Mohammad Javad Ranjbar

810101173

Homework 1

Machine learning, Fall 2022

سوال ۱)

محمد جواد رنجبر ۸۱۰۱۰۱۱۷۳

تمرین سری اول یادگیری ماشین

سوال ۱

$$R(w_1|x) = \lambda_{11} P(w_1|x) + \lambda_{12} P(w_2|x) = \lambda_{12} P(w_2|x)$$

$$R(w_2|x) = \cancel{\lambda_{22}} P(w_2|x) + \lambda_{21} P(w_1|x) = \lambda_{21} P(w_1|x)$$

$$\lambda_{12} P(w_2|x) = \lambda_{21} P(w_1|x) \Rightarrow \lambda_{12} P(x|w_2) P(w_2) = \lambda_{21} P(x|w_1) P(w_1)$$

$$\frac{P(w_2)}{P(w_1)} = \frac{\lambda_{21}}{\lambda_{12}} \frac{P(x|w_2)}{P(x|w_1)} \Rightarrow \frac{\lambda_{12} P(w_2)}{\lambda_{21} P(w_1)} = \frac{\frac{1}{\sqrt{2\pi}\sigma} \times e^{\frac{-1}{2\sigma^2}(x-\mu_2)^2}}{\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x-\mu_1)^2}}$$

$$\xrightarrow{\ln} \ln\left(\frac{\lambda_{12} P(w_2)}{\lambda_{21} P(w_1)}\right) = -\frac{1}{2\sigma^2}(x-\mu_2)^2 - \left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right) \Rightarrow$$

$$\cancel{-x^2} + 2\mu_2 x - \mu_2^2 + \cancel{x^2} - 2\mu_1 x + \mu_1^2 = 2\sigma^2 \ln\left(\frac{\lambda_{12} P(w_2)}{\lambda_{21} P(w_1)}\right)$$

$$x(2\mu_2 - 2\mu_1) + (\mu_1^2 - \mu_2^2) \Rightarrow x = \frac{2\sigma^2 \ln\left(\frac{\lambda_{12} P(w_2)}{\lambda_{21} P(w_1)}\right) - (\mu_1^2 - \mu_2^2)}{2\mu_2 - 2\mu_1}$$

$$\xrightarrow{\mu_2 = 0,\ \mu_1 = 1} x = \frac{1}{2} - \sigma^2 \ln\left(\frac{\lambda_{12} P(w_2)}{\lambda_{21} P(w_1)}\right)$$

سوال ۲)

سوال ۲

الف)

$$P(error) = P(y \neq \hat{y}) = 1 - P(y = \hat{y}) = 1 - \sum_{i=1}^{m} \overbrace{P(x \in R_i, w_i)}^{P(x \in R_i | w_i) P(w_i)}$$

$$P(error) = 1 - \sum_{i=1}^{m} P(x \in R_i | w_i) P(w_i) = 1 - \sum_{i=1}^{m} \left(\int_{R_i} P(x|w_i) P(w_i) \, dx \right)$$

از آنجا که برای هر x کلاس وقتی انتخاب می‌شود که شرطش در آن کلاس بیشترین باشد به عبارتی کلاس را انتخاب می‌کند اگر $P(x|w_i)P(w_i) > P(x|w_j)P(w_j) \quad \forall j \neq i$ پس در عبارت بالا $\int_{R_i} P(x|w_i)P(w_i)dx$ بزرگترین عدد ممکن را خواهد داشت که در نتیجه $P(error)$ را به کوچکترین حالت ممکن می‌رساند

