گزارش تکلیف ۳ درس یادگیری ماشین

کسرا سینایی
شماره دانشجویی ۸۱۰۶۹۶۲۵۴

۲۲ آذر ۱۴۰۰

۱ تمارین آماری

۱.۱ سوال یک

$$P(\theta|D^0) = U(0,10) \xrightarrow{D_1=\{x:4\}} P(\theta|D^1) \propto P(x|\theta)P(\theta|D^0) = \begin{cases} 1/\theta & \theta \in [4,10] \\ 0 & ow \end{cases}$$

$$\xrightarrow{D_2=\{x_1=4,\, x_2=7\}} P(\theta|D^2) \propto P(x|\theta)P(\theta|D^2) = \begin{cases} \frac{1}{\theta^2} & \theta \in [7,10] \\ 0 & ow \end{cases} \longrightarrow \dots$$

$$\longrightarrow P(\theta|D^n) \propto \begin{cases} \frac{1}{\theta^n} & \max_x [D^n] \leqslant \theta \leqslant 10 \\ 0 & ow \end{cases} \longrightarrow P(\theta|D^4) \propto \begin{cases} \frac{1}{\theta^4} & 8 \leqslant \theta \leqslant 10 \\ 0 & ow \end{cases}$$

$$\xrightarrow{\text{Max Likelihood}} \hat{\theta} = 8 \qquad P(x|D) \sim U(0,8)$$

۲.۱ سوال دو

$$P(D|\theta) = \prod_{k=1}^{n} P(x|\theta) = \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ik}} (1-\theta_i)^{(1-x_{ik})} \xrightarrow{\log}$$

$$\ell(\theta) = \sum_{k=1}^{n}\sum_{i=1}^{d} x_{ik} \ln \theta_i + (1-x_{ik}) \ln (1-\theta_i) \qquad \boxed{\underline{\theta} = \frac{1}{n}\sum_{k=1}^{n} \underline{x}_k}$$

$$\frac{d\ell(\theta)}{d\theta} = 0 \longrightarrow \sum_{k=1}^{n}\left[\frac{x_{ik}}{\theta_i} + \frac{x_{ik}-1}{1-\theta_i}\right] = \sum_{k=1}^{n}\left[x_{ik} - \theta_i\right] = 0 \longrightarrow \theta_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik}$$

vector notation

۳.۱ سوال چهار

الف

$$\mu_n = \left[\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right]\frac{n}{\sigma^2}\hat{\mu}_n + \left[\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right]\frac{\mu_0}{\sigma_0^2} \qquad \left(\text{eliminating } \sigma_n^2 \text{ using: } \frac{1}{\sigma_n^2} = \frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)$$

$$\xrightarrow{n_0 = \sigma^2/\sigma_0^2} \mu_n = \left[\frac{1}{1+\frac{\sigma^2}{n\sigma_0^2}}\right]\hat{\mu}_n + \left[\frac{1}{1+\frac{n\sigma_0^2}{\sigma^2}}\right]\mu_0 = \left[\frac{1}{1+n_0/n}\right]\frac{1}{n}\sum_{k=1}^{n} x_k + \left[\frac{1}{1+n/n_0}\right]\frac{1}{n_0}\sum_{k=-n_0+1}^{0} x_k$$

$$= \frac{1}{n+n_0}\sum_{k=1}^{n} x_k + \frac{1}{n+n_0}\sum_{k=-n_0+1}^{0} x_k = \frac{1}{n+n_0}\sum_{k=-n_0+1}^{n} x_k \quad \checkmark\checkmark$$

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2}+\frac{1}{\sigma_0^2} \longrightarrow \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} = \frac{\sigma^2}{n+\frac{\sigma^2}{\sigma_0^2}} \xrightarrow{n_0 = \sigma^2/\sigma_0^2} \sigma_n^2 = \frac{\sigma^2}{n+n_0} \quad \checkmark\checkmark$$

ب

$$\mu_n = \frac{n}{n+n_0}\left[\frac{1}{n}\sum_{k=0}^{n} n\right] + \frac{n_0}{n+n_0}\left[\frac{1}{n_0}\sum_{k=-n_0+1}^{0} x_k\right]$$

* μ_n میانگین وزن‌دار داده‌های واقعی و داده‌های ساختگی باشد.

