فردین عباسی 810199456 به نام خدا تمرین سری دوم یادگیری ماشین

سوال 1

آ)

$$y = \beta_0 + \varepsilon \rightarrow \varepsilon = y - \beta_0$$

$$\sum_i \varepsilon_i^{\ 2} = \sum_i (y_i - \beta_0)^2 \rightarrow \frac{d\sum_i \varepsilon_i^{\ 2}}{d\beta_0} = -\ 2\sum_i (y_i - \beta_0) \rightarrow \beta_0 = \frac{\sum_i y_i}{n} = 59.6$$

ب)

$$y = \beta_1 x + \varepsilon \rightarrow \varepsilon = y - \beta_1 x$$

$$\sum_i \varepsilon_i^{\ 2} = (y_i - \beta_1 x_i)^2 \rightarrow \frac{d\sum_i \varepsilon_i^{\ 2}}{d\beta_1} = -\ 2\sum_i x_i (y_i - \beta_0)^2 \rightarrow \beta_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^{\ 2}} = \frac{9774}{2498} = 3.91$$

ج)

$$\widehat{Y} = b_0 + b_1 X$$
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

در معادله بالایی $\widehat{Y}$ تخمین رگرسیون خطی به ازای X است که مقادیر b_1, b_0 پارامتر های بهترین رگرسیون خطی روی X می باشد اما معادله پایینی رابطه خطی برای داده های **مشاهده شده** X است که ε خطا حاصل از این رابطه هست همچنین پارامتر های β_1, β_0 را از طریق روش های آماری می توان بدست آورد.

د)

$$y_i = 59.6 + \varepsilon_i \rightarrow SSE = \sum_i \varepsilon_i = \sum_i (y_i - 59.6)^2 = 3318.4$$

$$Var = MSE = \frac{SSE}{n-1} = 368.7$$

$$y_i = 3.91x_i + \varepsilon_i \rightarrow SSE = \sum_i \varepsilon_i^2 = \sum_i (y_i - 3.91x_i)^2 = 597$$

$$Var = MSE = \frac{SSE}{n-1} = 66.34$$

ه)

نه الزاما زیرا داده ورودی جدید یک متغیر است که ما با رگرسیون خطی سعی داریم آن را تخمین بزنیم بنابراین احتمال خطا وجود دارد.

سوال 2

الف)

در L1 Regularization ازنرم 1 استفاده می کنیم در نتیجه جواب اسپارس و ممکن است feature selection رخ دهد ولی L2 Regularization از نرم 2 استفاده می کنیم که پاسخ non-sparse می دهد.

$$L(\beta) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2$$

$$L_2(\beta) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda||\beta||_2^2$$

$$L_1(\beta) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda||\beta||_1$$

ب)

$$\widehat{\beta} = argmin \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda||\beta||_2^2 = argmin\,(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

$$\frac{d}{d\beta} \rightarrow 2(X^TX\widehat{\beta} - X^TY + \lambda\widehat{\beta}) = 0 \rightarrow \widehat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

سوال 3

آ)

اگر کلاس k را 1 و بقیه را صفر درنظر بگیریم گویا حالت باینری است و درنتیجه میتوان از logestic regression استفاده کرد:

$$\sum_{k=1}^{K} P(Y = y_k|X) = 1 \Rightarrow P(Y = y_K|X) = 1 - \sum_{k=1}^{K-1} P(Y = y_i|X) \Rightarrow$$

$$P(Y = y_K|X) = \frac{1}{1+\sum_{k=1}^{k-1} exp(w_{k0}+\sum_{i=1}^{d} w_{ki}X_i)}$$

$$for\ k \in \{1,..., K - 1\}:\ P(Y = y_k|X) = \frac{exp(w_{k0}+\sum_{i=1}^{d} w_{ki}X_i)}{1+\sum_{k=1}^{K-1} exp(w_{k0}+\sum_{i=1}^{d} w_{ki}X_i)}$$

ب)

برای طبقه بندی نیز از روش زیر استفاده می کنیم:

$$y = y_k^* \leftrightarrow k^* = argmax\, P(Y = y_k|X)$$

سوال 4

$$\overline{P_n} = E[P_n(x)] = \frac{1}{n}\sum_{i=1}^{n} E[\frac{1}{h_n}\phi(\frac{x-x_i}{h_n})] = \int \frac{1}{h_n}\phi(\frac{x-x_i}{h_n})P(x_i)dx_i \xrightarrow{n\to\infty}$$

$$\frac{1}{h_n}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}exp(\frac{-1}{2}(\frac{x-x_i}{h_n})^2)\frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-1}{2}(\frac{x_i-\mu}{\sigma})^2)dx_i =$$

$$\frac{1}{2\pi h_n\sigma}exp(\frac{-1}{2}(\frac{x^2}{h_n^2}+\frac{\mu^2}{\sigma^2}))\int_{-\infty}^{\infty}exp(\frac{-x_i^2}{2}(\frac{1}{h_n^2}+\frac{1}{\sigma^2}) - 2x_i(\frac{x}{h_n^2}+\frac{\mu}{\sigma^2}))dx_i =$$

$$\frac{1}{\sqrt{2\pi}\sqrt{h_n^2+\sigma^2}}exp(\frac{-1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}) = N(\mu, h_n^2+\sigma^2)$$