ب) از آنجا که $\sum_{i=1}^{m} P(w_i|x) = 1$ حداقل یک $P(w_i|x)$ باید موجود باشد که بزرگتر مساوی $\frac{1}{m}$ باشد وگرنه این جمع یک نخواهد شد حال فرض می‌کنیم $P(w_k|x)$ نتیجه تصمیم ما است پس داریم که $P(w_k|x) = \max_i P(w_i|x)$ و با توجه به اینکه داریم $P(w_k|x) \geq \frac{1}{m}$ می‌توان نتیجه گرفت

$$P(error) = 1 - \max_i P(w_i|x) \leq 1 - \frac{1}{m} = \frac{m-1}{m}$$

ج) ROC به صورت پایه برای دو کلاس است. برای اینکه برای تعداد بیشتری کلاس استفاده کنیم می‌توانیم از روش‌های همچون one-vs-rest استفاده کنیم. به عبارتی هربار منحنی را برای یک کلاس در برابر سایر کلاس‌ها رسم می‌کنیم. همچنین روش‌های دیگری مانند میانگین یک کلاس در برابر بقیه یا باهم در نظر گرفتن تمام نتایج مثبت و منفی در یک کلاس نیز برای این کار موجود است

د) خیر لزوماً عملکرد بهینه نخواهد داشت. naive bayse در صورتی بهینه عمل خواهد کرد که ویژگی‌ها با شرط کلاس از هم مستقل باشند (conditionally independent) ولی در اینجا فقط استقلال داده ای ویژگی‌ها ذکر شده است که لزوماً جواب بهینه نخواهیم داشت مثلاً دو ویژگی A و B را کاملاً مستقل بگیرید و C کلاس ما باشد که داریم $A+B=C$ حال دیگر A و B با شرط C از یکدیگر مستقل نیستند

سوال ۳)

سوال ۳:

$$l(\mu) = \left(\prod_{i=1}^{N} P(x_i|\mu,\sigma^2)\right) P(\mu) \xrightarrow{\log} \ln l(\mu) = \ln P(\mu) + \sum_{i=1}^{N} \ln \left(P(x_i|\mu,\sigma)\right)$$

$$\xrightarrow[P(x_i|\mu)=\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}]{P(\mu)=\cdots} L(\mu) = \ln\left(\frac{\mu e^{-\frac{\mu^2}{2\sigma_\mu^2}}}{\sigma_\mu^2}\right) + \sum_{i=1}^{N} \left(\ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)\right) \Longrightarrow$$

$$L(\mu) = \ln(\mu) - \frac{\mu^2}{2\sigma_\mu^2} - \ln(\sigma_\mu^2) + \sum_{i=1}^{N}\left(\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \ln(\mu) - \frac{\mu^2}{2\sigma_\mu^2} - \ln(\sigma_\mu^2) + N \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

حال برای حداکثر شدن L نسبت به μ مشتق می‌گیریم و برابر صفر قرار می‌دهیم تا μ_{MAP} بدست آوریم:

$$\frac{dL}{d\mu} = \frac{1}{\mu} - \frac{2\mu}{2\sigma_\mu^2} - 0 + 0 + \overbrace{\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x-\mu)}^{\frac{1}{\sigma^2}\sum_{i=1}^{N} x - \frac{N}{\sigma^2}\mu} = 0 \xrightarrow{\times\mu}$$

$$\overbrace{\left(\frac{1}{\sigma_\mu^2} + \frac{N}{\sigma^2}\right)}^{R}\mu^2 - \overbrace{\left(\frac{1}{\sigma^2}\right)\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)}^{Z}\mu - 1 = 0$$

حال داریم

$$R = \frac{N}{\sigma^2} + \frac{1}{\sigma_\mu^2}, \quad Z = \frac{1}{\sigma^2}\sum_{k=1}^{N} x_k$$

$$\to R\mu^2 - Z\mu - 1 = 0 \Longrightarrow \frac{+Z \pm \sqrt{Z^2 + 4R}}{2R} = \frac{Z}{2R}\left(1 \pm \sqrt{1 + \frac{4R}{Z}}\right)$$

که فقط جواب مثبت قابل قبول است

سوال ۴)

سوال ۴)

الف)

$$P(x; p_i) = \prod_{j=1}^{N} p_i^{x_j}(1-p_i)^{1-x_j} \Longrightarrow L(p_i) = \ln(P(x; p_i)) = \sum_{j=1}^{N} x_j \ln p_i + (1-x)\ln(1-p_i)$$

