تمرین اول درس شناسایی الگو

نام : حسین سیم چی

۹۸۴۴۳۱۱۹

استاد : آقای دکتر احمدعلی آبین

۱۳۹۹/۰۹/۲۱

سوال ۱) فرض کنید c کلاس w_c , ... , w_2 , w_1 و یک بردار ویژگی x داریم. قانون بیز برای دسته بندی، به شرط احتمال پیشین کلاس ها و چگالی احتمال شرطی برای x را بیان کنید.

مسئله ی کلاس بندی را می توان به ۲ روش کلی دسته بندی کرد. دسته اول، روش های Discriminative هستند که در این روش ها از یک خط برای جدا سازی کلاس ها استفاده می کنیم مثل روش های perceptron، LDA، SVM و ....

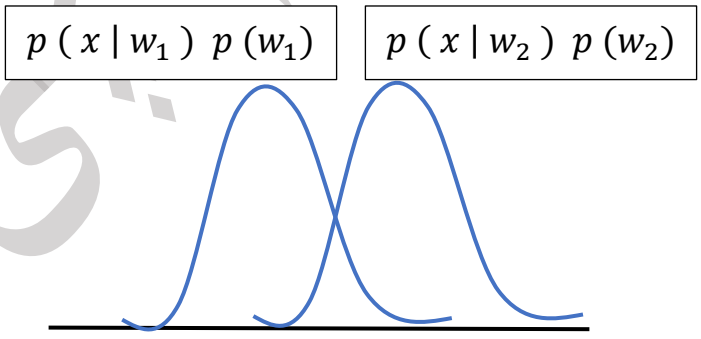
دسته دوم روش های Generative هستند که در این روش ها مثل Naïve bayes برای کلاس بندی از چگالی شرط استفاده می کند به این صورت که فضا را ناحیه بندی می کنیم

چیزی که در روش دسته بند بیز برای ما مهم این است که داده ای را بدهیم و نوع کلاس آن را مشخص نماییم که این را به صورت $p(w_c \mid x)$ نشان می دهیم. برای این که بتوان این احتمال را بدست آوریم ما فرض می کنیم که مسئله ی ما به صورت Parametric می باشد، یعنی توزیع را می دانیم (می دانیم که توزیع داده ها به چه صورت است و داده های قبلی که وارد سیستم شده اند در کدام کلاس قرار گرفته اند).در این صورت می توانیم با درنظر گرفتن توزیع گوسی داده ها و یا هر نوع توزیع مثل پوآسن احتمال بالا را بدست آورد :

$$p(w_c \mid x) = \frac{p(x \mid w_c)\ p(w_c)}{p(x)}$$

احتمال اینکه بردار ویژگی x در هر کلاس باشد

احتمال پیشین هر کلاس



برای دو کلاسه بودن می توان شکل فوق را رسم کرد.

سوال ۲) یک مسئله ی دو کلاسه ($A, \sim A$) با یک ویژگی تک مقداره را درنظر بگیرید. فرض کنید احتمال پیشین $p\,(A) = 0.33$ باشد. توزیع نمونه در جدول مقابل نشان داده شده است با استفاده از قانون بیز مقدار پسین کلاس ها را محاسبه کنید.

	A	$\sim A$
X	248	167
$\sim X$	82	503

با استفاده از فرمول زیر می توانیم مقدار پسین کلاس ها را بدست بیاوریم.

$$p\,(w_c \mid x\,) = \frac{p\,(x \mid w_c\,)\;p\,(w_c)}{p\,(x)}$$

$$p\,(A \mid x\,) = \frac{p\,(x \mid A)\,p(A)}{p(x)} = \frac{0.75 \times 0.33}{0.41} = 0.60$$

$$p\,(\sim A \mid x\,) = \frac{p\,(x \mid \sim A)\,p(\sim A)}{p(x)} = \frac{0.24 \times 0.67}{0.41} = 0.40$$

$$p\,(A \mid \sim x\,) = \frac{p\,(\sim x \mid A)\,p(A)}{p(\sim x)} = \frac{0.25 \times 0.33}{0.59} = 0.14$$

$$p\,(\sim A \mid \sim x\,) = \frac{p\,(\sim x \mid \sim A)\,p(\sim A)}{p(\sim x)} = \frac{0.76 \times 0.67}{0.59} = 0.86$$

