دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر

گزارش تکلیف چهارم درس شناسائی آماری الگو

دانشجو:

سید احمد نقوی نوزاد

ش-د: ۹۴۱۳۱۰۶۰

استاد:

دکتر رحمتی

پائیز ۹۵

جواب سوال ۱

کدهای مربوط این سوال در فایل ex01 قرار دارد.

قسمت الف)

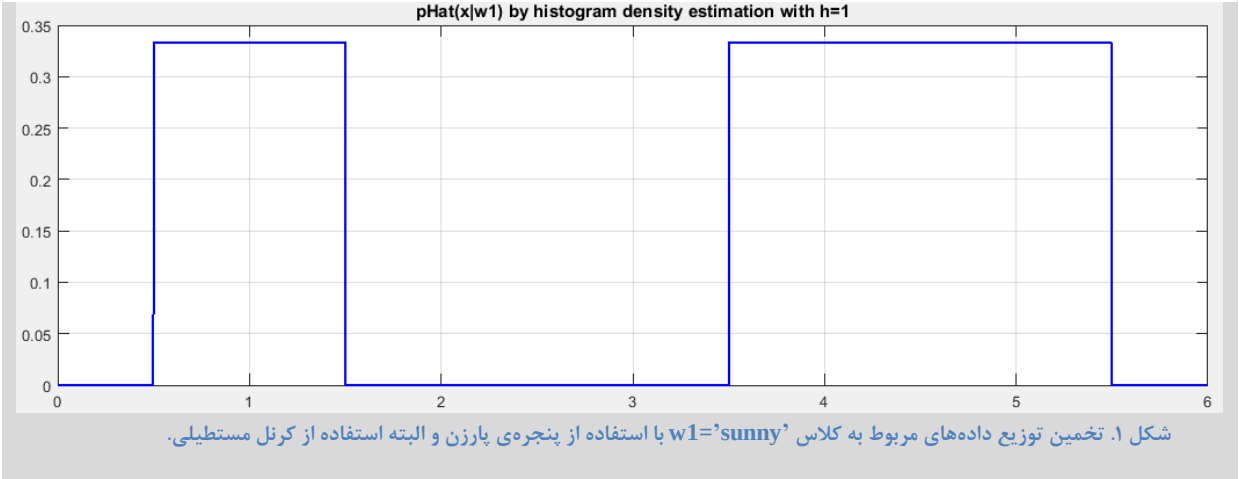
با توجه به صورت سوال، مجموعه‌داده‌ی مورد استفاده به صورت زیر می‌باشد:

light level	label (w1='sunny' , w2='cloudy')
1	sunny
2	cloudy
3	cloudy
4	sunny
5	sunny

معادلات پنجره‌ی پارزن[1] نیز به صورتی که در ادامه می‌آید می‌باشند:

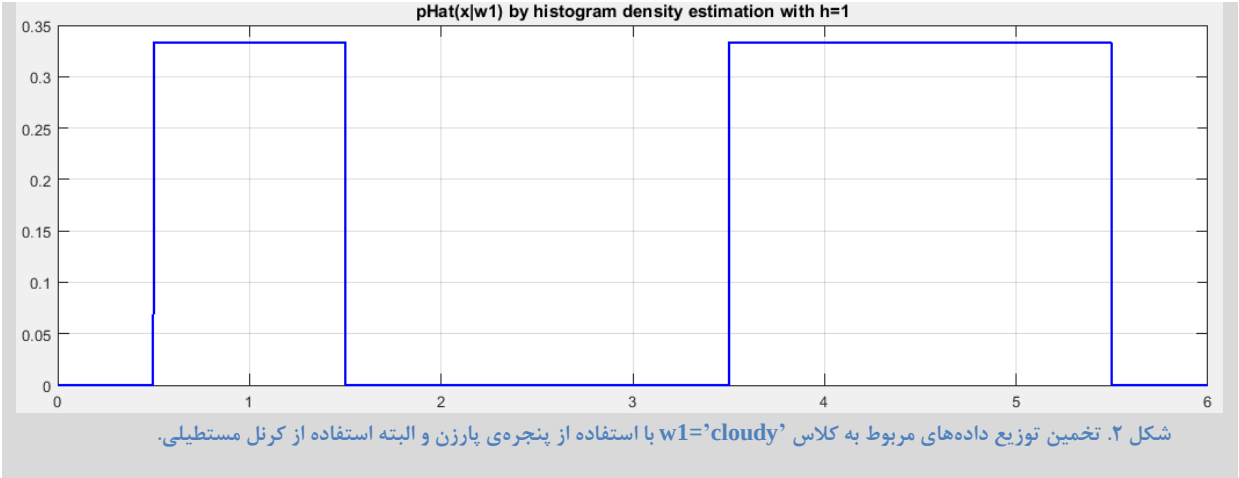
$$p(x)=\frac{1}{nV}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h}\right);\quad \varphi\left(\frac{x}{h}\right)=\begin{cases}1 & |x|<\frac{h}{2}\\ 0 & \text{otherwise}\end{cases}$$

قسمت الف)



شکل ۱. تخمین توزیع داده‌های مربوط به کلاس w1='sunny' با استفاده از پنجره‌ی پارزن و البته استفاده از کرنل مستطیلی.