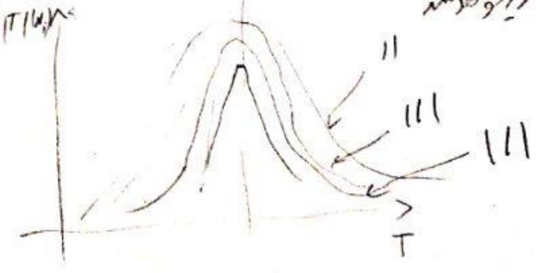
برای پیدا کردن تخمین ML نسبت به p_i مشتق می‌گیریم

$$\frac{\sum_{j=1}^{N} x_j}{p_i} - \frac{\sum_{j=1}^{N}(1-x_j)}{1-p_i} = 0 \Longrightarrow -\frac{N}{1-p_i} + \sum_{j=1}^{N} x_j\left(\frac{1}{p_i} + \frac{1}{1-p_i}\right) = 0$$

$$\left(\frac{1-p_i+p_i}{p_i(1-p_i)}\right)\sum_{j=1}^{N} x_i = +\frac{N}{1-p_i} \Rightarrow \frac{1}{p_i} = \frac{N}{\sum_{j=1}^{N} x_i} \Rightarrow \hat{p} = \frac{1}{N}\sum_{j=1}^{N} x_i$$

ب) با نزدیک شدن α به بی نهایت احتمال به P خواهد رفت و عبارت $\hat{p} = \frac{\sum x_i}{\alpha} = \frac{p\alpha}{\alpha} = p$ خواهد شد و از آنجا که توزیع برنولی فقط با P توصیف می‌شود با بی نهایت شدن تعداد نمونه‌ها به صورت دقیق می‌توانیم مقدار p حساب کنیم در نتیجه خطا صفر خواهد شد

ج) مشخص است که با افزایش مقدار T نیز به P نزدیک و نزدیکتر خواهد شد و درصورتی که این توزیع نورمال داشته باشند شکل نهایی به صورت زیر خواهد شد



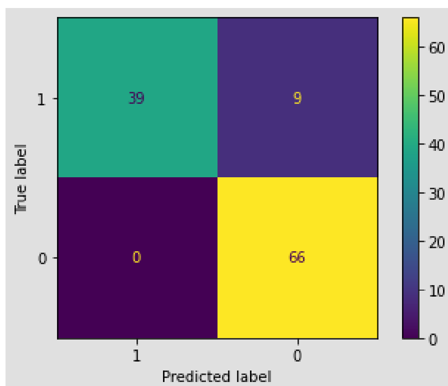
سوال ۵)

الف) قضیه بیز روشی برای محاسبه احتمالات شرطی ارائه می دهد که احتمال پسین نامیده می شود. که محتمل ترین فرضیه را که مجموعه داده آموزشی را توصیف می کند، پیدا می کند. Optimal baye یک مدل احتمالی است که محتمل‌ترین پیش‌بینی را با استفاده از داده‌های آموزشی و فضای فرضیه‌ها برای پیش‌بینی یک نمونه داده جدید پیدا می‌کند. در naïve bayes ما فرض برا conditional independence ویژگی‌ها می‌گذاریم یعنی برای هر ویژگی $P(X|Y,Z)=P(X|Z)$ برقرار است، اما در optimal bayes ویژگی‌ها ممکن است دارای وابستگی شرطی باشند.

ب)

با پیاده سازی naïve bayes، جدول زیر بدست می‌آید. که دارای درصدهای زیر می‌باشد:

```
Naive Bayes classification accuracy 0.9210526315789473
Precision score: 0.88
Recall score: 1.0
```