سوال ۳) فرض کنید یک مسئله ی دسته بندی دو کلاسه داریم که احتمال پیشین هر دو کلاس باهم برابر است. ورودی دسته بند یک بردار ویژگی $X = (x_1\,, x_2)^T$ با دو المان که غیر منفی و از هم مستقل هستند. این مسئله با مشخصات زیر وجود دارد. احتمال درستی تصمیم کلاس بند با خطای کمینه چقدر است ؟

$$p\,(\,x_k \mid C_i\,) = \begin{cases} \lambda_{ik} e^{-\lambda_{ik} x_k}, & 0 \le x_k \\ 0, & \end{cases}$$

$$\lambda_{ik} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

همانطور که از صورت سوال و طرح مسئله مشخص است، هزینه ی خطا کردن در این دسته بندی برای تمام خطاها یکسان نیست و باید برای خطاهای خود هزینه ای را درنظر بگیریم که به این هزینه اصطلاحا ریسک کردن نیز می گویند. پس برای حداقل خطا باید ریسک هرکدام را درنظر بگیریم در نتیجه داریم:

$$\alpha\,(x) = argmin \sum \lambda_{ik}\,p\,(\,C_i \mid x\,)$$

در این سوال λ_{ik} به معنای هزینه ی انتصاب برچسب k درصورتیکه i برچسب یا کلاس درست بوده است.

برای ناحیه ی اول داریم :

$$R_1: \quad \lambda_{11} p(x|c_1) p(c_1) + \lambda_{12} p(x|c_2) p(c_2) = e^{-x} p(c_1) + 2e^{-2x} p(c_2)$$

و به همین ترتیب برای ناحیه ی دوم داریم :

$$R_2: \quad \lambda_{21} p(x|c_1) p(c_1) + \lambda_{22} p(x|c_2) p(c_2) = 2e^{-2x} p(c_1) + e^{-x} p(c_2)$$

و در نتیجه برای بدست آوردن احتمال درستی تصمیم کلاس بند با خطای کمینه یکی از نواحی باید محدوده ی بزرگتری را نسبت به دیگری داشته باشد و این بدین معنا است که مرز تصمیم را به سمت ناحیه ای جابه جا می کنیم که خطا در آن ناحیه ریسک بزرگتری دارد در نتیجه:

$$R_1 > R_2 \quad or \quad R_1 < R_2$$

کافی است یکی از معادلات فوق را حل نماییم تا احتمال درستی دسته بند با کمترین خطا را بدست آوریم.

سوال ۴) فرض کنید یک مسئله ی دسته بندی دودویی با مشخصات زیر وجود داشته باشند:

دسته بند با حداقل مقدار ریسک را بدست آورید

$$p(x \mid c_1) = x + \frac{1}{2} \quad , \quad p(x \mid c_2) = \frac{3x^2}{4} + \frac{3}{4}$$

$$\lambda_{ik} = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}$$

خب با استفاده از فرمول زیر ناحیه اول کلاس را بدست می آوریم:

$$\frac{p(x \mid c_1)}{p(x \mid c_2)} > \frac{(\lambda_{22} - \lambda_{12})\, p(c_2)}{(\lambda_{11} - \lambda_{21})\, p(c_1)}$$

با جایگذاری روابط فوق داریم :

$$\frac{x + \frac{1}{2}}{\frac{3x^2}{4} + \frac{3}{4}} > \frac{\frac{1}{4}(2-1)}{\frac{3}{4}(3-1)}$$