##

$$P(x) - \overline{P}_n(x) = \frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-1}{2}(\frac{x-\mu}{\sigma})^2) - \frac{1}{\sqrt{2\pi}\sqrt{h_n^2+\sigma^2}}exp(\frac{-1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}) =$$

$$\frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2})[1 - \frac{\sigma}{\sqrt{\sigma^2+h_n^2}}exp(\frac{-1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}+\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})] =$$

$$P(x)[1 - \frac{1}{\sqrt{1+(\frac{h_n}{\sigma})^2}}exp(-\frac{(x-\mu)^2}{2}(\frac{1}{h_n^2+\sigma^2}-\frac{1}{\sigma^2}))] =$$

$$P(x)[1 - \frac{1}{\sqrt{1+(\frac{h_n}{\sigma})^2}}exp(\frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2})] \to$$

$$\frac{1}{\sqrt{1+(\frac{h_n}{\sigma})^2}} \approx 1 - \frac{1}{2}(h_n/\sigma)^2 \quad , \quad exp(\frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}) \approx 1 + \frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}$$

از درجات بزگتر h_n^2 نیز صرف نظر می کنیم:

$$\approx P(x)[1 - 1 + \frac{h_n^2}{2\sigma^2} - \frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}] \approx P(x)[1 - \frac{(x-\mu)^2}{h_n^2+\sigma^2}]\frac{h_n^2}{2\sigma^2} \approx P(x)[1 - (\frac{x-\mu}{\sigma})^2]\frac{h_n^2}{2\sigma^2}$$

سوال 5

$$D(x, y) = \sqrt{\sum_{k=1}^{d} (x_k - y_k)^2}$$

$$x'_k = a_k x_k\,,\ y_k' = a_k y_k$$

ویژگی های فاصله استاندارد:

$$I:\ D(x', y') = D(y', x'): D(x', y') = \sqrt{\sum_{k=1}^{d} (x'_k - y_k')^2} = \sqrt{\sum_{k=1}^{d} a_k^{\,2} (x_k - y_k)^2} =$$

$$\sqrt{\sum_{k=1}^{d} a_k^{\,2} (y_k - x_k)^2} = \sqrt{\sum_{k=1}^{d} (y_k' - x_k')^2} = D(y', x')$$

$$II: D(x', y') = 0 \leftrightarrow x' = y': D(x', y') = \sqrt{\sum_{k=1}^{d} (x_k' - y_k')^2} =$$

$$\sqrt{\sum_{k=1}^{d} a_k^{\,2} (x_k - y_k)^2} = 0 \Rightarrow x_k = y_k \,\&\, x'_k = y'_k$$

$$III:\ D(x', y') > 0 \leftrightarrow x' \neq y': \sqrt{\sum_{k=1}^{d} (x'_k - y_k')^2} = \sqrt{\sum_{k=1}^{d} a_k^{\,2} (x_k - y_k)^2} \Rightarrow^{x' \neq y'}$$

$$D(x', y') > 0$$

$IV:\ D(x', y')\ +\ D(y', z')\ \geq\ D(x', z')$

$$D^2(x', z')\ =\ \sum_{k=1}^{d} a_k (x\ -\ z)^2 = \sum_{k=1}^{d} a_k [(x\ -\ y)\ +\ (y\ -\ z)]^2 =$$

$$\sum_{k=1}^{d} a_k [(x\ -\ y)^2 +\ 2(x\ -\ y)(y\ -\ z)\ +\ (y\ -\ z)^2] =$$

$$D^2(x', y')\ + D^2(y', z')\ +\ 2\sum_{k=1}^{d} a_k (x\ -\ y)(y\ -\ z)$$

حال براساس نامساوی کوشی-شوآرتز:

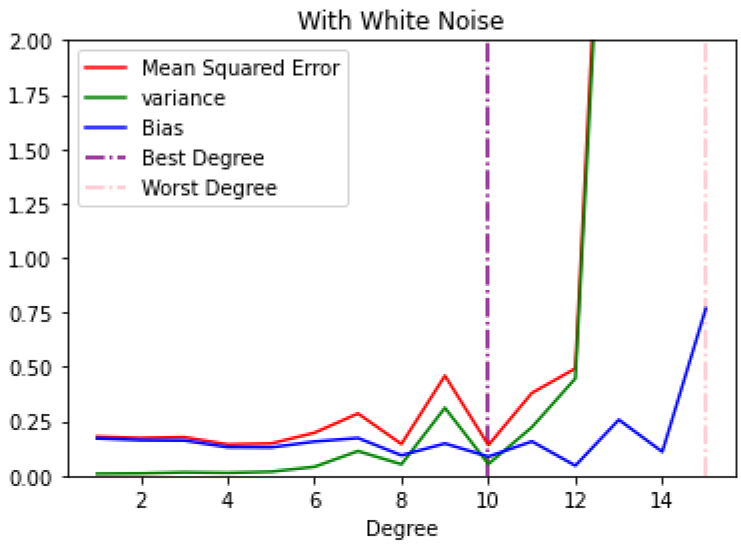
$$|\sum_{k=1}^{d} a_k (x\ -\ y)(y\ -\ z)| \leq \sqrt{\sum_{k=1}^{d} a_k (x\ -\ y)^2}\sqrt{\sum_{k=1}^{d} a_k (y\ -\ z)^2} = D(x', y')D(y', z')$$

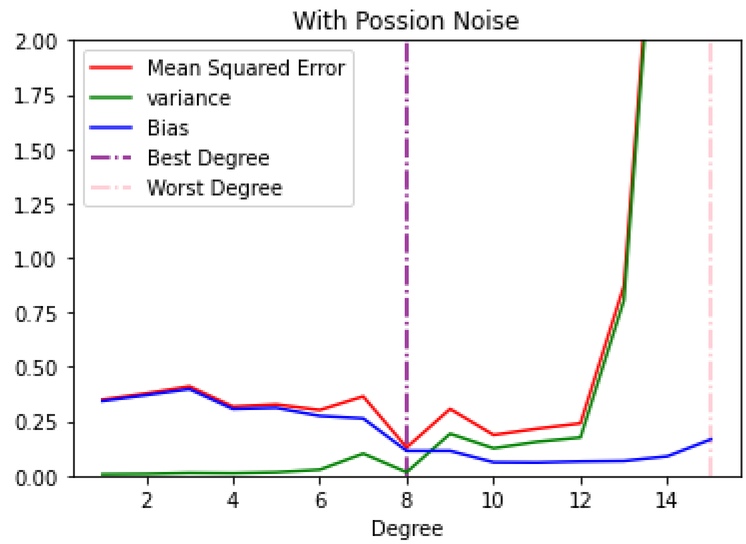
$$\Rightarrow D(x', y')\ +\ D(y', z')\ \geq\ D(x', z')$$

در KNN با توجه به K همسایه اطراف تصمیم گیری می کنیم و پنجره به قدری بزرگ یا کوچک می شود که K همسایه داخل آن بیفتد، با ضرب این ثابت در عناصر متریک میتوان اندازه پنجره را بزرگ یا کوچک کرد.

سوال 6

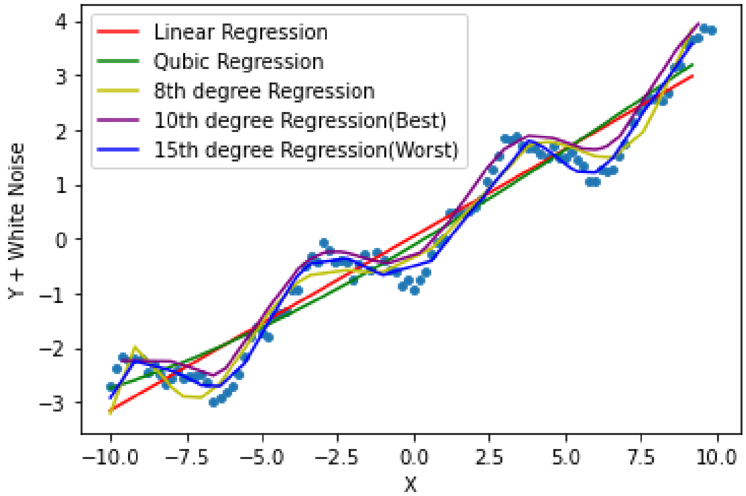
الف)