* اگر $\frac{1}{\sigma_n^2}$ را معیاری از دقت در نظر بگیریم؛ داریم $\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2}+\frac{n_0}{\sigma^2}$، که نشان می‌دهد دقت تخمین به صورت خطی با افزایش داده‌های واقعی بیشتر می‌شود.

۴.۱ سوال هفت

الف

$$\begin{cases} P(a_1) = P(a_4) = 0.5 \longrightarrow P(a_2) = P(a_3) = 0 \\ P(b_1) = 1 \rightarrow P(b_2) = 0 \\ P(d_2) = 1 \longrightarrow P(d_1) = 0 \end{cases}$$

$$P(x_1) = P(x_1 | a_1, b_1) P(a_1) P(b_1) + P(x_1 | a_4, b_1) P(a_4) P(b_1)$$

$$= \frac{(0.9)(0.65)}{0.9 \times 0.65 + 0.1 \times 0.35} \times 0.5 + \frac{(0.8)(0.65)}{(0.8)(0.65) + (0.2)(0.35)} \times 0.5$$

$$\longrightarrow \boxed{P(x_1) = 0.913 \quad (\text{fish is salmon})} \; *$$

$$P(x_2) = P(x_2 | a_1, b_1) P(a_1) P(b_1) + P(x_2 | a_4, b_1) P(a_4) P(b_1) = \frac{0.35 \times 0.1}{(0.35)(0.1) + (0.9)(0.65)} \times 0.5 + \frac{0.2 \times 0.35}{0.8 \times 0.65 + 0.2 \times 0.35} \times 0.5$$

$$\longrightarrow \boxed{P(x_2) = 0.087 \quad (\text{sea bass}),} \; *$$

evidence is only based on wide/thin:

$$P(x_1) = P(e | d_1) P(d_1) + P(e | d_2) P(d_2) = 0 + 1 \times 0.6 = 0.6 \quad **$$

Normalize $*, ** \longrightarrow P(x_1 | e) = \frac{0.913 \times 0.6}{0.913 \times 0.6 + 0.087 \times 0.4} = 0.992$

$$P(x_2 | e) = \frac{0.087 \times 0.4}{0.913 \times 0.6 + 0.087 \times 0.4} = 0.008$$

$\longrightarrow$ fish is salmon

$P_{error} = 0.008$

ب)

$$\begin{cases} P(e_D|d_1)=0 \rightarrow P(e_D|d_2)=1 \\ P(e_C|c_1)=0 \longrightarrow P(e_C|c_2)=1 \end{cases}$$

$$P(x_1) = P(e_C|x_1)\,P(e_D|x_1) = \left[P(e_C|c_1)P(c_1|x_1) + P(e_C|c_2)P(c_2|x_1) + P(e_C|c_3)P(c_3|x_1)\right]$$
$$\times\left[P(e_D|d_1)\,P(d_1|x_1) + P(e_D|d_2)\,P(d_2|x_1)\right]$$

$$\longrightarrow P(x_1) = (0+1\times0.33+0)(0+1\times0.6) = 0.198$$

$$P(x_2) = (0+1\times0.1+0)(0+1\times0.05) = 0.005$$

$$P(x_1|e) = \frac{0.198}{0.198+0.005} = 0.975\ ,\quad P(x_2|e) = \frac{0.005}{0.198+0.005} = 0.025 \Longrightarrow \boxed{\text{salmon } (x_1)}$$

$$P(a_1)=P(a_2)=P(a_3)=P(a_4)=\frac{1}{4} \Rightarrow P(a_i|\overset{\text{salmon}}{x_1}) = \frac{P(x|a_i)\,P(a_i)}{\sum_{i=1}^{4} P(x|a_i)\,P(a_i)} \longrightarrow \begin{cases} P(a_1)=0.375 \\ P(a_2)=0.125 \\ P(a_3)=0.167 \\ P(a_4)=0.333 \end{cases}$$

(winter) a_1

پ) مانند رابطه قسمت ب $\longrightarrow$

$$\left.\begin{matrix} P(x_1|e)=0.975 \\ P(x_2|e)=0.025 \end{matrix}\right\} \longrightarrow \text{salmon } (x_1)$$