قسمت ب)



شکل ۲. تخمین توزیع داده‌های مربوط به کلاس w1='cloudy' با استفاده از پنجره‌ی پارزن و البته استفاده از کرنل مستطیلی.

[1] Parzen Window

جواب سوال ۲

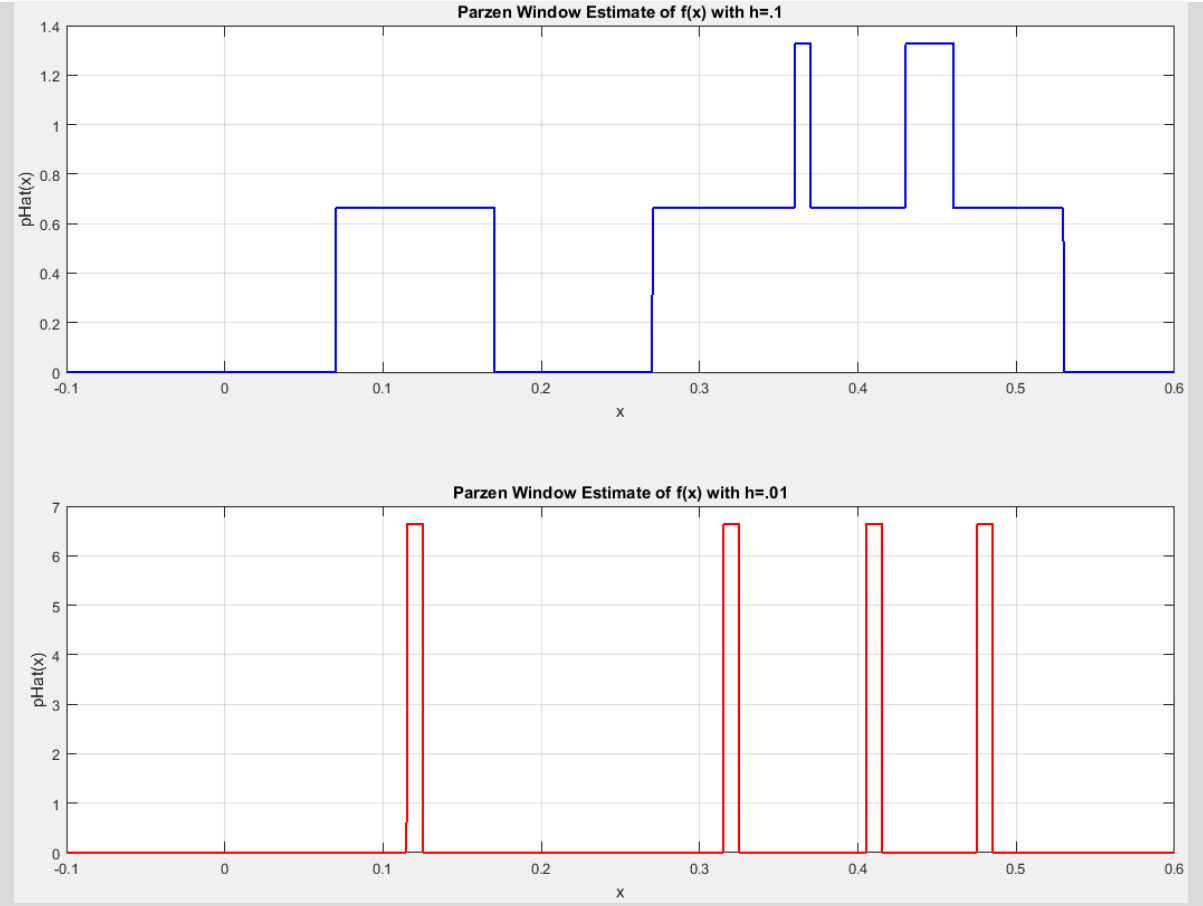
کدهای مربوط این سوال در فایل‌های ex02_a تا ex02_b قرار دارند.

$$X = \{.01, .12, .19, .32, .41, .48\}$$

قسمت الف)

Kernel Type: Gaussian $\rightarrow$ $\varphi(X) \sim N(0,1)$

Find and sketch the Parzen window estimate of f(x) for values of $h_n = 0.1$ and $h_n = 0.01$.