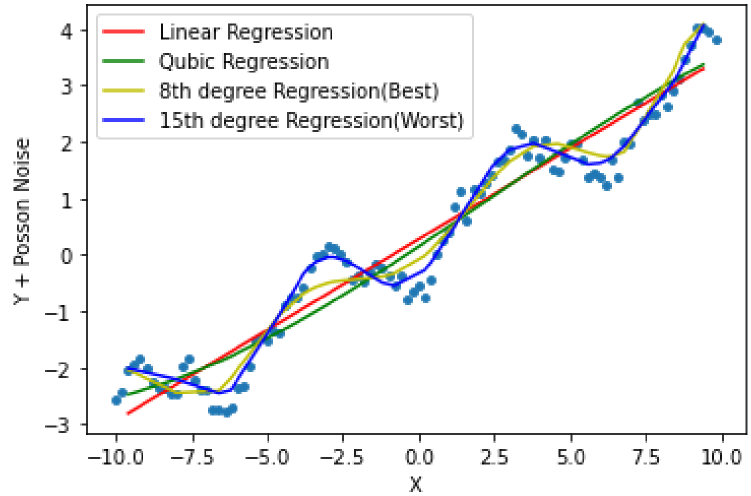




بهترین و بدترین درجه براساس خطای MSE انتخاب می شود. همانگونه که مشاهده می کنید ابتدا به دلیل underfit خطا MSE بیشتر و سپس با افزایش درجه کاهش می یابد و در انتها به دلیل خطا overfit خطا MSE افزایش می یابد. Overfit به دلیل این است که مدل حتی نویز موجود در دیتا را هم آموزش میبیند.
بهترین درجه برای نویز سفید، درجه 10 و بدترین،درجه 15 است. برای نویز پوآسون نیز بهترین درجه 8 و بدترین 15 است.

ب)






```
MSE for White Noise= [0.18319204722067156, 0.17393602021203486, 0.1782841796510229, 0.13495926076568537, 0.13804496359186347,
0.17514674767480995, 0.25898550986552077, 0.14384893364572132, 0.32177592331986576, 0.12967348188694242, 0.24032525000949057,
0.47710290869564254, 2.469162984865544, 8.774991484377372, 394.67166957236805]
```

```
MSE for Poisson Noise= [0.18319204722067156, 0.17393602021203486, 0.1782841796510229, 0.13495926076568537, 0.13804496359186347,
0.17514674767480995, 0.25898550986552077, 0.14384893364572132, 0.32177592331986576, 0.12967348188694242, 0.24032525000949057,
0.47710290869564254, 2.469162984865544, 8.774991484377372, 394.67166957236805]
```

ج)

```
Bias for Poisson Noise= [0.17575820476551307, 0.16492859085180647, 0.16488013871625234, 0.12398592155537279, 0.122935077355418
7, 0.14404270545740358, 0.16706922131305543, 0.0867907255702771, 0.10232230462839947, 0.08379980913484375, 0.10225727478286695,
0.029779360947962966, 0.13060785473515305, 0.033213279421375135, 0.6124111532811658]
```

```
Variance for Poisson Noise= [0.007433842455158635, 0.009007429360228317, 0.013404040934770345, 0.010973339210312473, 0.01510988
6236444851, 0.031104042217406397, 0.09191628855246546, 0.05705820807544416, 0.2194536186914664, 0.04587367275209861, 0.13806797
522662365, 0.44732354774767963, 2.338555130130392, 8.741778204955995, 394.0592584190869]
```

```
Bias for White Noise= [0.17575820476551307, 0.16492859085180647, 0.16488013871625234, 0.12398592155537279, 0.1229350773554187,
0.14404270545740358, 0.16706922131305543, 0.0867907255702771, 0.10232230462839947, 0.08379980913484375, 0.10225727478286695, 0.
029779360947962966, 0.13060785473515305, 0.033213279421375135, 0.6124111532811658]
```

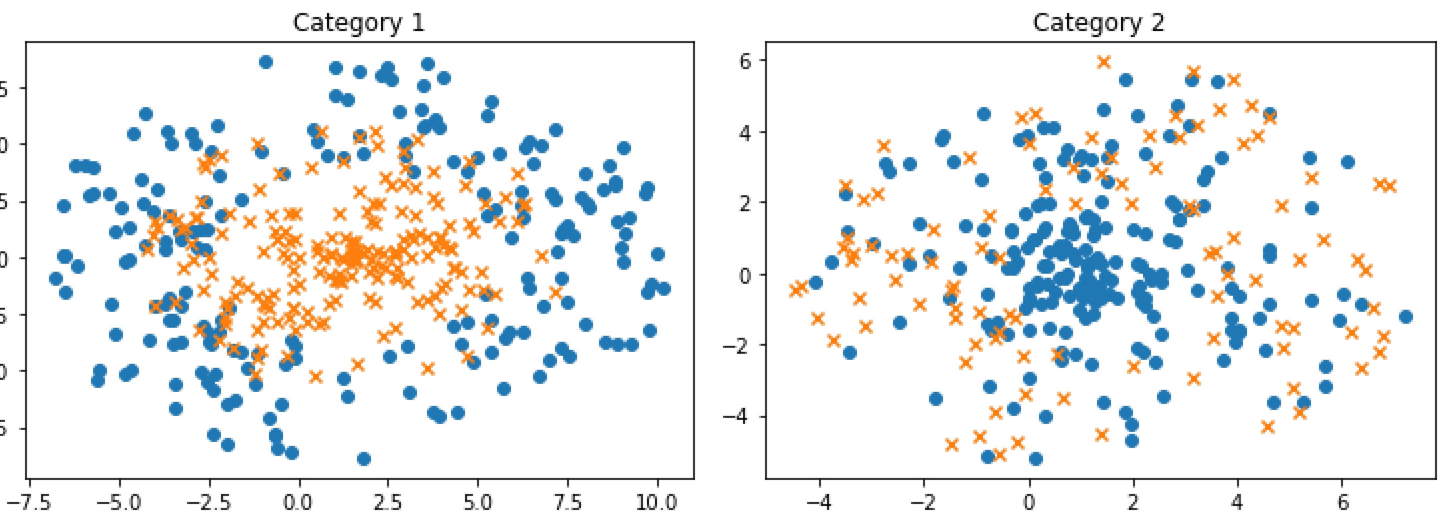
```
Variance for White Noise= [0.007433842455158635, 0.009007429360228317, 0.013404040934770345, 0.010973339210312473, 0.0151098862
36444851, 0.031104042217406397, 0.09191628855246546, 0.05705820807544416, 0.2194536186914664, 0.04587367275209861, 0.1380679752
2662365, 0.44732354774767963, 2.338555130130392, 8.741778204955995, 394.0592584190869]
```

