به نام خدا

سیستم‌های هوشمند (دکتر رشاد حسینی)

تمرین سری **پنجم**

نیم‌سال اول ۱۴۰۲-۱۴۰۱

عرفـان پـنـــاهی ۸۱۰۱۹۸۳۶۹

این فایل شامل گزارش و نتایج شبیه سازی های انجام شده است.

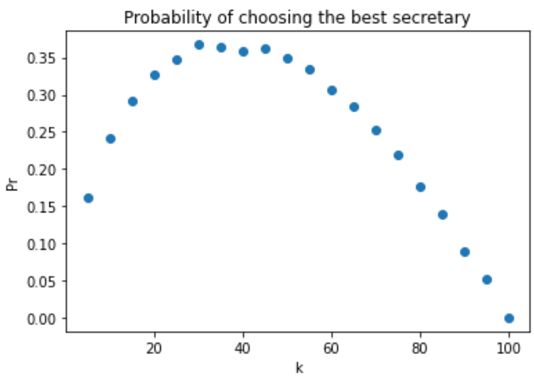
***** فایل شبیه سازی با پایتون مربوط به هر قسمت این تمرین با عنوان HW5_Q#x_810198369.ipynb پیوست شده است.**

سوال ۱: سوالات آماری (پیاده‌سازی)

*** بخش ۱: مسئله منشی

سوال ۱- بدست آوردن مناسب ترین مقدار k برای تعداد کاندیدای ثابت

ابتدا برای الگوریتم راه حل مسئله، تابعی با نام find_Pr() می‌نویسیم. این تابع در حقیقت تعداد آزمایش ها، مقادیر n و k را می‌گیرد و در نهایت احتمال را برای مقادیر n و k برمی‌گرداند. در این سوال مقدار n را به صورت ثابت $n = 100$ در نظر می‌گیریم و مقدار k را به صورت $k = range(5, 101, 5)$ تعریف می‌کنیم. در نهایت نیز به ازای مقادیر k، احتمالات آن را با استفاده از scatter plot رسم می‌کنیم. تصویر ۱-۱ این نمودار را نشان می‌دهد. همانطور که مشاهده می‌شود، بهترین مقدار k تقریباً در حدود ۳۰ تا ۴۰ می‌باشد. (در سوال ۳، ثابت می‌شود که بهترین مقدار k برابر $k = \frac{n}{e} = \frac{100}{e} \cong 36.78$ با احتمال $Pr = \frac{1}{e} = 0.3678$ خواهد بود.)

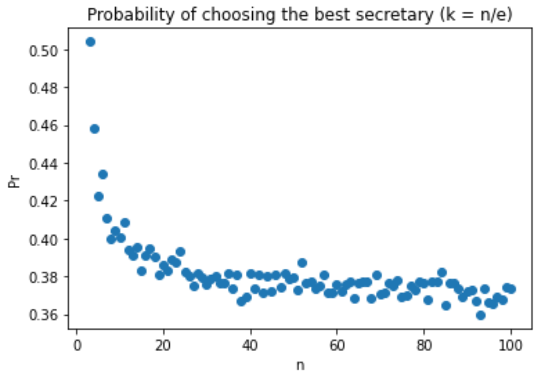


تصویر ۱-۱: نمودار احتمال انتخاب بهترین کاندید برحسب مقادیر k

سوال ۲- بدست آوردن احتمال انتخاب بهترین کاندید به ازای $k = \frac{n}{e}$

در این سوال مقادیر n را از ۳ تا ۱۰۰ متغیر و به صورت $n = range(3, 100)$ و مقدار k را به صورت $k = \frac{n}{e}$ در نظر می‌گیریم. در نهایت به ازای مقادیر n، احتمالات آن را با استفاده از scatter plot رسم می‌کنیم. تصویر ۱-۲ این نمودار را

نشان می‌دهد. همانطور که مشاهده می‌شود با افزایش مقدار n احتمال انتخاب بهترین کاندید کاهش می‌یابد اما به ازای n های بزرگ تقریباً احتمال ثابت و برابر $\frac{1}{e} = 0.3678$ می‌ماند.