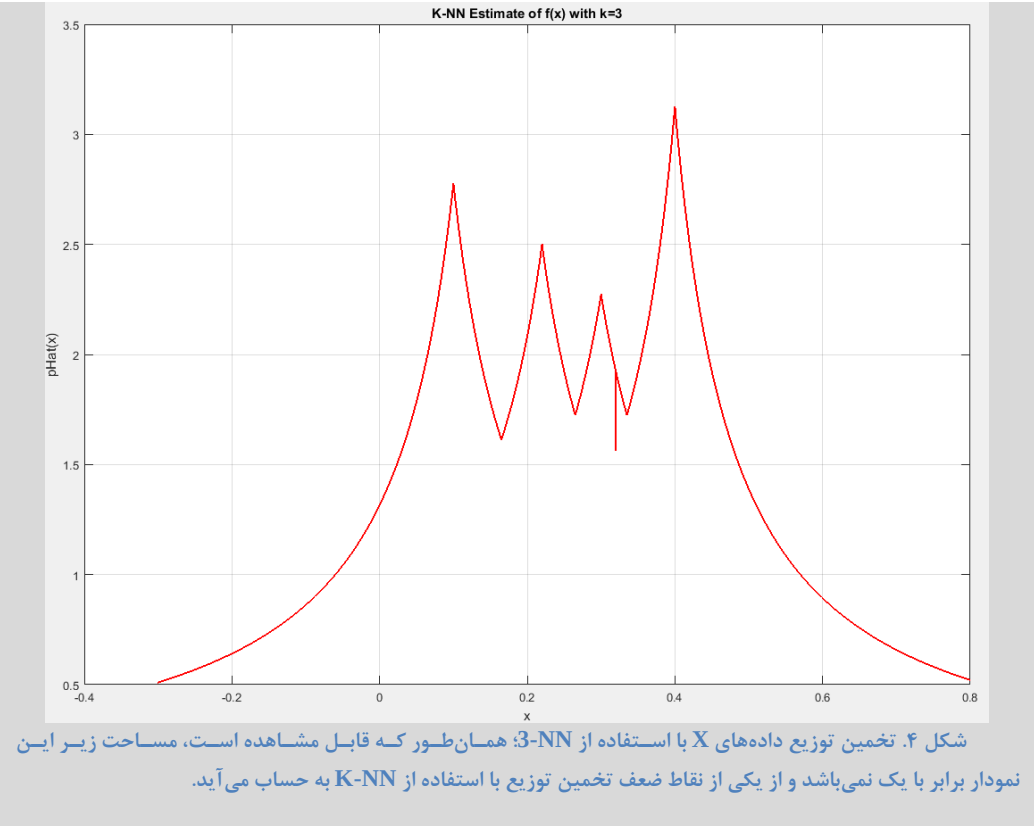


شکل ۳. تخمین توزیع داده‌های X با استفاده از مقادیر متفاوت برای عرض پنجره برابر با ۰٫۱ (تصویر بالا) و ۰٫۰۱ (تصویر پایین)؛ همان‌طور که قابل مشاهده است، به ازای مقدار کمتر برای عرض پنجره شاهد تخمین ناهموارتری از پنجره‌ی پارزن می‌باشیم، که علت این مسئله نیز کم‌بودن تعداد داده‌های آموزشی می‌باشد.

قسمت ب)

شکل زیر تخمین تابع f(x) را با استفاده از ۳ نزدیک‌ترین همسایه[۲] یا همان 3-NN نشان می‌دهد:

[2] 3-nearest neighbours



شکل ۴. تخمین توزیع داده‌های X با استفاده از 3-NN؛ همان‌طور که قابل مشاهده است، مساحت زیر این نمودار برابر با یک نمی‌باشد و از یکی از نقاط ضعف تخمین توزیع با استفاده از K-NN به حساب می‌آید.

جواب سوال ۳

در این مسئله، با توجه به این‌که داده‌های مسئله دوبُعدی بوده و فاصله‌ی مدنظر نیز از نوع فاصله‌ی اقلیدسی می‌باشد، قصد داریم تا برچسب داده‌ها را با با استفاده از مقادیر آموزشی موجود پیش‌بینی نمائیم. داریم:

قسمت الف)

در این قسمت، قصد داریم تا leave-one-out cross-validation error را برای 1-NN محاسبه نمائیم. با توجه به تصویر بالا مشاهده می‌نمائیم که در مورد داده‌های سمت راست با استفاده از رویه‌ی loo cv دچار اشتباه نمی‌شویم، چرا که اولین نزدیک‌ترین همسایه‌ی هر داده با خود داده هم‌کلاس می‌باشند. اما در مورد داده‌های سمت چپ، مشاهده می‌کنیم که نزدیک‌ترین همسایه‌ی هر داده از کلاس مخالف بوده و در نتیجه در مورد همگی آن‌ها دچار خطا خواهیم شد. در نتیجه میزان خطا برای این قسمت برابر $\frac{5}{10}$ می‌باشد.

قسمت ب)

در این قسمت نیز مانند قبل به قضیه می‌نگریم و مشاهده می‌کنیم که در مورد داده‌های سمت راست و حتی به ازای *3-NN* باز هم دچار خطا نمی‌شویم. اما در مورد داده‌های سمت چپ، به ازای داده‌های کلاس مثبت، شاهدیم که از ۳ نزدیک‌ترین همسایه‌ی هر کدام از آن‌ها، دو مورد هم‌کلاس و یکی از کلاس مخالف می‌باشد. در نتیجه کلاس نسبت‌داده‌شده به هر یک صحیح خواهد بود. اما در مورد تنها داده‌ی از کلاس منفیِ سمت چپ، هر ۳ نزدیک‌ترین همسایه‌ی آن از کلاس مخالف بوده و در نتیجه برچسب اشتباه به آن داده نسبت داده خواهد شد. بنابراین میزان خطا برای این قسمت برابر $\frac{1}{10}$ می‌باشد.

