی که دو کلاس دارای دو توزیع گاوسی $P(x, w_1) \sim N(0,1)$ و $P(x, w_2) \sim N(1,2)$ باشند، رسم کنید.

سوال ۲) با یک مجموعه داده یک کلاسه‌بند را دو بار آموزش می‌دهیم. بار اول داده‌ها را به صورت تصادفی به نسبت ۵۰-۵۰ به دو دسته آموزش و تست تقسیم کرده و عملیات آموزش کلاسه‌بند را انجام می‌دهیم و بار دوم داده‌ها را به صورت تصادفی به نسبت ۲۰-۸۰ به دو دسته آموزش و تست تقسیم می‌کنیم (مجموعه آموزشی، بزرگتر از مجموعه تست) و عملیات آموزش کلاسه‌بند را انجام می‌دهیم. کارایی کلاسه‌بند اول بر روی مجموعه آموزشی‌اش ۸۰ درصد و کارایی کلاسه‌بند دوم بر روی مجموعه آموزشی‌اش ۹۰ درصد بوده است. آیا برای یک مجموعه داده جدید کلاسه‌بند دوم بهتر از کلاسه‌بند اول عمل می‌کند؟ نظر خود را توضیح دهید.

سوال ۳) دو کلاس با توزیع‌های گاوسی و با پارامترهای زیر در نظر بگیرید:

$$P_1 = P_2 = 0.5, \mu_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} +1 \\ 0 \end{bmatrix}, \Sigma_1 = \Sigma_2 = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$$

الف) میزان خطای کمینه را برای این مساله کلاسه‌بندی بیابید. خطای تشخیص کلاس اول و دوم را جداگانه نیز ارائه دهید.
ب) اگر بخواهیم خطای تشخیص کلاس اول، از نصف حالت قبل بیشتر نشود، مرز کلاسه‌بندی به چه صورت تغییر خواهد کرد (از روش Neyman-Pearson استفاده کنید). در این صورت خطای تشخیص کلاس دوم چقدر خواهد شد و میزان کل خطا نسبت به حالت قبل به چه صورت تغییر می‌کند؟

پ) مرز تصمیم‌گیری را برای این مساله در صورتی که بخواهیم ریسک را کمینه کنیم، در حالتی که O-1 loss داشته باشیم با حالتی که ماتریس loss بصورت زیر باشد، را بدست آورده و بر روی یک نمودار هر دوی آنها را نمایش دهید:

$$L = \begin{bmatrix} 0 & 0.5 \\ 1 & 0 \end{bmatrix}$$

سوال ۴) یک مسئله دسته‌بندی سه کلاسه در دو بعد با توزیع‌های زیر را در نظر بگیرید:

$$P(x|w_1) \sim N(0, I)$$

$$P(x|w_2) \sim N([1 \ 1]^t, I)$$

$$P(x|w_3) \sim 0.5 * N([0.5 \ 0.5]^t, I) + 0.5 * N([-0.5 \ 0.5]^t, I)$$

$$P(w_1) = P(w_2) = P(w_3)$$

الف) با محاسبه احتمال‌های پسین، نقطه‌ی $x = [0.3 \ 0.3]^t$ را برای حالت کمترین احتمال خطا کلاسه‌بندی نمایید.

ب) فرض کنید برای یک نقطه خاص، ویژگی اول را نداریم (یعنی $x = [0.3 \ *]^t$). این نقطه را کلاسه‌بندی نمایید.

پ) فرض کنید برای یک نقطه دیگر، ویژگی دوم را نداریم (یعنی $x = [0.3 \ *]^t$). حال این نقطه را هم دسته‌بندی نمایید.

سوال ۵) یک مساله کلاسه‌بندی دو کلاسه چند بعدی را در نظر بگیرید که هر کدام از کلاس‌ها از یک توزیع نرمال تبعیت می‌کنند. کوواریانس این دو توزیع نیز با هم برابر است.

الف) نشان دهید که لگاریتم نسبت تشابه (likelihood ratio) نسبت به بردار ویژگی‌ها خطی است.

ب) اگر کوواریانس کلاس اول a برابر کوواریانس کلاس دوم شود (a یک عدد اسکالر می‌باشد)، مرز این دو کلاس را پیدا کرده و به ساده‌ترین فرم ممکن بنویسید.

پ) شرایطی را بر حسب احتمال‌های پیشین این دو توزیع پیدا کنید که مرز تصمیم‌گیری بدست آمده بوسیله بی‌ز، بین دو میانگین واقع نشود.

سوال ۶) فرض کنید دو کلاس c_1 و c_2 دارای توزیع گاوسی با میانگین $\mu_1 = [-2, -2]^t$ و $\mu_2 = [2, 2]^t$ که دارای ماتریس‌های کوواریانس زیر هستند، را داریم:

$$\Sigma_1 = \begin{bmatrix} \alpha & 0 \\ 0 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

این دو توزیع را رسم کنید و سطح بهینه جداساز بی‌زی را برای حالت $\alpha = 3$ و $\alpha = 1$ مشخص کنید. در هر دو مورد فرض کنید احتمال‌های پیشین یکسان هستند و هزینه کلاسه‌بندی صحیح برابر صفر و هزینه کلاسه‌بندی اشتباه برای هر دو یکسان است.

سوال ۷) برای یک مساله دو کلاسه یک بعدی قانون تصمیم‌گیری زیر را در نظر بگیرید:

یک نمونه را به کلاس w_1 نسبت بده اگر $x > \theta$ در غیر این صورت نمونه را به w_2 نسبت بده.

الف) نشان دهید که احتمال خطا برای این قانون مطابق رابطه زیر می‌باشد:

$$p(\text{error}) = P(w_1) \int_{-\infty}^{\theta} p(x|w_1) dx + P(w_2) \int_{\theta}^{+\infty} p(x|w_2) dx$$

ب) نشان دهید که شرط لازم برای کمینه شدن احتمال خطا این است که $p(\theta|w_1)P(w_1) = p(\theta|w_2)P(w_2)$ باشد.

سوال ۸) الف) مرز کلاسه‌بندی خطای کمینه بیز را برای مساله دو کلاسه زیر بیابید:

$$P(x|w_1) = \frac{1}{\pi b} \left(\frac{1}{1 + \left(\frac{x - a_1}{b}\right)^2} \right)$$
$$P(w_1) = P(w_2)$$

ب) $P(w_1|x)$ را برای $a_1=3$ ، $a_2=5$ و $b=1$ رسم کنید. رفتار $P(w_1|x)$ وقتی که x به سمت بی‌نهایت میل می‌کند، چگونه است؟

سوال ۹) دو توزیع یکنواخت یک بعدی زیر را در نظر بگیرید:

$$P(x|w_1) \sim U(0,2)$$

$$P(x|w_2) \sim U(1,4)$$

$$P(w_1) = P(w_2)$$

الف) نمودار ROC را رسم نمائید.

ب) مرز Neyman-Pearson را برای $E_2=0.25$ بدست آورید.

پ) مرز minimax را پیدا کنید.