North Atlantic $\longrightarrow P(b_1|e_B)=1,\ P(b_2|e_B)=0$

$$\Rightarrow \begin{cases} P(a_1|x_1,b_1) = P(x_1|a_1)\,P(x_1|b_1)\,P(a_1)P(b_1) = 0.4\times0.65\times0.25 = 0.146 \\ P(a_2|x_1,b_1) = 0.3\times0.65\times0.25 = 0.049 \\ P(a_3|x_1,b_1) = 0.4\times0.65\times0.25 = 0.065 \\ P(a_4|x_1,b_1) = 0.8\times0.65\times0.25 = 0.13 \end{cases}$$

$\Rightarrow$ winter , $P(x_1|e)P(a_1|x_1) = 0.975\times0.146 = \boxed{0.367}$

۲ تمارین شبیه سازی

۱.۲ سوال هشت

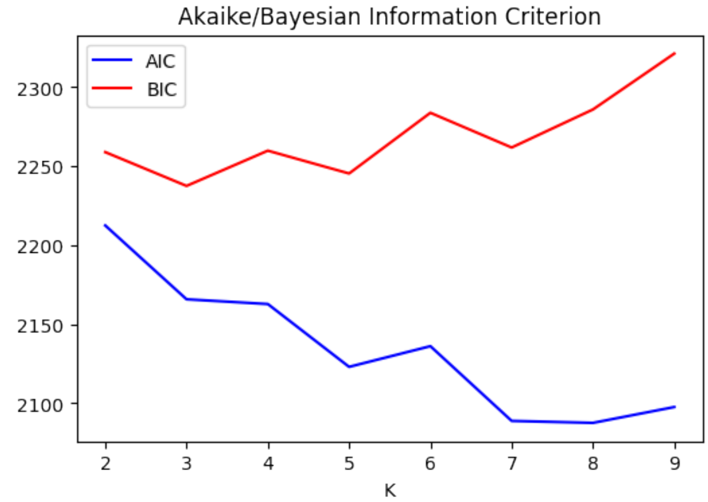
در این تمرین ابتدا با استفاده از کتابخانه OpenCV عکس‌های موجود در پوشه مربوط به این سوال را لود کرده و تنها مقادیر مربوط به میانگین کانال R و B را به عنوان فیچر استخراج شده و به همراه لیبل‌ها باز گردانده می‌شوند. تابع مورد استفاده در این سوال مشابه تمرین یک بوده و در فایل load pics py پیاده سازی شده است.
برای پیاده سازی طبقه بند در هر سه سوال این تکلیف یک کلاس به نام GMM پیاده سازی شده است که سه متد برای فیت کردن مدل و پیش‌بینی دارد:

- constructor: ابعاد فیچرها، تعداد کلاس‌ها و تعداد المان‌های میکسچر مدل را از کاربر می‌گیرد و یک لیست آبجکت mixture GaussianMixture (از کتابخانه scikit learn) با تعداد میکسچر تنظیم شده در آرگومان ورودی می‌سازد.

- update: داده‌ها و لیبل‌ها را گرفته و برای هر کلاس پارامترهای میکسچر مدل را آپدیت می‌کند. (با استفاده از متد fit)

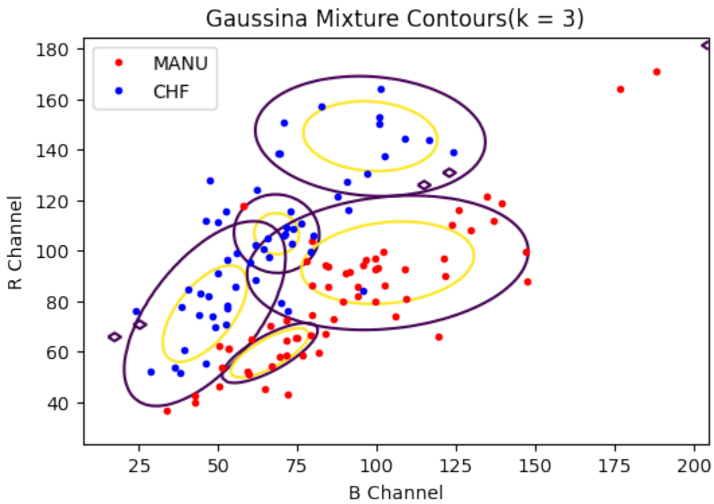
- predict: این تابع یک بردار با اندازه دلخواه از فیچرها می‌گیرد و با توجه به میکسچر مدل‌های یادگرفته شده لیبل آن‌ها را پیش‌بینی می‌کند. مبنای تصمیم گیری در این طبقه بند loglikelihood میکسچر مدل‌ها برای هر یک از کلاس‌ها است. با استفاده از متد score samples این مقدار به دست می‌آید.