قسمت ج)

جهت تعیین مقدار بهینه‌ی K در رویه‌ی دسته‌بندی K-NN، شاید بهترین راه، استفاده از همین رویه‌ی Leave-One-Error Cross-Validation باشد. چرا که به این صورت، در واقع داریم رویه‌ی آزمایش را شبیه‌سازی می‌نمائیم. بدین معنی که هر بار یک داده را به عنوان داده‌ی آزمایشی کنار گذاشته و بررسی می‌نمائیم که به ازای چه مقادیری از K برچسب صحیحی به داده‌ی مورد نظر نسبت داده خواهد شد. حال به ازای همه‌ی داده‌ها تعدادی مقادیر K داریم. در ادامه برای تعیین مقدار بهینه‌ی K، می‌بایست بررسی نمائیم که به ازای کدام مقدار K در کل مجموعه‌داده کمترین میزان خطا را داشته‌ایم. آن مقدار K، مقدار بهینه برای استفاده در دسته‌بند K-NN و به ازای مجموعه‌داده‌ی فعلی می‌باشد.

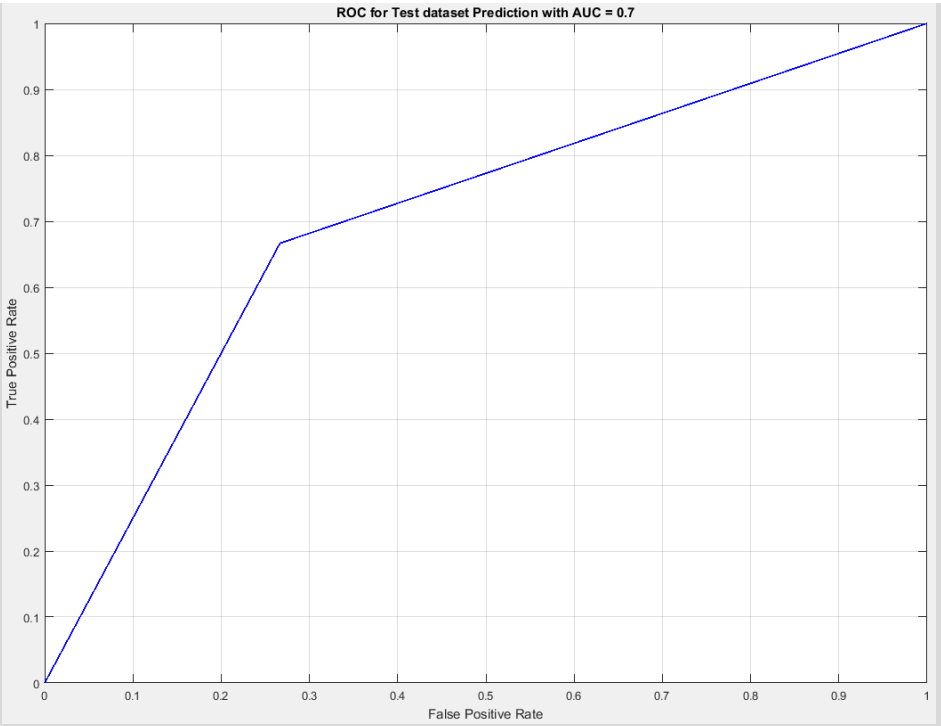
جواب سوال ۴

کدهای مربوط به این سوال در فایل ex04 قرار دارد.

در این قسمت، نتایج حاصل از اجرای کد متلب مربوط به دسته‌بند K-NN برای دو قسمت اول سؤال به صورت زیر می‌باشند:

```
The best value for K is:        5
The error rate for test dataset with the best value of K = 5, gained by Leave-One-Out CV is:        29.63 %
```

در مورد قسمت سوم سؤال نیز به جای اعلان برچسب‌های پیش‌بینی‌شده به ازای داده‌های آزمایشی، منحنی ROC مربوط به کیفیت دسته‌بندی داده‌های تست را با استفاده از پارامترهای بهینه‌ی حاصله از دو قسمت قبل نشان می‌دهیم. داریم:



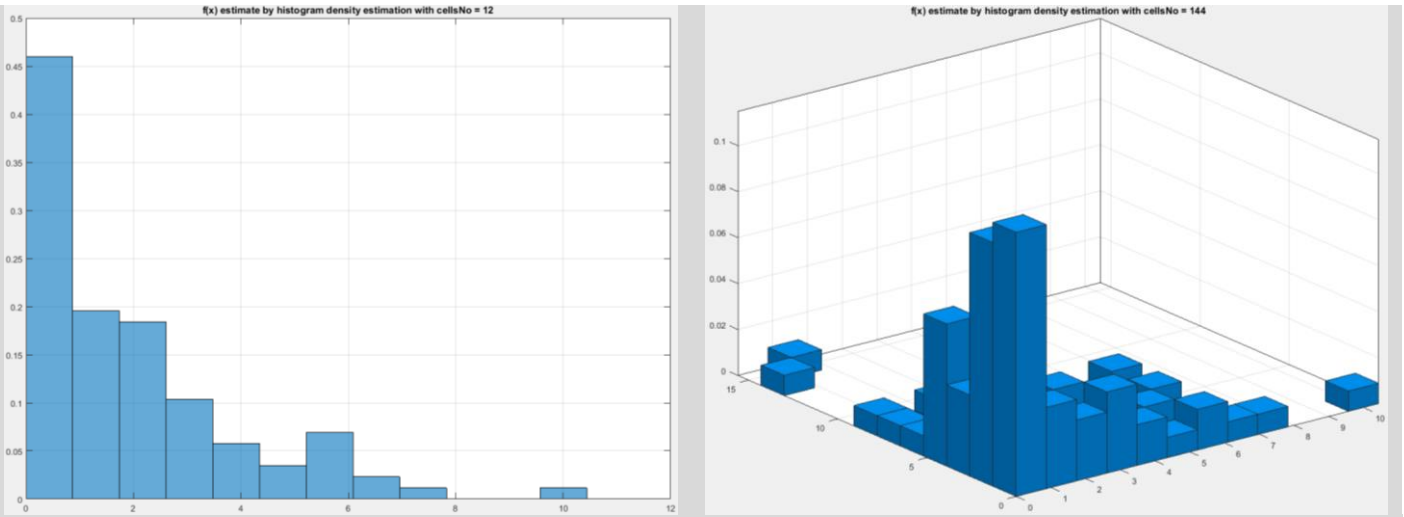
شکل ۵. منحنی ROC مربوط به کیفیت دسته‌بندی داده‌های تست با استفاده از مقدار بهینه‌ی K=5 حاصله از قسمت قبل.