و در نهایت بازه ای که اگر سمپل وارد شده در این بازه باشد متعلق به کلاس اول است به شرح زیر است:

$$\frac{8 - \sqrt{76}}{2} < x < \frac{8 + \sqrt{76}}{2}$$

خب به همین ترتیب بازه ای را برای ناحیه ی دوم نیز بدست می آوریم، اگر بازه ی مربوط به هرکدام از ناحیه ها بزرگتر باشد بدین معنا است که ریسک خطا در این ناحیه از ناحیه ی دیگر کمتر است درنتیجه دسته بند با کمترین مقدار ریسک برابر با دسته بند با بیشتر بازه می باشد.

برای ناحیه ی دوم نیز به همین ترتیب عمل می کنیم با این تفاوت که بجای فرمول اول داریم :

$$\frac{p(x \mid c_2)}{p(x \mid c_1)} > \frac{(\lambda_{11} - \lambda_{21})\, p(c_1)}{(\lambda_{22} - \lambda_{12})\, p(c_2)}$$

سوال ۵) مسئله ی دسته بندی مقابل را درنظر بگیرید در ابتدا برای توزیع $Y \sim bernouli\ (\frac{1}{2})$ برچسب یک با احتمال ½ انتخاب شده است. اگر Y = 1 باشد آنگاه $X \sim bernouli\ (p)$ در غیر این صورت $X \sim bernouli\ (q)$. فرض کنید که $p > q$. دسته بند بهینه برای این توزیع چیست؟ و ریسک آن چیست ؟

فرض کنید کلاس ما y=1 است در این صورت اگر از x توزیع برنولی بگیریم در برنولی p قرار می گیرد.در غیر این صورت در برنولی q قرار می گیرد. در نتیجه چون p از q بزرگتر است در نتیجه می توان به عنوان مثال p را برابر کلاس یک و دیگری را برابر q درنظر گرفت. همچنین در صورت سوال احتمال پیشین برابر ½ در نظر گرفته شده است. در نتیجه برای این دسته بندی داریم :

$$P = \frac{1}{2}\ (1 - p) + \frac{1}{2}\ q$$

سوال ۶) در یک دسته بندی دو کلاسه و دوبعدی، بردارهای ویژگی به وسیله ی دو توزیع نرمال مشترک به صورت زیر تولید شده اند

$$\boldsymbol{\delta = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}}$$

هم چنین بردارهای میانگین به ترتیب $\mu_2 = [3 \quad 3]$, $\mu_1 = [0 \quad 0]$ هستند. بر اساس دسته بندی بیز، بردار $[1 \quad 2.2]$ در کدام کلاس قرار می گیرد؟

در ابتدا باید معکوس ماتریس کواریانس فوق را بدست آوریم و سپس چون ماتریس کواریانس برای ما مشخص شده است و برای هر دو کلاس یکسان نیست از فرمول زیر میزان فاصله را بدست می آوریم که این فاصله برابر فاصله ی ماهالانوبیس میباشد. فاصله ی نمونه ی داده شده تا هر دو کلاس را محاسبه می کنیم و هرکدام از این فواصل کمتر باشد، نمونه به آن کلاس تعلق دارد:

$$\delta^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$$

سپس فاصله ی ماهالانوبیس را برای هر کلاس با استفاده از فرمول زیر بدست می آوریم:

$$(x - \mu_i)^T\ \delta^{-1}\ (x - \mu_i)$$

با استفاده از فرمول فوق فواصل را به ترتیب برای هر دو کلاس محاسبه می کنیم که برای کلاس اول فاصله برابر 2.952 و برای کلاس دوم فاصله 3.672 بدست می آید. در نتیجه چون فاصله ی ماهالانوبیس در کلاس اول کمتر از کلاس دوم است، نمونه به کلاس اول تعلق دارد.