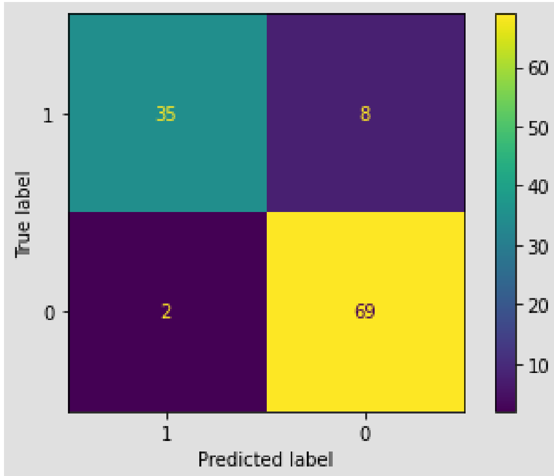


در کلاس ۰ داده‌های بیشتری موجود هستند و این مدل تعدادی از داده‌های دسته‌ی یک را به کلاس صفر نسبت داده است در حالی که هیچ داده‌ای از کلاس صفر به کلاس یک نسبت داده نشده است. در کل در این دیتاست اگر ۱ به معنی داشتن سرطان باشد، ماتریس بالا به شدت بد بوده زیرا ۹ نفر از کسانی که سرطان داشته‌اند به اشتباه نداشته اعلام شده اند.

پیاده‌سازی با کتابخانه:

```
Naive Bayes classification accuracy 0.9122807017543859
Precision score: 0.8961038961038961
Recall score: 0.971830985915493
```

که نتیجه دقیقا مانند حالت بالاست.

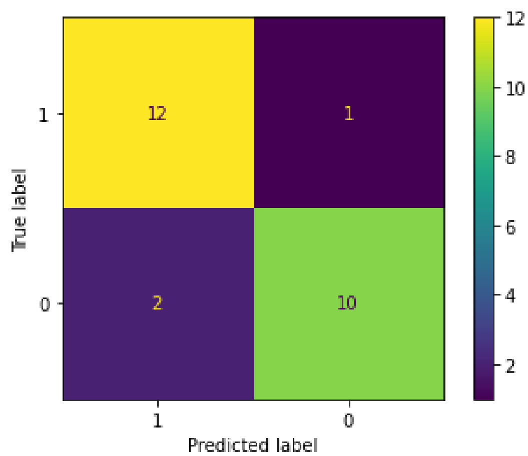


ج) پیاده‌سازی بیزین آپتیمال امکان پذیر نیست، زیرا باید توزیع داده را به صورت دقیق داشته باشیم و رابطه‌ی بین ویژگی‌ها برای شخصی که علم ندارد نیز به راحتی قابل تشخیص نیست.

سوال ۶) هر عکس از سه ماتریس آبی، سبز، قرمز تشکیل شده است. حال می‌توان برای سادگی ویژگی‌های ماتریس را کم کرد و تنها میانگین اعداد هرکدام از این سه ماتریس را برای آموزش به مدل داد.

در اینجا از دو مدل استفاده کرده‌ایم یکی مدل naïve bayes که با استفاده از کتابخانه پیاده شده است و نتیجه زیر را می‌دهد:

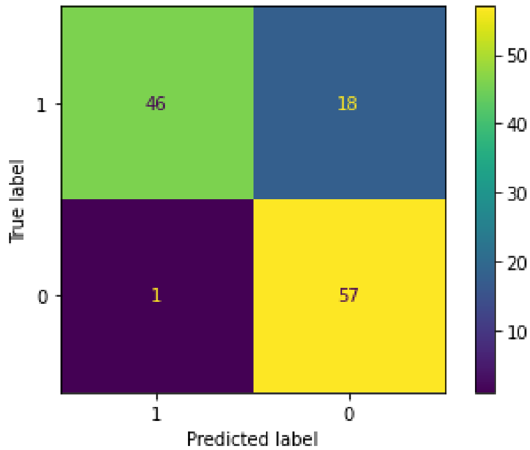
```
Naive Bayes classification accuracy 0.88
Precision score: 0.9090909090909091
Recall score: 0.8333333333333334
```



که در این مورد خطای کلاس صفر که لیبل یک پیش‌بینی می‌کند زیاد است. از آنجا که این مدل برای دیتاست عکس فوتبال آموزش دیده است. این خطا زیاد موردی ندارد.

مدل دوم که به سادگی هرجا میانگین رنگ مربوط به تیم بالاتر بود آن تیم انتخاب می‌شود، نتیجه زیر را خواهد داد:

```
Color classification accuracy: 0.8442622950819673
Precision score: 0.76
Recall score: 0.9827586206896551
```



با وجود پاسخ نسبتا مناسب این نوع پیاده سازی در عکس‌های موجود در دیتاست، برای عکس‌های نوع دیگر این روش قطعا خطای بالایی خواهد داشت.