جواب سوال ۵

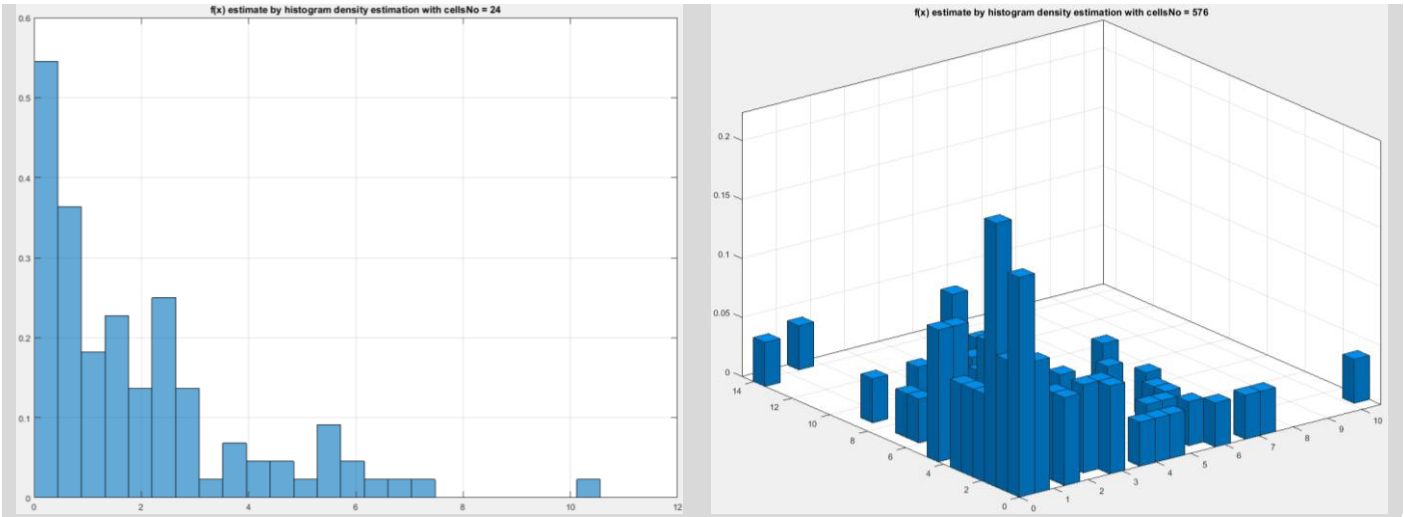
کدهای مربوط به این سوال در فایل‌های ex05، histEst، KNN-Est و parzenWindEst قرار دارند.

در این‌جا توزیع نمائی[3] را به عنوان توزیع دلبخواهی تک‌بُعدی و دوبُعدی برمی‌گزینیم (لازم به ذکر است که نمونه‌دادهای توزیع مربوطه را با استفاده از ابزار randtool متلب تولید نمودیم). نتایج حاصله به ازای مقادیر مختلف برای پارامترهای مسئله در ادامه می‌آید. داریم:

نتایج مربوط به تخمین هیستوگرام:



شکل ۶. سمت چپ: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین هیستوگرام و تعداد سلول‌ها برابر با ۱۲؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی دوبُعدی با استفاده تخمین هیستوگرام و تعداد سلول‌ها در هر بُعد برابر با ۱۲.