تصویر ۱-۲: نمودار احتمال انتخاب بهترین کاندید برحسب مقادیر n

سوال ۳- بدست آوردن احتمال انتخاب بهترین کاندید (تحلیلی)

برای حل این مسئله از احتمال شرطی استفاده می‌کنیم. در حقیقت ما می‌خواهیم احمال انتخاب بهترین کاندید از بین $n-k$ کاندید بعد از k کاندید اول را بدست آوریم.

$$Pr\left(\text{احتمال بهترین کاندید با مرز } k\right) = \sum_{i=1}^{n} \Pr\left(\text{کاندید } i\text{ام بهترین باشد } \cap \text{ کاندید } i\text{ام انتخاب شود}\right)$$

$$= \sum_{i=1}^{n} \Pr\left(\text{کاندید } i\text{ام انتخاب شود} \mid \text{کاندید } i\text{ام بهترین باشد}\right) . \Pr\left(\text{کاندید } i\text{ام بهترین باشد}\right)$$

با توجه به اینکه اگر بهترین کاندید بین $k-1$ نفر اول باشد، احتمال انتخاب آن صفر است یعنی:

$$\sum_{i=1}^{k-1} \Pr\left(\text{کاندید } i\text{ام انتخاب شود} \mid \text{کاندید } i\text{ام بهترین باشد}\right) = 0$$

به عبارت دیگر می‌توانیم بگوییم، کاندید i ام باید فقط بهتر از بهترین کاندید بین $k-1$ کاندید اول باشد (یعنی بهترین کاندید بین $i-1$ کاندید اول و $k-1$ کاندید اول یکی باشد.) پس اگر $A_{k,i}$ را به صورت زیر تعریف کنیم.

$$A_{k,i} = \text{بهترین کاندید } i-1 \text{ کاندید اول همان بهترین کاندید بین } k-1 \text{ کاندید اول باشد}$$

$$Pr\left(\text{انتخاب بهترین کاندید}\right) = \sum_{i=k}^{n} \underbrace{\Pr\left(A_{k,i} \mid \text{کاندید } i\text{ام بهترین باشد}\right)}_{\frac{\binom{k-1}{1}}{\binom{i-1}{1}} = \frac{k-1}{i-1}} . \underbrace{\Pr\left(\text{کاندید } i\text{ام بهترین باشد}\right)}_{\frac{1}{n}} = \frac{1}{n}\sum_{i=r}^{n}\frac{k-1}{i-1}$$

$$\rightarrow \boldsymbol{Pr(k) = \frac{k-1}{n}\sum_{i=r}^{n}\frac{1}{i-1}}$$

در حالت $n \to \infty$ ، احتمال پیوسته به صورت زیر تعریف می‌شود.

$$Pr(k) = \lim_{n\to\infty} \frac{k-1}{n} \sum_{i=k}^{n} \frac{1}{i-1} \xrightarrow{\frac{k}{n}=x} \Pr(x) = x\int_{x}^{1} \frac{1}{t}dt = -xln(x) \to \boldsymbol{P(x) = -xln(x)}$$

حال می‌خواهیم ببینیم به ازای چه مقدار x احتمال $P(x)$ ماکزیمم می‌شود.

$$\frac{d}{dx}P(x) = \frac{d}{dx}\big(-xln(x)\big) = -lnx - 1 = 0 \to \ln(x) = -1 \to x^{opt} = \frac{1}{e}$$

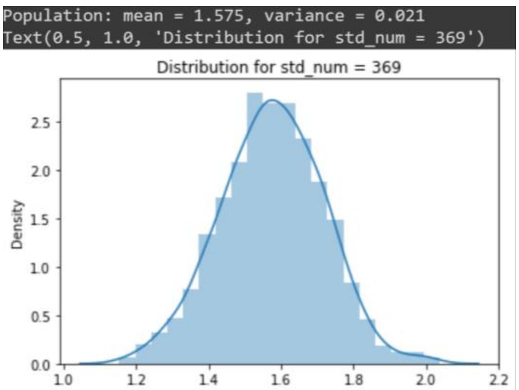
$$x = \frac{k}{n} \to k^{opt} = nx^{opt} = \frac{n}{e} \to \boldsymbol{k^{opt} = \frac{n}{e}}$$