مقادیر بایاس و واریانس به شرح بالا است و همچنین نمودار رشد و نمو آنها در قسمت الف ذکر شد.

براساس Variance & Bias Trade-off در درجات کم با سادگی مدل واریانس کم ولی بایاس زیاد است اما هرچه مدل پیچیده تر می شود واریانس افزایش و بایاس کاهش می یابد.

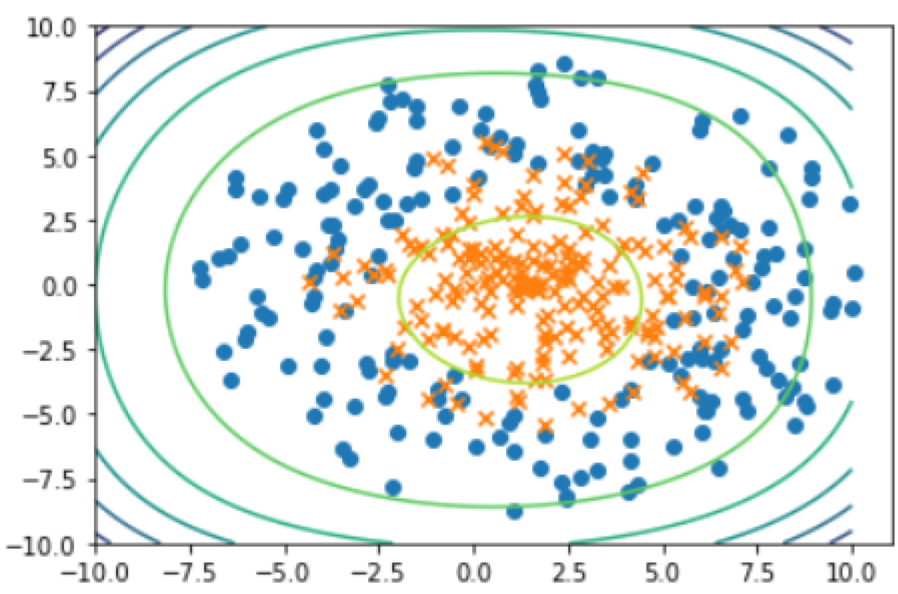
سوال 7

الف)



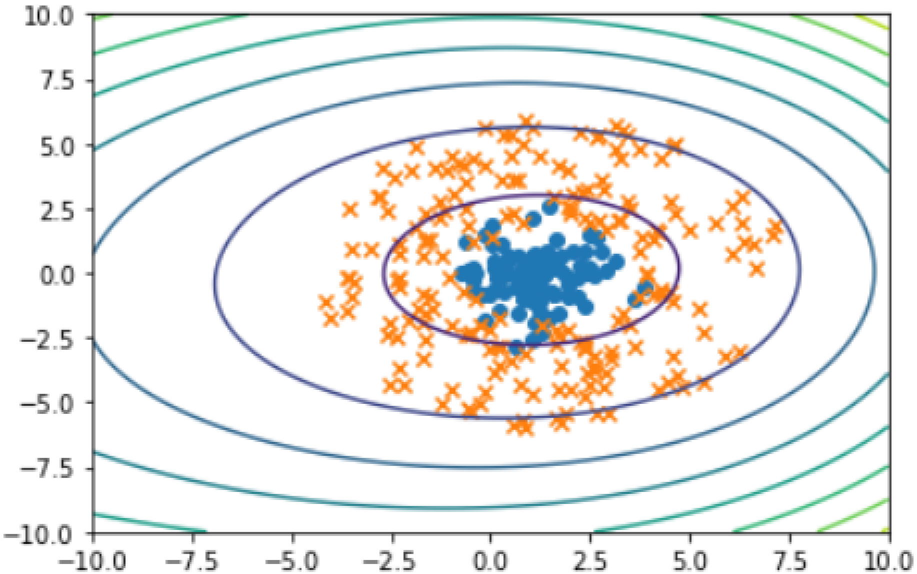
ب)

برای دیتاست اول:



```
model with degree 7 has accuracy: 0.81
```

برای دیتاست دوم:




```
model with degree 4 has accuracy: 0.86
```

ج)