تابع update حاصل جمع مقادیر AIC و BIC را بازمی‌گرداند. میانگین‌ها و ماتریس کواریانس همه کلاس‌ها نیز پابلیک می‌باشد و می‌توان خراج از کلاس به آن‌ها دسترسی داشت و از آن‌ها برای ترسیم کانتورهای میکسچرمدل‌های کلاس‌ها استفاده کرد.
تابع plot برای ترسیم این کانتورها در هر سه سوال شبیه سازی این تکلیف استفاده شده‌است. این تابع چهار آرایه numpy در آرگومان‌های ورودی خود دریافت می‌کند. آرایه اول در هر سطر خود میانگین یک توزیع گوسی را دارد، آرایه دوم ماتریس کواریانس متناظر با میانگین توزیع‌های آرایه اول است، آرایه سوم و چهارم داده‌های سوال هستند که به صورت نقاط رنگی اسکتر می‌شوند. برای رسم کانتور نیز یک مش گرید در راستای x و y تشکیل می‌دهیم و با استفاده از تابع Multivariate Normal از کتابخانه scipy احتمال آن را حساب می‌کنیم و برای هر توزیع نرمال دو خط کانتور رسم می‌کنیم.
نمودارهای AIC و BIC را پس از فیت کردن میکسچر مدل‌هایی با تعداد مدل ۲-۹ رسم می‌کنیم تا بر اساس آن بهینه ترین مدل برای این سوال را بیابیم.



شکل ۱: نمودارهای Akaike/Bayes Information Criterion به ازای مقادیر مختلف k

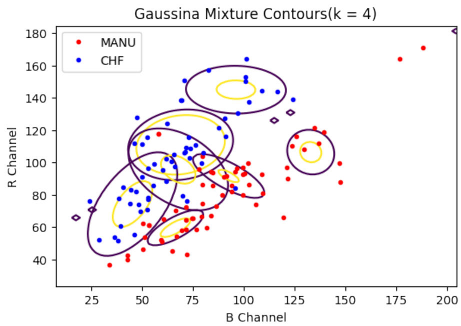
با توجه به نمودار می‌توان نتیجه گرفت مقدار ۳ و ۷ برای مدل مناسب به نظر می‌رسند. حال کانتورهای مربوط به میکسچر مدل را برای این طبقه بند رسم می‌کنیم.



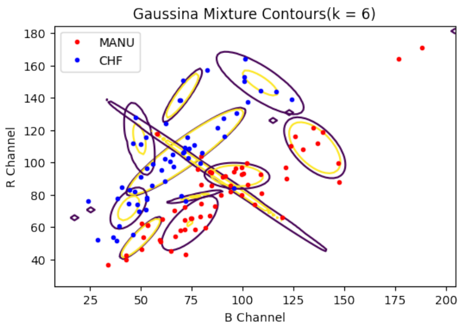
شکل ۲: کانتور توزیع‌های گوسی مربوط به میکسچر مدل‌های دو کلاس

برای این حالت پارامترهای مدل‌های گوسی فیت شده برای هر کلاس در جدول زیر گزارش شده‌اند. همچنین طبقه‌بند با تعداد مؤلفه‌های بالاتر (۴ و ۶) نیز آپدیت شد و کانتور مدهای گاوسی آن پس از یادگیری در شکل زیر نشان داده شده‌اند. با توجه به نمدار BIC تعداد مؤلفه بهینه برای این مسئله ۳ انتخاب شدُ تاثیر افزایش مؤلفه‌ها بر روی شکل

کانتور توزیع‌های گاوسی طبقه بند برای هر یک از کلاس‌ها در شکل دیده می‌شود. تطابق بیش از اندازه بر روی داده‌های نویزی (عکس‌هایی که هم رنگ آبی و هم رنگ قرمز در آن‌ها به یک مقدار دیده می‌شد) لزوماً به بهبود عملکرد طبقه بند نمی‌انجامد.