*** بخش ۲: بررسی قضیهٔ حد مرکزی

قضیهٔ حد مرکزی: دنباله $X_1, X_2, \dots, X_n$ از متغیر های تصادفی مستقل با توزیع یکنواخت پیوسته D را که بر یک فضای احتمال تعریف شده اند، در نظر بگیرید. فرض کنید میانگین D برابر μ و انحراف معیار از آن σ است. حالا سری میانگین S_n را به صورت $S_n = X_1 + X_2 + \cdots + X_n$ در نظر بگیرید. در این صورت میانگین S_n برابر μ و انحراف معیار از آن $\frac{\sigma}{\sqrt{n}}$(واریانس $\frac{\sigma^2}{n}$) و براساس قضیهٔ حدمرکزی S_n در بی‌نهایت به توزیع نرمال $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ میل می‌کند.

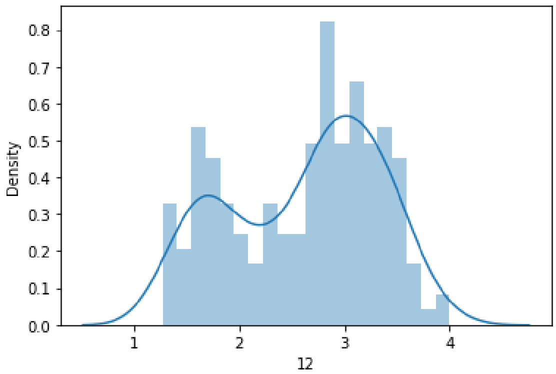
در این بخش می‌خواهیم با استفاده از توزیع اولیه X، دنباله $X_1, X_2, \dots, X_n$ را تعریف کرده و از هر عضو این این دنباله میانگین بگیریم و توزیع S_n را بدست آوریم. در این صورت انتظار داریم طبق قضیه حدمرکزی S_n در بی‌نهایت به توزیع نرمال $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ میل کند.

ابتدا ارقام ۳۶۹ را به تابع Distribution() می‌دهیم. در نهایت با استفاده از تابع Dist() که برای به دست آوردن میانگین هر مرحله (s مرحله) با استفاده از n نمونه نوشته شده است (بدست آوردن S_n)، توزیع جمعیت را به همراه میانگین و واریانس آن مطابق با تصویر ۱-۳ نشان می‌دهیم. همانطور که مشاهده می‌شود توزیع شکل نرمال پیدا کرده است.