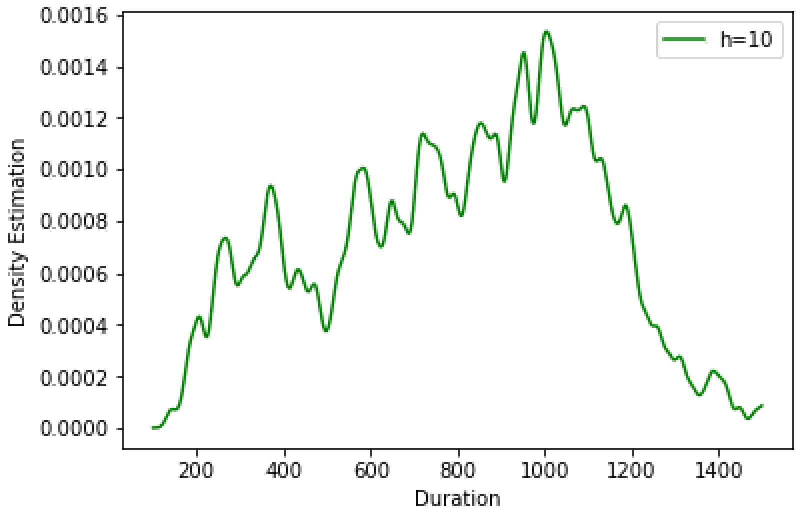
با توجه به فضای دو بعدی فیچر ها و غیرخطی بودن آنها باید فضای ویژگی را به بعد های بالاتر تغیر دهیم ولی البته به دلیل وجود L2 Penalty از یک درجه ای به بعد تقریبا افزایش بعد تاثیری نخواهد داشت!

در درجات پایین تر مدل دچار underfit و در درجات بالاتر مدل دچار overfit می شود!

سوال 8

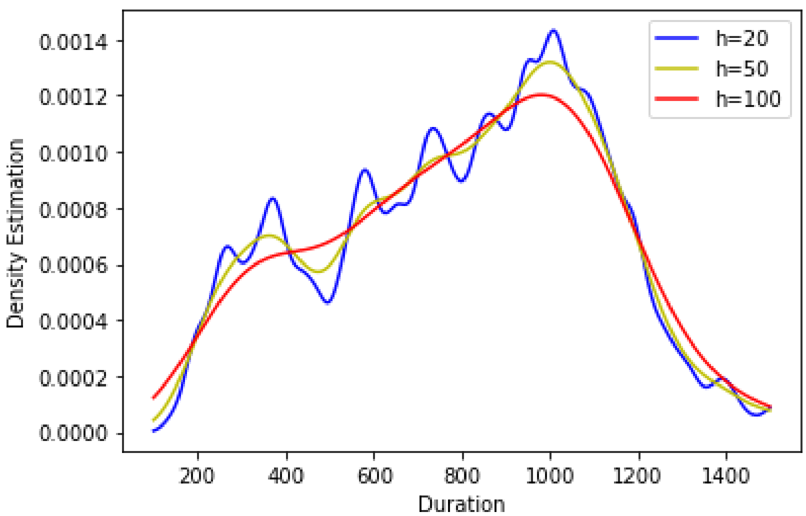
الف)

در این روش با فیکس بودن اندازه پنجره (h=10) براساس نحوه توزیع گوسی سمپل ها اطراف پنجره، توزیع کلی را به دست می آورد.