(أ) k = 4



(ب) k = 6

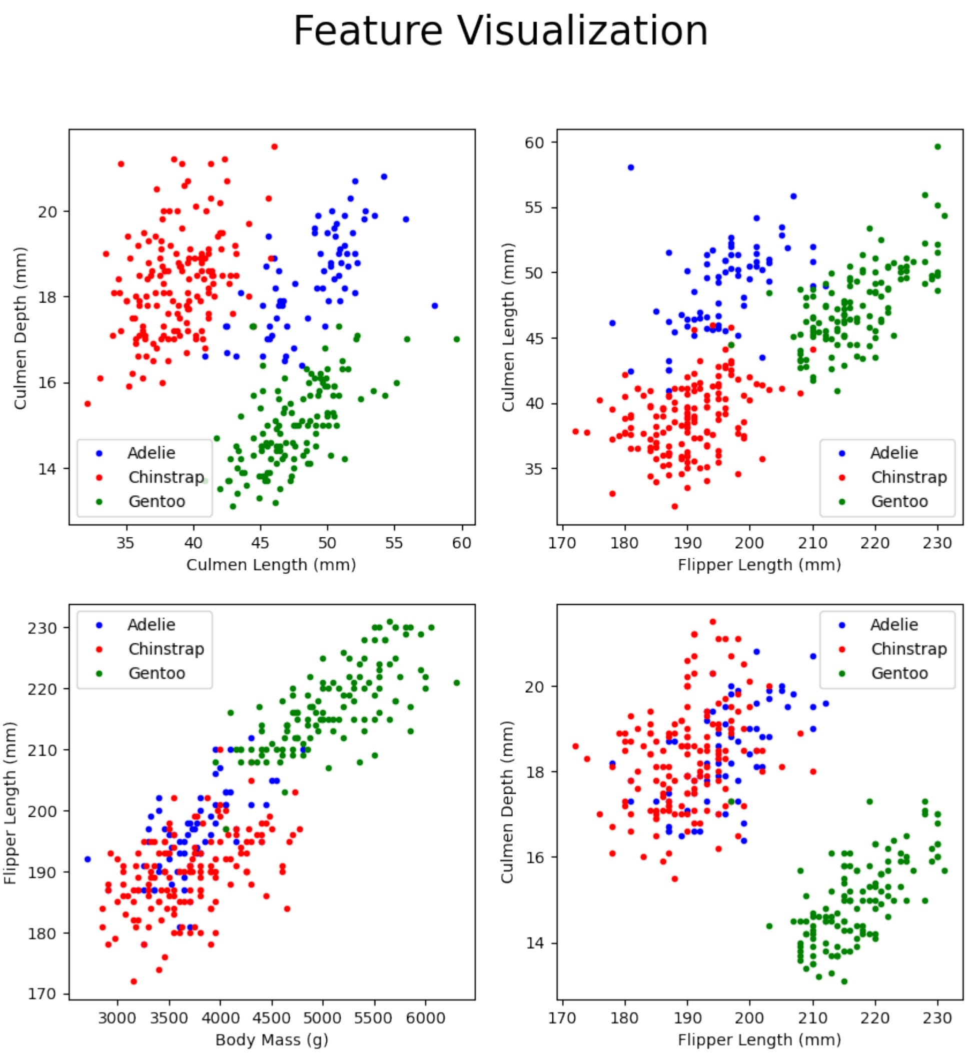
شکل ۳: کانتورهای به دست آمده از طبقه‌بند با تعداد مؤلفه‌های بیشتر

CHF			ManU			comp
σ	μ	weight	σ	μ	weight	
$\begin{pmatrix} 33.6 & 20.7 \\ 20.7 & 12.8 \end{pmatrix}$	$\begin{pmatrix} 182.4 & 167.4 \end{pmatrix}$	0.03	$\begin{pmatrix} 169.1 & -4.9 \\ -4.9 & 226.12 \end{pmatrix}$	$\begin{pmatrix} 68.4 & 106.6 \end{pmatrix}$	0.46	1
$\begin{pmatrix} 232.9 & 138.5 \\ 138.5 & 144.8 \end{pmatrix}$	$\begin{pmatrix} 66.3 & 59.4 \end{pmatrix}$	0.44	$\begin{pmatrix} 99.9 & 87.3 \\ 87.3 & 207.4 \end{pmatrix}$	$\begin{pmatrix} 45.8 & 74.9 \end{pmatrix}$	0.34	2
$\begin{pmatrix} 536.8 & 54.6 \\ 54.6 & 193.0 \end{pmatrix}$	$\begin{pmatrix} 103.5 & 95.1 \end{pmatrix}$	0.53	$\begin{pmatrix} 250.2 & -15.8 \\ -15.8 & 107.1 \end{pmatrix}$	$\begin{pmatrix} 98.0 & 145.2 \end{pmatrix}$	0.19	3