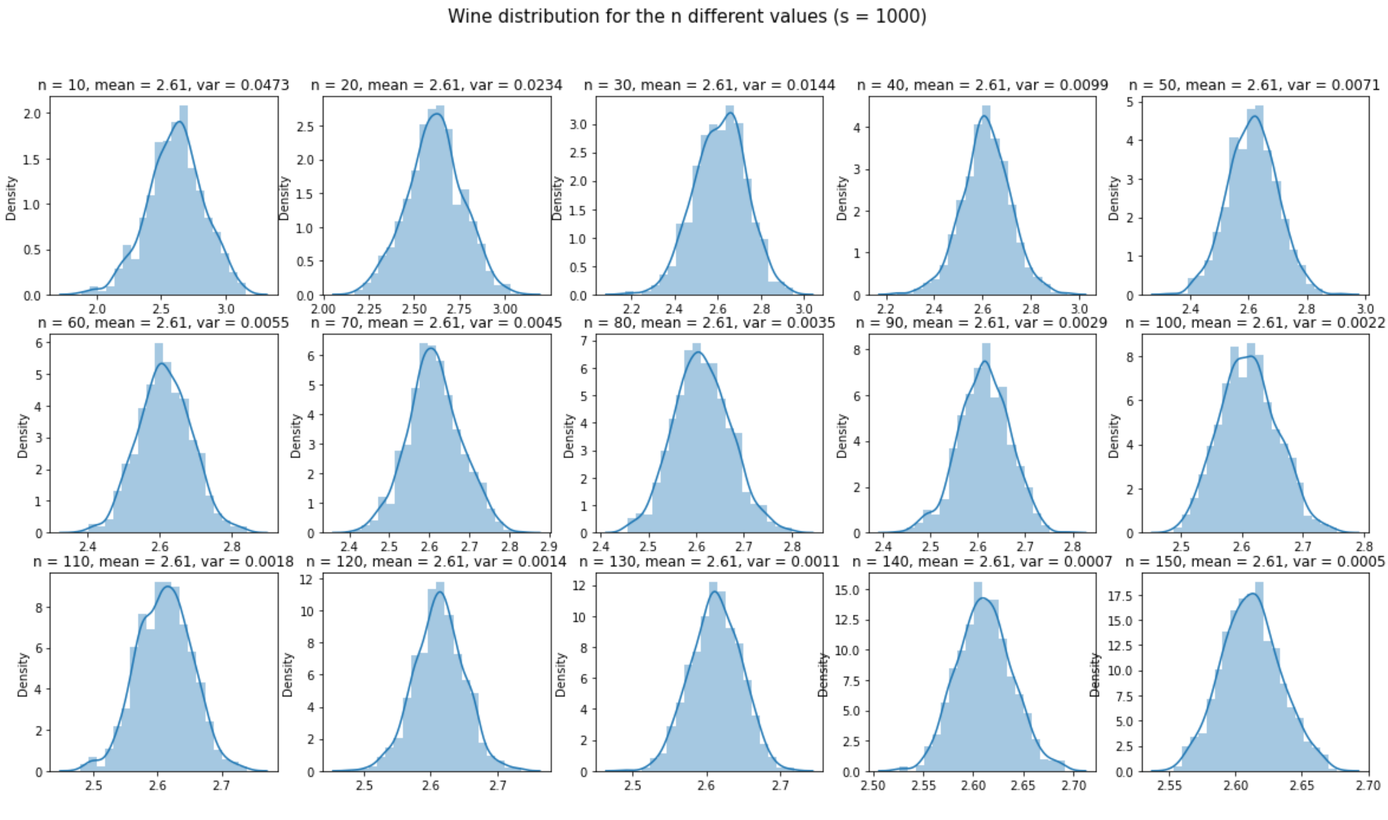
تصویر ۱-۳: توزیع میانگین نمونه های جمعیت ($s = 1000$ و $n = 1000$)

حال فایل پیوست شده در تمرین با عنوان wine.csv را در پایتون می‌خوانیم و توزیع ستون ۱۲ آنرا مطابق با تصویر ۱-۴ نشان می‌دهیم. همانطور که مشاهده می‌شود این توزیع در ابتدا نرمال نیست.



تصویر ۱-۴: توزیع داده *wine*

حال با استفاده از تابع Dist() به ازای مقادیر مختلف n از داده ها نمونه می‌گیریم تا توزیع های مستقل بسازیم و در نهایت از آن‌ها در ۱۰۰۰ بار تکرار میانگین می‌گیریم. تصویر ۱-۵ نمودار توزیع های گفته شده به ازای n های مختلف را نشان می‌دهد.



تصویر ۱-۵: توزیع میانگین نمونه های *wine* ($s = 1000$ و n های مختلف)

همانطور که در تصویر ۱-۵ نیز مشاهده می‌شود، توزیع میانگین نمونه ها نرمال شده است و میانگین آن به ازای همهٔ n ها شکل نرمال دارد. همچنین مشاهده می‌شود که واریانس با افزایش n نزولی است. به این منظور مطابق با تصویر ۱-۶ میانگین و واریانس توزیع ها به ازای n های مختلف نشان می‌دهیم. همانطور که مشاهده می‌شود رابطهٔ $\sigma_{new}^2 = \frac{\sigma^2}{n}$ کاملاً برقرار است.

```
Central Limit Theorem Analysis:
n = 10:
        mean = 2.598, variance = 0.04849
CLT:    mean = 2.612, variance = 0.05013
n = 20:
        mean = 2.615, variance = 0.02041
CLT:    mean = 2.612, variance = 0.02506
n = 30:
        mean = 2.612, variance = 0.01391
CLT:    mean = 2.612, variance = 0.01671
n = 40:
        mean = 2.616, variance = 0.00994
CLT:    mean = 2.612, variance = 0.01253
n = 50:
        mean = 2.612, variance = 0.00716
CLT:    mean = 2.612, variance = 0.01003
n = 60:
        mean = 2.612, variance = 0.00575
CLT:    mean = 2.612, variance = 0.00835
n = 70:
        mean = 2.615, variance = 0.00434
CLT:    mean = 2.612, variance = 0.00716
n = 80:
        mean = 2.614, variance = 0.00337
CLT:    mean = 2.612, variance = 0.00627
n = 90:
        mean = 2.608, variance = 0.00292
CLT:    mean = 2.612, variance = 0.00557
n = 100:
        mean = 2.612, variance = 0.00204
CLT:    mean = 2.612, variance = 0.00501
n = 110:
        mean = 2.612, variance = 0.00161
CLT:    mean = 2.612, variance = 0.00456
n = 120:
        mean = 2.613, variance = 0.00134
CLT:    mean = 2.612, variance = 0.00418
n = 130:
        mean = 2.612, variance = 0.001
CLT:    mean = 2.612, variance = 0.00386
n = 140:
        mean = 2.611, variance = 0.00075
CLT:    mean = 2.612, variance = 0.00358
n = 150:
        mean = 2.611, variance = 0.00053
CLT:    mean = 2.612, variance = 0.00334
```