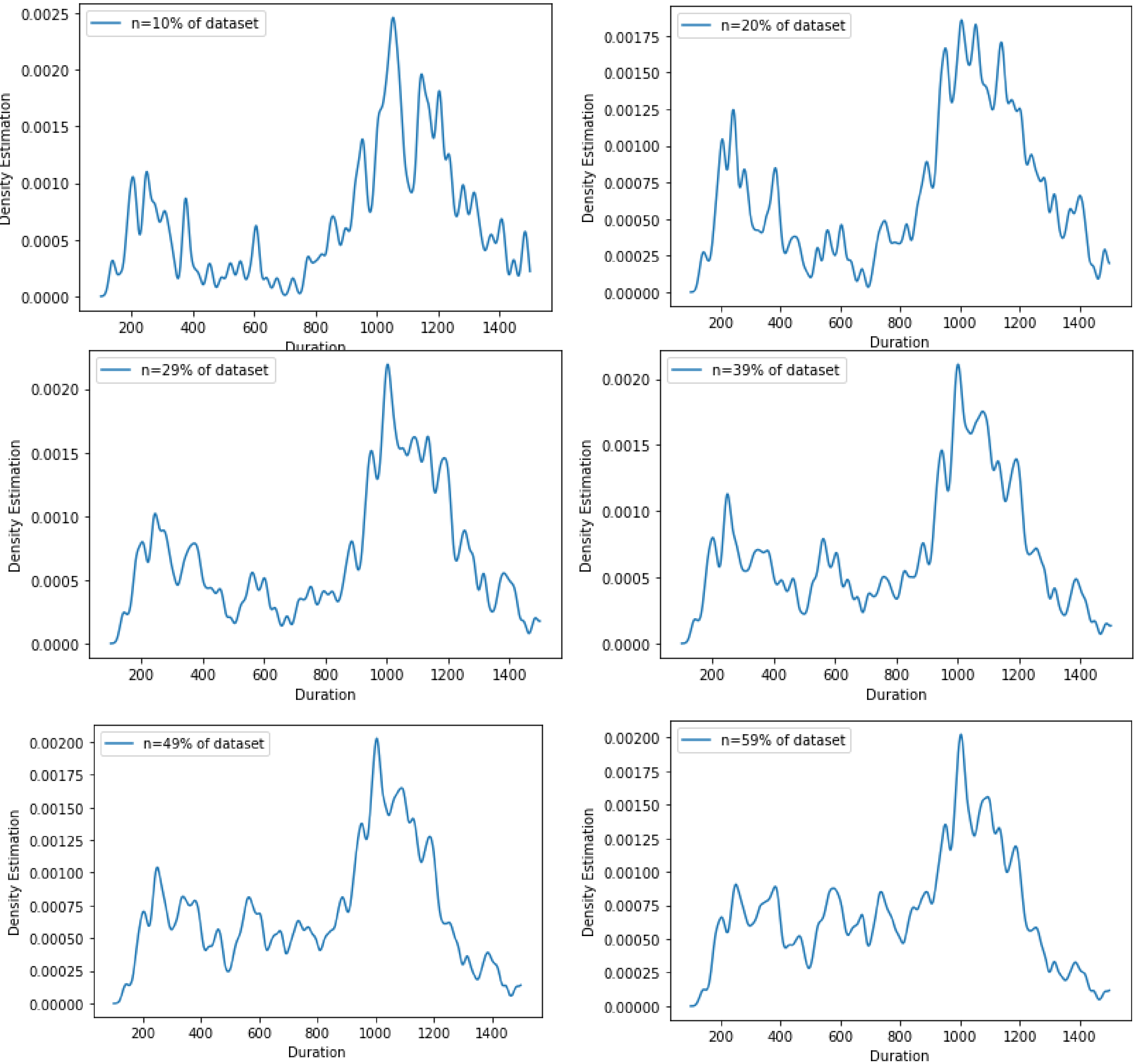


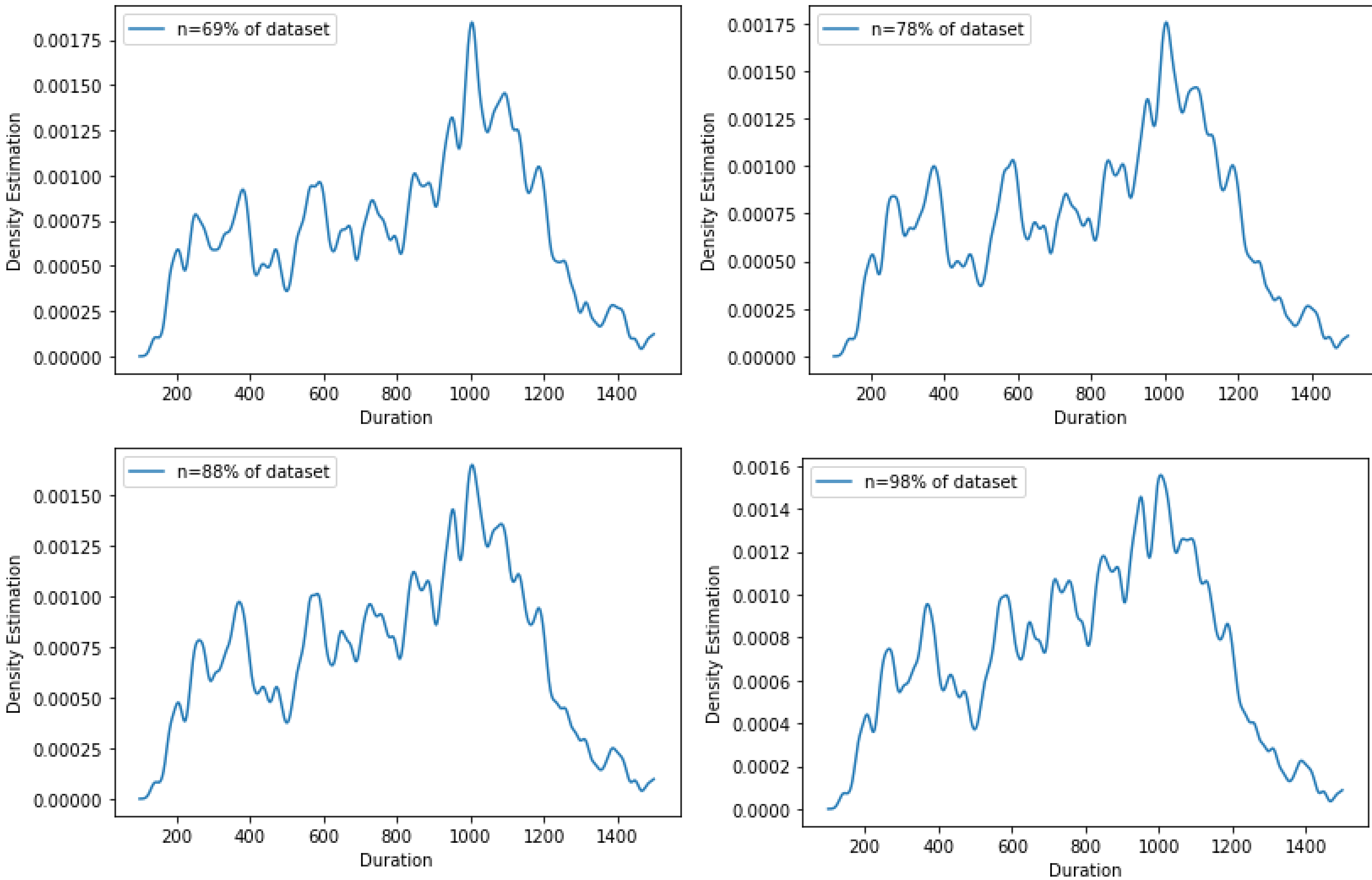
ب)