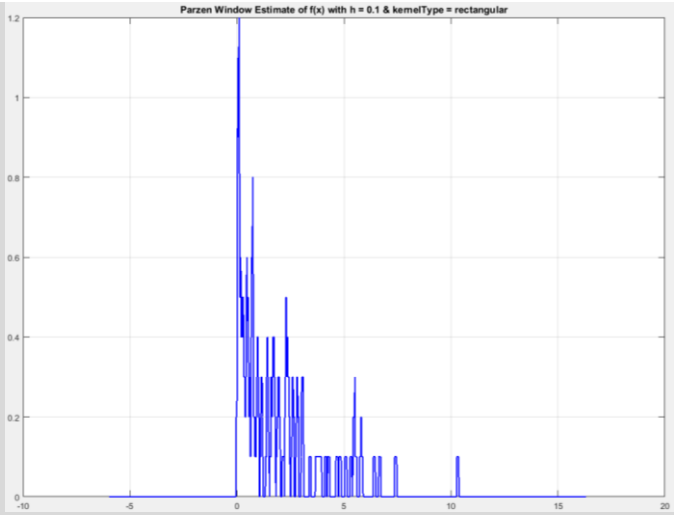


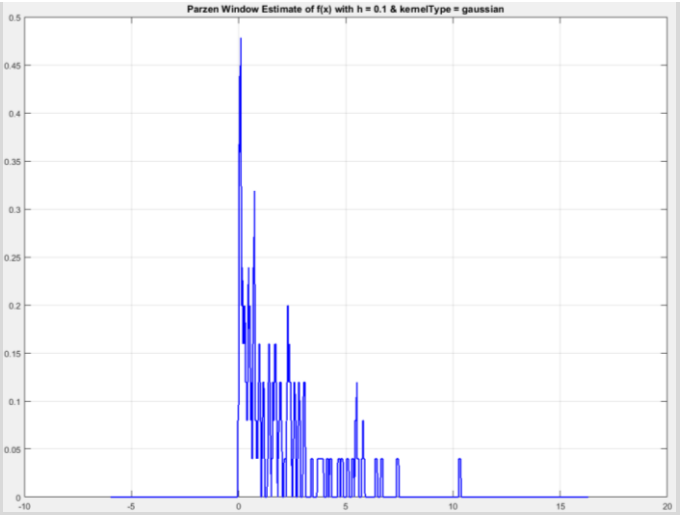
شکل ۷. سمت چپ: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین هیستوگرام و تعداد سلول‌ها برابر با ۲۴؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی دوبُعدی با استفاده تخمین هیستوگرام و تعداد سلول‌ها در هر بُعد برابر با ۲۴.

همان‌طور که از نتایج مربوطه به تخمین هیستوگرام پیداست، به ازای افزایش تعداد سلول‌ها در هر بعد شاهد تخمین هموارتر و با جزئیات بیشتری خواهیم بود.

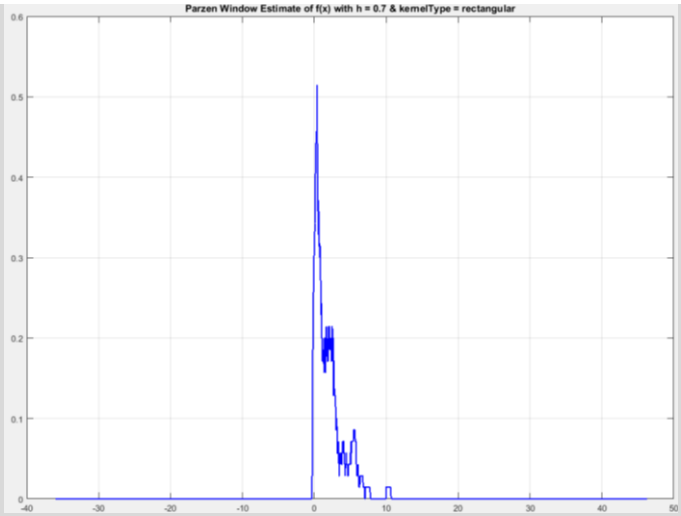
نتایج مربوط به تخمین پنجره‌ی پارزن:

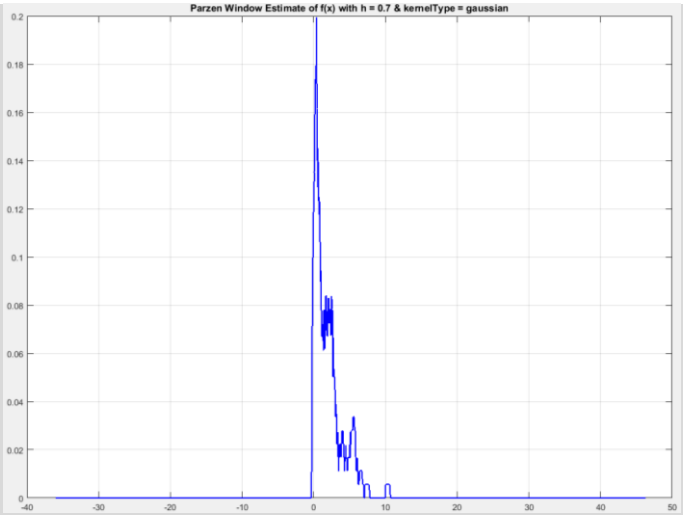
[3] Exponential Distribution





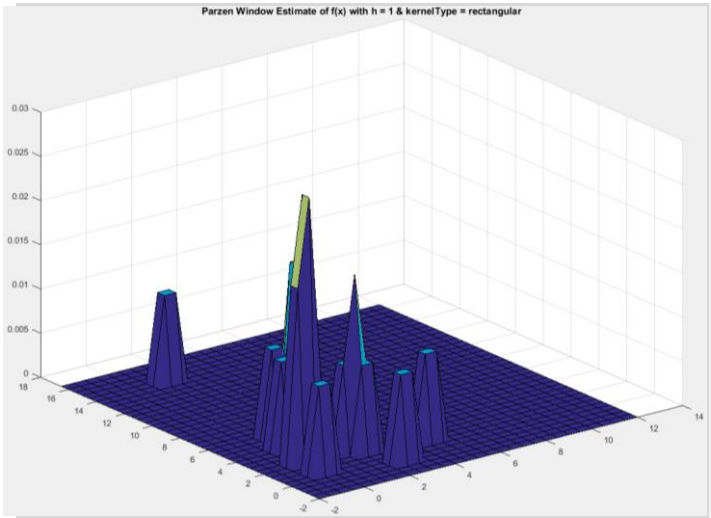
شکل ۸. سمت چپ: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۰٫۱ و تابع کرنل مستطیلی؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۰٫۱ و تابع کرنل گاوسین.