تصویر ۱-۶: میانگین و واریانس توزیع میانگین نمونه های $wine$ ($s = 1000$ و n های مختلف)

سوال ۲: سوالات آماری (تحلیلی)

*** بخش ۱: فرض استقلال در بیز ساده انگارانه

تعریف استقلال: اگر دو متغیر تصادفی مستقل باشند، رابطه زیر بین چگالی توام آن‌ها برقرار است.

$$f_{XY}(x,y) = f_X(x)f_Y(y) \rightarrow X\ and\ Y\ are\ independent$$

ناهمبستگی: در صورتی که ضریب همبستگی دو متغیر تصادفی که به صورت زیر تعریف می‌شود، صفر باشد، این دو متغیر ناهمبسته هستند.

$$\rho = E\left[\frac{(X - E(X))}{\sqrt{E\left[(X - E(X))^2\right]}} \frac{(Y - E(Y))}{\sqrt{E\left[(Y - E(Y))^2\right]}}\right] = 0 \rightarrow X\ and\ Y\ are\ uncorrelated$$

تعامد: تعامد مفهومی جبری است که در آمار و احتمالات با همان ناهمبستگی مدل می‌شود.

حال باید به دو نکته توجه می‌کنیم:

۱- استقلال دو متغیر تصادفی ناهمبسته بودن آن‌ها را نشان می‌دهد.

۲- ناهمبسته بودن دو متغیر تصادفی لزوماً استقلال آن‌ها را نتیجه نمی‌دهد.

← در فضای اولیه ویژگی ها مستقل نیستند. در فضای جدید ویژگی ها متعامد و ناهمبسته اند اما الزامی وجود ندارد که در این فضا ویژگی‌ها مستقل باشند. در نتیجه فرض بیز ساده انگارانه در فضای جدید نیز صحیح نیست و استدلال احمد اشتباه است.

*** بخش ۲: محاسبه پارامترهای نمونه‌های $i.i.d$ توزیع نرمال با استفاده از MLE

تخمین Maximum Likelihood این مسئله به صورت زیر تعریف می‌شود.

$$i.i.d\ \ variables \rightarrow\ (X_1, X_2, \dots, X_n)$$

$$\Rightarrow f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_n}(x_n)$$

$$\Rightarrow f_{X_1,X_2,\dots,X_n|\theta}(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta) = \mathcal{L}(\theta)$$

$$\hat{\theta}_{ML} = \max_{\theta} \mathcal{L}(\theta) = \max_{\theta} \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

که برای بدست آوردن $\hat{\theta}_{ML}$ در این بخش، مشتق $\mathcal{L}(\theta)$ برحسب θ را صفر قرار می‌دهیم. برای سادگی محاسبات مشتق ها از لگاریتم طبیعی $\mathcal{L}(\theta)$ (که آن‌را به صورت $\mathcal{LL}(\theta)$ تعریف می‌کنیم.) استفاده می‌کنیم.