گزارش تمرین برنامه نویسی

در ابتدا نیاز داریم با استفاده از کتابخانه ی PANDAS مجموعه دادگان را بخوانیم.

```
iris = pd.read_csv("C:\\Users\\Lenovo\\Desktop\\iris.csv")
```

	sepal length	sepal width	iris
0	5.1	3.5	Iris-setosa
1	4.9	3.0	Iris-setosa
2	4.7	3.2	Iris-setosa
3	4.6	3.1	Iris-setosa
4	5.0	3.6	Iris-setosa
..	...	...	...
145	6.7	3.0	Iris-virginica
146	6.3	2.5	Iris-virginica
147	6.5	3.0	Iris-virginica
148	6.2	3.4	Iris-virginica
149	5.9	3.0	Iris-virginica

زمانی که مجموعه دادگان را به استفاده از دستور فوق می خوانیم متوجه می شویم که مقادیر ویژگی های مجموعه دادگان فوق اعشاری و با یک رقم اعشار است. به عنوان مثال عدد ۳.۶ را درنظر بگیرید. برای اینکه بتوانیم در ادامه مقادیر ویژگی ها را بین ۲۰ bin تقسیم کنیم در ابتدا با استفاده از دو دستور زیر مقادیر ویژگی ها را در عدد ده ضرب می کنیم.

```
iris['sepal length'] = iris['sepal length'].apply(lambda x: x*10)
iris['sepal width'] = iris['sepal width'].apply(lambda x: x*10)
```

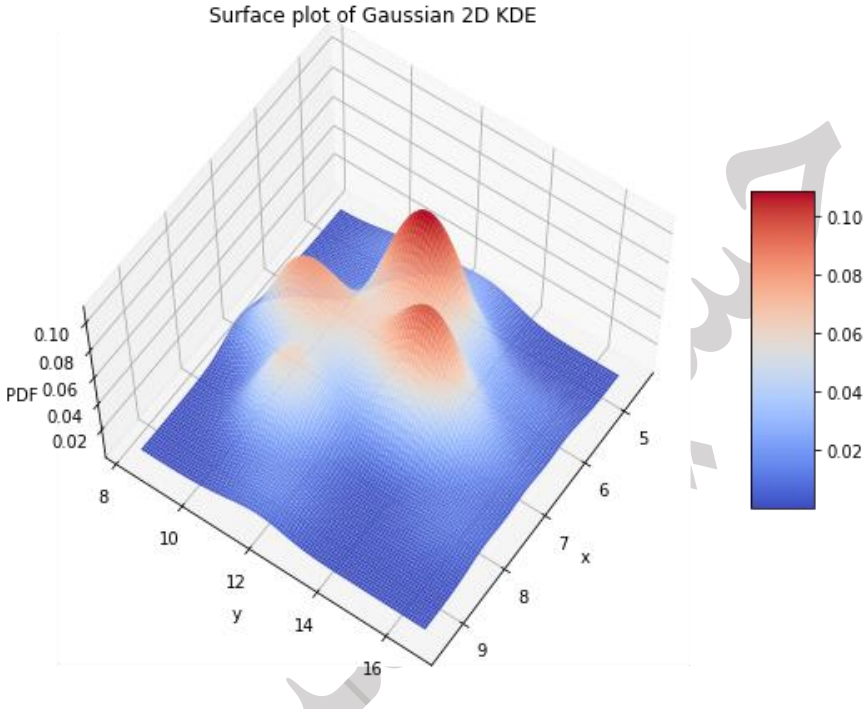
	sepal length	sepal width	iris
0	51.0	35.0	Iris-setosa
1	49.0	30.0	Iris-setosa
2	47.0	32.0	Iris-setosa
3	46.0	31.0	Iris-setosa
4	50.0	36.0	Iris-setosa
..	...	...	...
145	67.0	30.0	Iris-virginica
146	63.0	25.0	Iris-virginica
147	65.0	30.0	Iris-virginica
148	62.0	34.0	Iris-virginica
149	59.0	30.0	Iris-virginica

سپس اگر مقدار هر کدام از ویژگی ها را بین ۲۰ bin پخش می کنیم. به این صورت که به عنوان مثال اگر مقدار یک ویژگی بین عدد ۱ تا ۵ باشد در bin اول قرار می گیرد و اگر بین ۵ تا ۱۰ باشد در bin دوم قرار می گیرد و ... .