اصطلاحا اندازه پنجره، smoothing factor است یعنی با افزایش h نحوه توزیع نرم تر میشود.



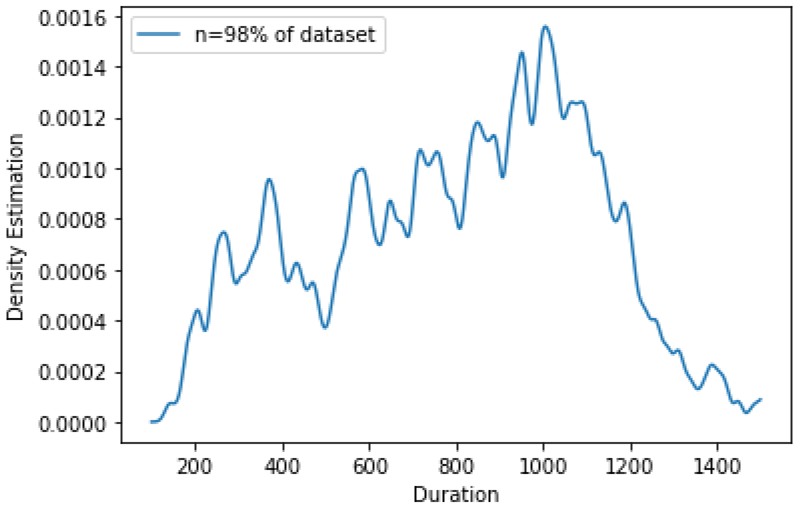
ج)

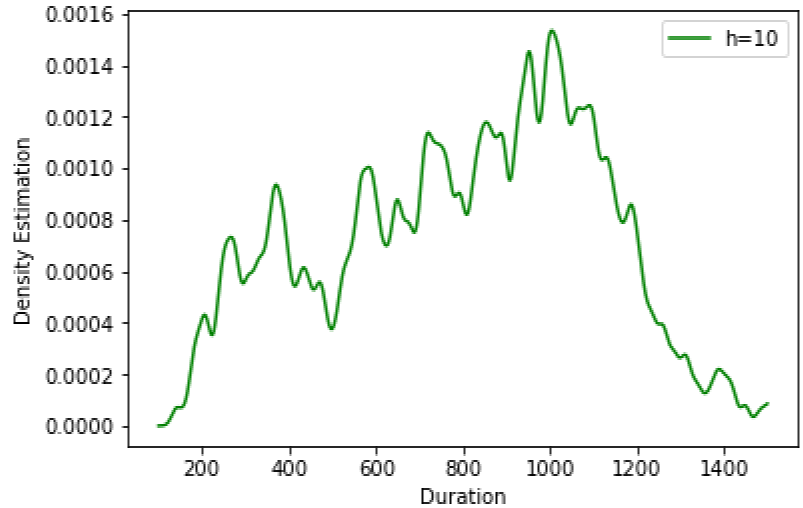

n=10% of dataset
n=20% of dataset
n=29% of dataset
n=39% of dataset
n=49% of dataset
n=59% of dataset
Density Estimation
Duration



با افزایش n توزیع نرم تر شده تا در نهایت به مدل واقعی همگرا میشود.

د)





نمودار سمت راست حاصل از محاسبه دستی و نمودار سمت چپ با استفاده از کتابخانه KernelDensity از sklearn.neighbors رسم شده است.

همانطور که مشاهده می کنید به ازای طول پنجره یکسان روش دستی همانند مدل آماده است.