در این بخش می‌خواهیم پارامتر های توزیع نرمال یعنی میانگین و واریانس را بدست آوریم. به این منظور یک‌بار به جای θ میانگین و بار دیگر واریانس قرار دهیم.

$$X_i \sim \mathcal{N}(\mu, \sigma) \rightarrow f_{X_i}(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\mathcal{LL}(\mu, \sigma) = \ln\left(\prod_{i=1}^{n} f_{X_i}(x_i|\theta)\right) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$
$$= \sum_{i=1}^{n}\left(-\ln(\sqrt{2\pi}\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\hat{\mu}_{ML}: \quad \frac{\partial \mathcal{LL}(\mu, \sigma)}{\partial \mu} = 0 \rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML}) = 0 \rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}_{ML}: \quad \frac{\partial \mathcal{LL}(\mu, \sigma)}{\partial \sigma} = 0 \rightarrow \sum_{i=1}^{n}\left(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}\right) = 0 \rightarrow \hat{\sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

سوال ۳: بیز ساده‌انگارانه

در این قسمت ابتدا در خصوص طبقه بند بیز ساده انگارانه توضیحات مختصری می‌دهیم. (برای این توضیحات از یک ویدیو که با لینک زیر استفاده می‌کنیم.)

- https://www.youtube.com/watch?v=Po6lsacF5pw

❄❄❄ طبقه‌بند بیز ساده‌انگارانه (Naïve Bayes Classifier) :

در بیز ساده انگارانه، هدف استفاده از قضیه بیز در طبقه‌بندی داده ها است. قضیه بیز به صورت زیر تعریف می‌شود.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

حال در طبقه بندی داده ها، به فرض اگر داده X را به ما داده باشند، احتمال اینکه X در هر کدام از کلاس ها باشد را بدست آورده و با استفاده از بیشترین احتمال، کلاس داده X را مشخص می‌کنیم.

کلاس ها به صورت C_i و داده های هر ویژگی به صورت x_i مشخص می‌شود (فرض می‌کنیم d ویژگی متمایز داریم).

دو فرض در بیز ساده انگارانه که حائز اهمیت است به صورت زیر می‌باشد:

۱- داده ها در هر کلاس (برای هر ویژگی) از توزیع نرمال پیروی می‌کنند.

۲- داده های هر دو ویژگی متفاوت ناهمبسته اند.

احتمال اینکه داده X عضو کلاس C_i باشد به صورت زیر است.

$$P(C_i|X) = \frac{P(C_i)P(X|C_i)}{P(X)}$$

با توجه به ثابت بودن $P(X)$ و اینکه می‌خواهیم مقایسه بین احتمال های کلاس های مختلف انجام دهیم، از مخرج صرف نظر می‌کنیم.

$$P(C_i|X) \propto P(C_i)P(X|C_i)$$

به $P(C_i)$ احتمال پیشین (Prior) گفته می‌شود. برای بدست آوردن این احتمال صرفاً تعداد لیبل های کلاس C_i بر تعداد کل داده ها تقسیم می‌کنیم. (در داده های آموزش این احتمال را بدست می‌آوریم.)

به $P(X|C_i)$ احتمال راستی آزمایی (Likelihood) گفته می‌شود. برای به دست آوردن این احتمال از دو فرض گفته شده استفاده می‌کنیم.

$$Independent\ Fetures \rightarrow P(X|C_i) = P(x_1|C_i)P(x_2|C_i)\ldots P(x_d|C_i)$$

$$P(x_j|C_i)\ has\ a\ normal\ (gaussian) distribution \rightarrow P(x_j|C_i) = \frac{1}{\sigma_{x_j}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j-\mu_{x_j}}{\sigma_{x_j}}\right)^2}$$

که با استفاده از داده های آموزش باید توزیع همه $P(x_j|C_i)$ را بدست آوریم. یعنی پارامتر های σ_{x_j} را مشخص کنیم.

در نهایت برای داده های تست احتمال تعلق کلاس ها به صورت زیر بدست می‌آید.

$$P(C_i|X) \equiv P(C_i)P(x_1|C_i)P(x_2|C_i) \ldots P(x_d|C_i)$$

$$P(x_j|C_i) = \frac{1}{\sigma_{x_j}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j-\mu_{x_j}}{\sigma_{x_j}}\right)^2}$$

سوال ۱. معرفی ویژگی ها و نحوه بدست آوردن احتمال‌ها

حال می‌خواهیم برای مسئله اشاره شده، طبقه بند بیز ساده انگارانه طراحی کنیم.