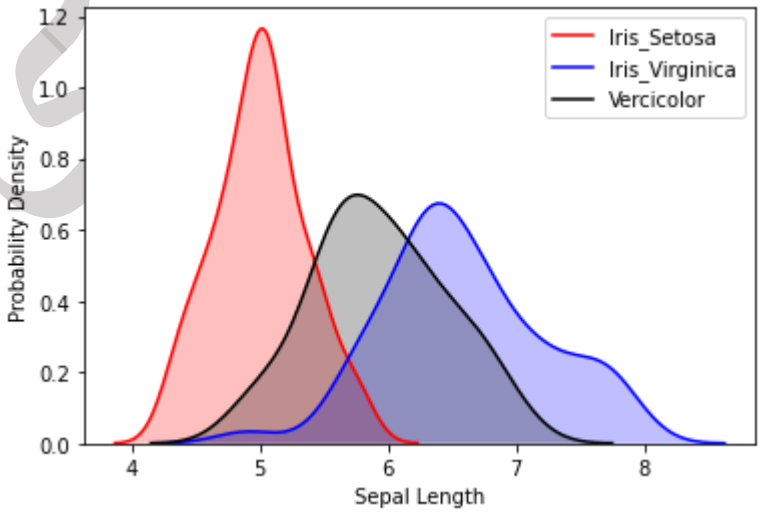
در شکل زیر می توان خروجی بین بندی را بر روی مجموعه ی دادگان مشاهده کنید عدد نوشته شده برای هر ویژگی به معنای شماره ی bin برای آن ویژگی است.

	sepal length	sepal width	iris
0	11.0	8.0	Iris-setosa
1	10.0	7.0	Iris-setosa
2	10.0	7.0	Iris-setosa
3	10.0	7.0	Iris-setosa
4	11.0	8.0	Iris-setosa
...	...	...	...
145	14.0	7.0	Iris-virginica
146	13.0	6.0	Iris-virginica
147	14.0	7.0	Iris-virginica
148	13.0	7.0	Iris-virginica
149	12.0	7.0	Iris-virginica

در ادامه با استفاده از تابع plot_2d_kde تابع توزیع کلاس ها را رسم می کنیم که در شکل زیر قابل مشاهده است. هرکدام از قله ها معرف یکی از کلاس ها است. کلاسی که داده ها در آن بیشترین توزیع را دارند بلندترین قله را دارد.



اگر به ازای تنها یک ویژگی نمودار تابع احتمال را برای هر سه کلاس رسم کنیم شکل زیر بدست می آید.



با مقایسه ی دو شکل فوق می توان دریافت قله ای که در شکل اول بالاترین ارتفاع را دارد معادل کلاس iris_Setosa است .

در ادامه با استفاده از کتابخانه ی Sklearn ستون مربوط به Iris که معادل لیبل است را به اعداد صحیح تبدیل می کنیم. مثلا کلاس اول معادل عدد ۱ ، کلاس دوم معادل عدد ۲ و کلاس سوم معادل عدد ۳ درنظر گرفته می شود تا در ادامه بتوانیم آموزش و تست را بر روی داده ها انجام دهیم.

در انتها باز هم با استفاده از کتابخانه ی Sklearn و استفاده از دسته بندی بیز و استفاده از Cross validation مجموعه ی ویزگی ها را به ۲۰ بازه تقسیم می کنیم که معادل تعداد bin های درنظر گرفته شده است. در آخر با استفاده از کتابخانه ی Matplotlib نمودار خواسته شده را برای هر bin بدست می آوریم.