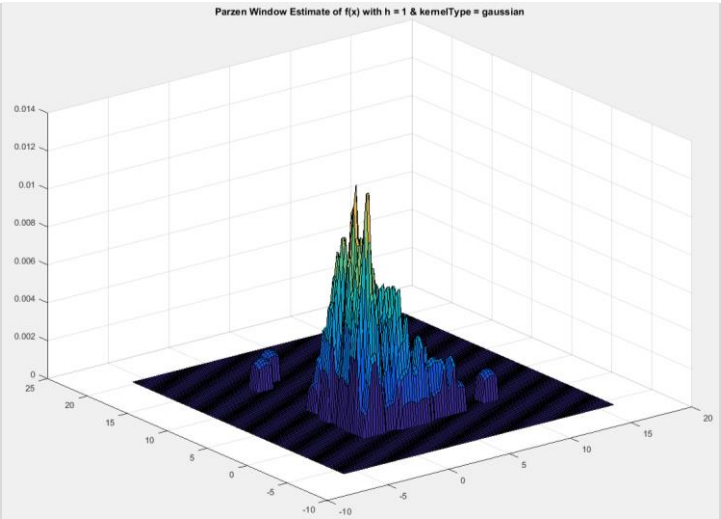




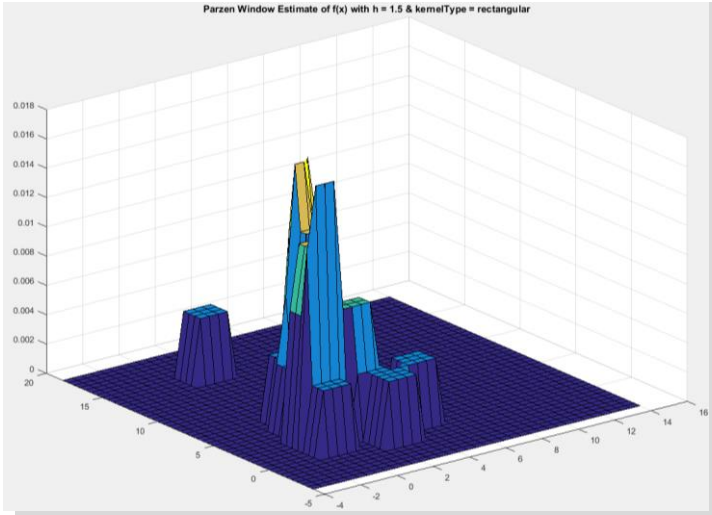
شکل ۹. سمت چپ: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۰٫۷ و تابع کرنل مستطیلی؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۰٫۷ و تابع کرنل گاوسین.

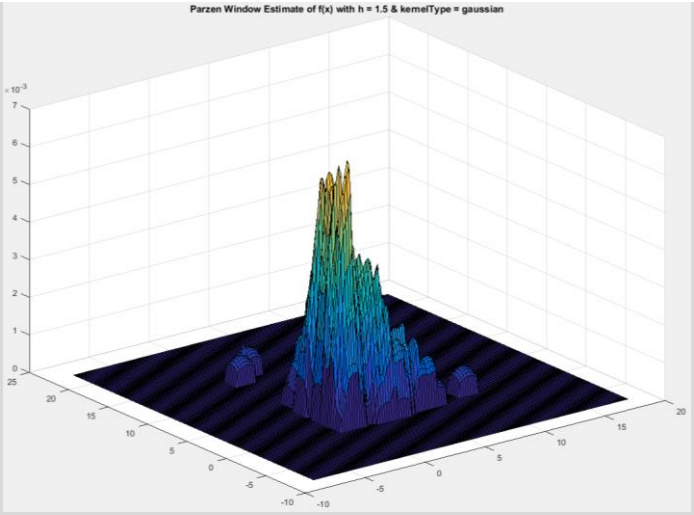
همان‌طور که از نتایج مربوطه به تخمین پنجره‌ی پارزن برای داده‌های یک‌بعدی پیداست، به ازای افزایش عرض پنجره شاهد تخمین هموارتری از توزیع نمائی خواهیم بود.





شکل ۱۰. سمت چپ: تخمین توزیع داده‌های توزیع نمائی دوبعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۱ و تابع کرنل مستطیلی؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی دوبعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۱ و تابع کرنل گاوسین.