جدول ۱: پارامترهای مدل فیت شده

۲.۲ سوال نه

در این سوال ابتدا داده‌های سه کلاس را بر حسب فیچرهای خواسته شده دو به دو در نمودار به تصویر می‌کشیم تا تشخیص بدهیم کدام فیچرها برای طبقه‌بندی مناسب‌تر می‌باشند.



شکل ۴: نمونه‌های مورد استفاده برای طبقه بندی پنگوئن‌ها

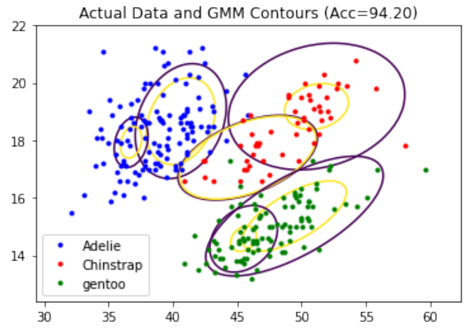
تفکیک‌پذیری داده‌ها بر حسب دو فیچر culmen length و culmen depth از سایر مجموعه فیچرهای خواسته شده در صورت سوال بهتر به نظر می‌رسند. داده‌های مربوط به سه گونه مختلف پنگوئن بر حسب این دو ویژگی از یکدیگر بیشتر فاصله داشته و نمونه‌های استثنایی کمتر به چشم می‌خورند.
در قسمت بعدی تمرین با استفاده از کلاس GMM که در سوال ۸ همین تمرین نیز از آن استفاده شد، طبقه بندهایی بر مبنای میکسچر مدل‌های گوسی با تعداد مؤلفه‌های مختلف فیت می‌کنیم. بر حسب دقت طبقه بند بر روی مجموعه

داده‌ی تست مشخص می‌کنیم که کدام مجموعه فیچر برای طبقه بندی بهتر می‌باشد. دقت حالات مختلف طبقه‌بندی در جدول زیر آورده شده‌اند.

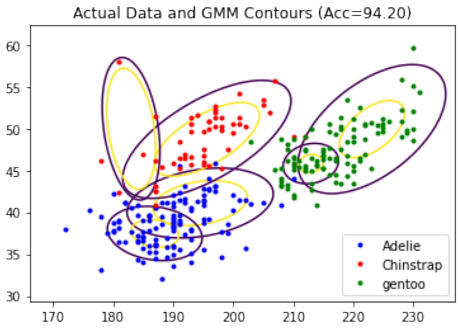
GMM Components	Feature Set	Accuracy
۲	۱	98.5
	۲	95.6
	۳	79.7
	۴	78.2
۳	۱	97.1
	۲	95.6
	۳	81.1
	۴	84.0
۴	۱	97.1
	۲	95.6
	۳	81.1
	۴	81.1
۵	۱	98.5
	۲	95.6
	۳	81.1
	۴	81.1