به این منظور در ابتدا ویژگی ها را به صورت زیر بدست می‌آوریم.

ویژگی ۱: ارتفاع عدد (Hight)

ویژگی ۲: پهنای عدد (Width)

ویژگی ۳: تعداد # نیمه بالایی عدد (Sharp_num_upper) - برای اعداد {5,7,9} این مقدار بزرگتر است.

ویژگی ۴: تعداد + نیمه بالایی عدد (Plus_num_upper) - برای اعداد {5,7,9} این مقدار بزرگتر است.

ویژگی ۵: تعداد # نیمه پایینی عدد (Sharp_num_below) - برای اعداد {2,4,6} این مقدار بزرگتر است.

ویژگی ۶: تعداد + نیمه پایینی عدد (Plus_num_below) - برای اعداد {2,4,6} این مقدار بزرگتر است.

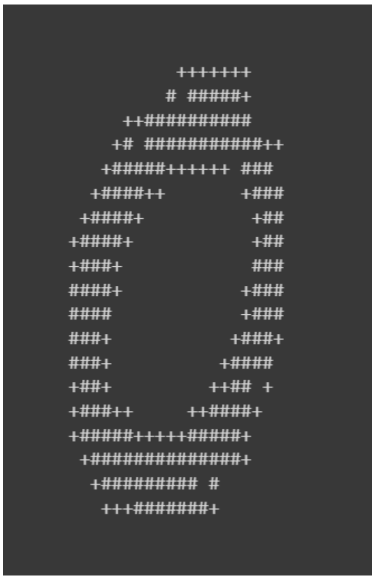
ویژگی ۷: بزرگترین نسبت تعداد # و + به فضا های خالی در سطرهای عکس ها (Row_limit) - تشخیص {1,7,4,5}

حال برای بدست آوردن احتمال های پسین هر کلاس و مقایسه احتمال ها و تشخیص کلاس داده های تست مطابق با توضیحات داده شده برای طبقه بند Naïve Bayes عمل می‌کنیم.

برای فرآیند آموزش تابع Training() نوشته شده است که در آن برای استخراج ویژگی ها از تابع Feature_Extraction() استفاده شده است.

همچنین برای طبقه بندی و فرآیند ارزیابی تابع Validation() نوشته شده است.

تصویر ۳-۱، یک تصویر تصادفی از دادگان را نشان می‌دهد.

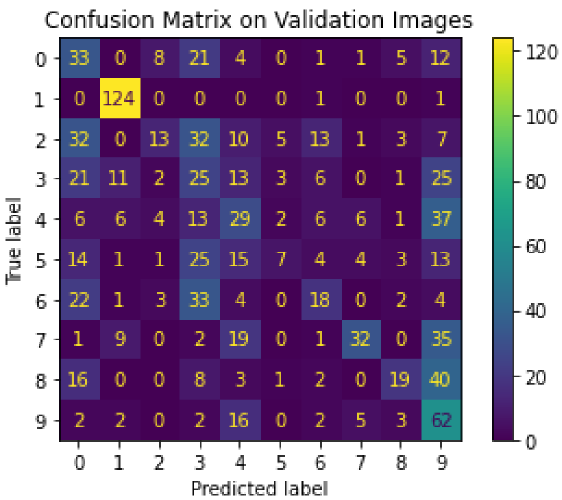


تصویر ۳-۱: یک تصویر تصادفی از دادگان آموزشی

سوال ۲. بدست آوردن درصد دقت تشخیص اعداد

درصد دقت تشخیص اعداد برای الگوریتم گفته شده روی داده های ارزیابی (Validation) به همراه ماتریس آشفتگی (Confusion Matrix) در تصویر ۳-۲ قابل مشاهده است.

```
Naive-Bayes Classifier:
        Accuracy = 36.199999999999996%
```



تصویر ۳-۲: دقت و ماتریس آشفتگی طبقه بند بیز ساده‌انگارانه