جدول ۲: دقت مدل‌های فیت بر روی مجموعه فیچرهای مختلف

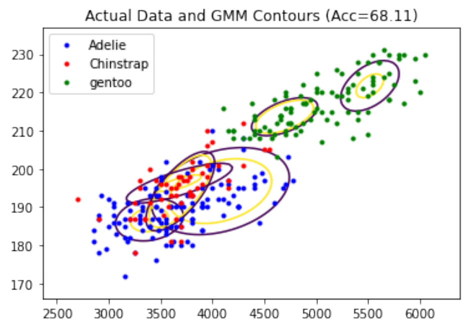
در داده‌های این سوال تعدادی از فیچرها در برخی از سمپل‌ها ناقص و نامعلوم بودند. با توجه به اینکه تعداد این موارد در کل دیتاست درصد قابل اغماضی بود با استفاده از imputer میانه، فیچرهای نامعلوم جاگذاری شدند. انتظار می‌رود تغییر چندانی در عملکرد نهایی طبقه بند مشاهده نشود، چون تعداد این موارد جمعاً به ده نمی‌رسید.
برای رسم کانتور توزیعذهای فیت شده با استفاده از زیرمجموعه‌های مختلف از دیتاست تعداد مؤلفه‌ها را برابر با ۲ در کلاس GMM تنظیم می‌کنیم و با دیتاست‌های مختلف مدل را آپدیت می‌کنیم. کانتورهای به دست آمده برای برای هر زیرمجموعه از دیتا ست و دقت طبقه‌بند بر روی مجموعه تست در تصاویر زیر نشان داده شده‌اند.



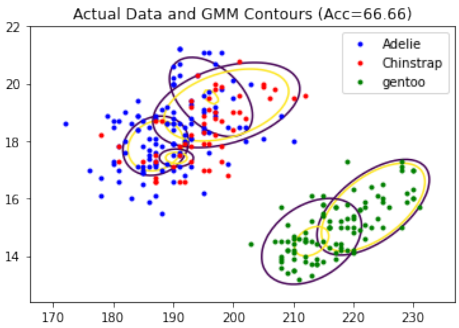
(آ) فیچرهای ۱



(ب) فیچرهای ۲



(ج) فیچرهای ۳

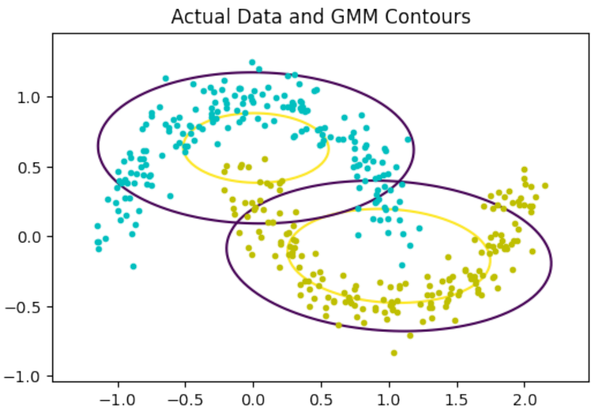


(د) فیچرهای ۴

شکل ۵: کانتور توزیع‌های گاوسی مربوط به میکسچرمدل با تعداد مؤلفه‌های متفاوت

۳.۲ سوال ده

در این سوال عملکرد طبقه‌بند ساده بیزی را با عملکرد طبقه‌بندی با مدل GMM مقایسه می‌کنیم.
ابتدا همانند تکالیف قبلی با استفاده از الگوریتم optimal bayes و بدون دانش اولیه (prior) داده‌ها را طبقه‌بندی می‌کنیم.
کانتور مدل‌های گاوسی به دست آمده برای تصمیم‌گیری بیز در شکل زیر نشان داده شده‌اند. دقت طبقه بند در این حالت برابر ۸۸ درصد می‌باشد.



شکل ۶: توزیع‌های به دست آمده توسط تخمین‌گر بیزی ساده

برای آنکه بتوان از کلاس GMM نوشته شده برای سؤالات قبلی استفاده کرد یک کلاس به نام GaussianMixture همانند آبجکت موجود در کلاس سای‌کیت‌لرن پیاده سازی می‌کنیم. نام متغییرهای غیرخصوصی این کلاس مانند means، covariances و weights می‌گذاریم. متدهای زیر برای اجرای الگوریتم EM باید پیاده سازی شوند.

- constructor: آرایه‌های نام‌پای با طول متناسب با ابعاد فیچر و تعداد مؤلفه‌های میکسچر مدل برای میانگین، واریانس و وزن توزیع‌ها تشکیل می‌دهد.
- fit: داده‌های ترین را به همراه تعداد ایتریشن می‌گیرد و به طور متوالی E Step و M Step را انجام می‌دهد.
- e step: در این متد متغییر z که همان latent variable در الگوریتم EM می‌باشد را با استفاده از تخمین کنونی پارامترهای میکسچر مدل به روز رسانی میذکنیم.
- m step: در این متد پارامترهای مدل (μ_i، Σ_i و p_i) را با استفاده از آخرین تخمین به دست آمده از متغییر دیده نشده محاسبه و به روزرسانی می‌کنیم.
- log likelihood: لوگ‌لایکلیهود را برای داده‌های آرگومان ورودی بر حسب پارامترهای فعلی مدل محاسبه می‌کند. این تابع برای محاسبه AIC و BIC لازم می‌شود.
- aic: مقدار AIC را با استفاده از رابطه

$$\mathrm{AIC}(G) = -2\log \ell_o\left(\hat{\theta}_G; G\right) + 2v_G$$

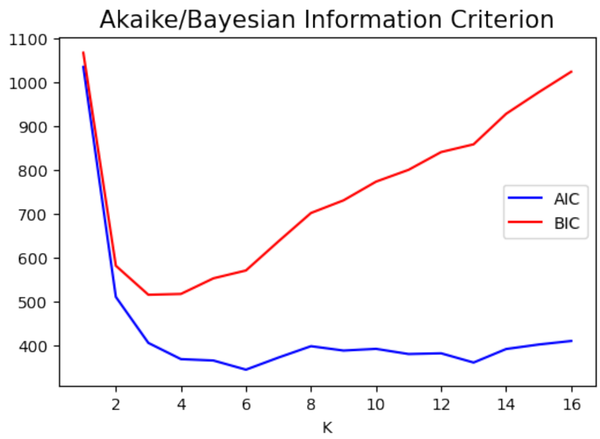
محاسبه می‌کند

- bic: مقدار BIC را با استفاده از رابطه

$$\mathrm{BIC}(G) = -2\log \ell_o\left(\hat{\theta}_G; G\right) + v_G \log(n)$$

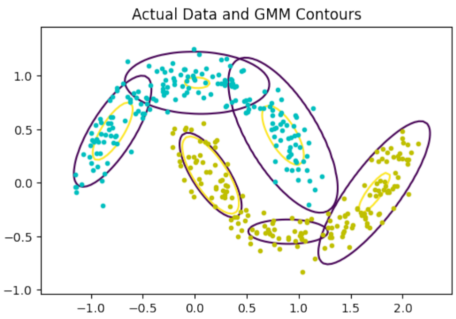
محاسبه می‌کند.

حال برای تشخیص دادن بهترین میکسچر مدل برای طبقه‌بند این سوال تعداد مؤلفه‌های مکیشچر مدل طبقه‌بند را از ۲ تا ۱۶ تغییر می‌دهیم و با رسم نمودار AIC و BIC بهینه‌ترین تعداد برای مؤلفه‌ها را می‌یابیم. این اطلاعات در شکل زیر نمایش داده شده‌اند.

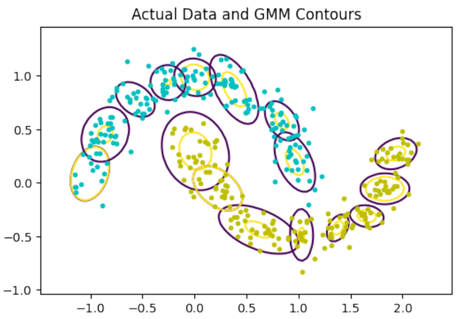


شکل ۷: تاثیر تعداد مؤلفه‌های میکسچرمدل گاوسی

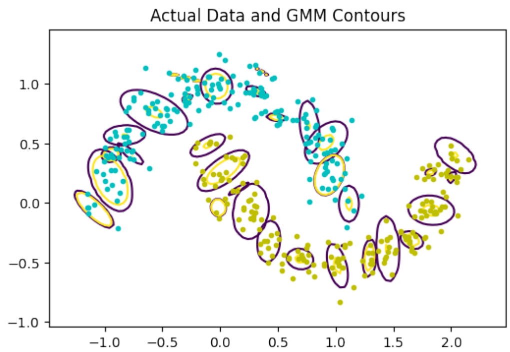
برای رسم نمودارها و طبقه‌بندی در این سوال نیز همان کلاس پیاده سازی شده در سوال ۸ استفاده شده است.



k = 3 (آ)



k = 8 (ب)



k = 16 (ج)

شکل ۸: کانتور توزیع‌های گاوسی مربوط به میکسچرمدل با تعداد مؤلفه‌های متفاوت