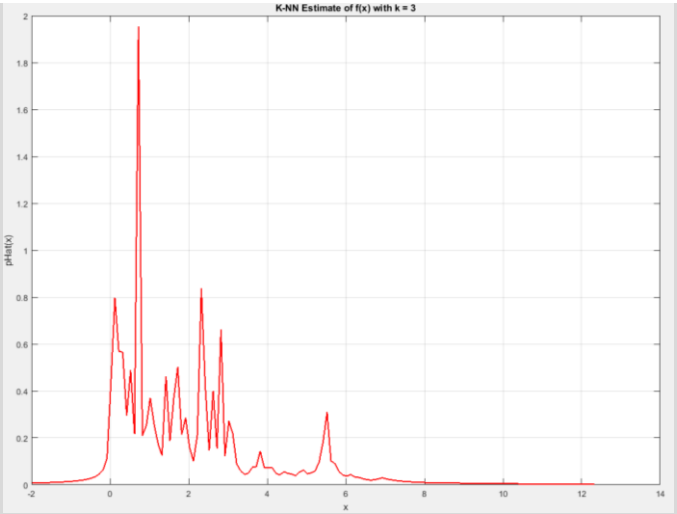


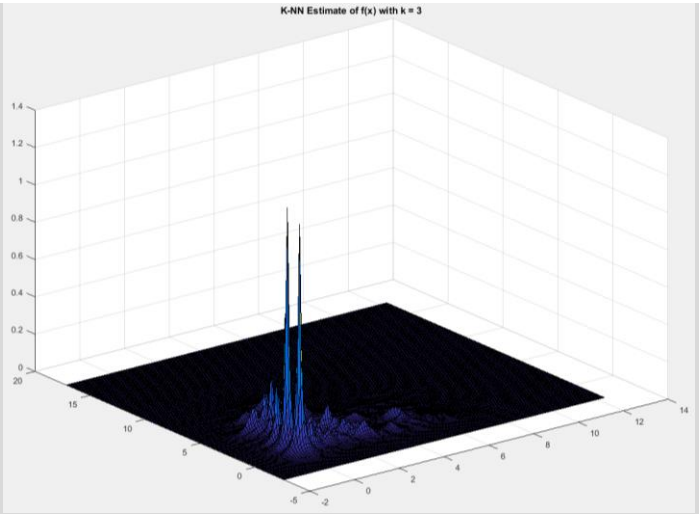


شکل ۱۱. سمت چپ: تخمین توزیع داده‌های توزیع نمائی دوبعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۱.۵ و تابع کرنل مستطیلی؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی دوبعدی با استفاده تخمین پنجره‌ی پارزن و مقدار عرض پنجره برابر ۱.۵ و تابع کرنل گاوسین.

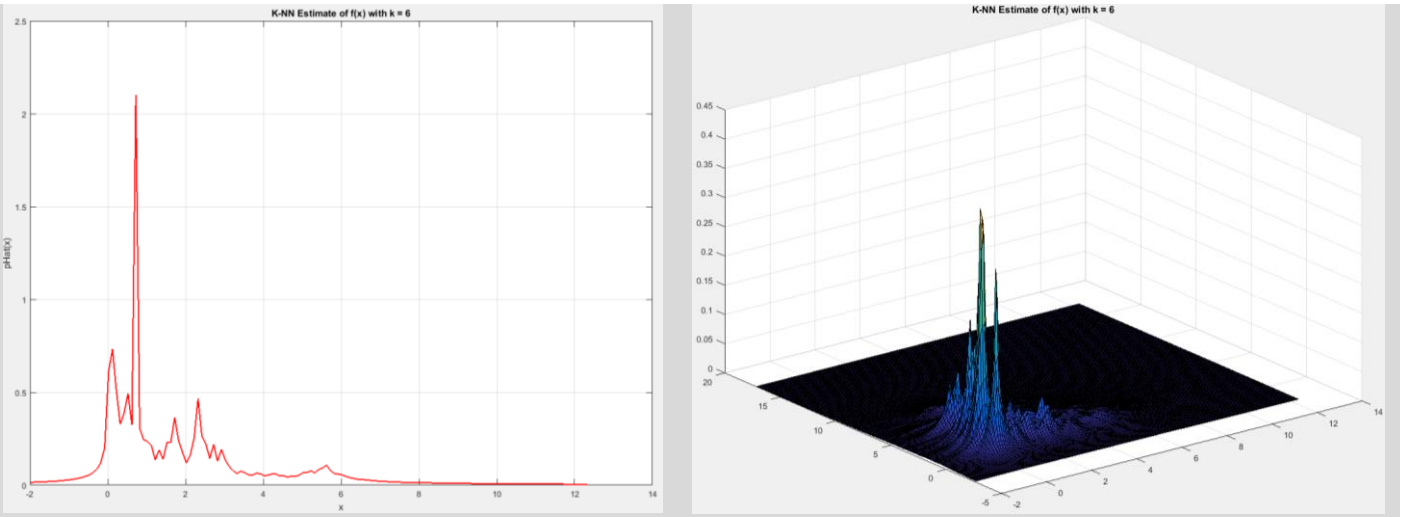
همان‌طور که از نتایج مربوطه به تخمین پنجره‌ی پارزن برای داده‌های دوبعدی پیداست، مانند داده‌های یک‌بعدی به ازای افزایش عرض پنجره شاهد تخمین هموارتری از توزیع نمائی خواهیم بود.

نتایج مربوط به تخمین K-NN:





شکل ۱۲. سمت چپ: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین K-NN و مقدار K برابر ۳ ؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی دوبعدی با استفاده تخمین K-NN و مقدار K برابر ۳.



شکل ۱۳. سمت چپ: تخمین توزیع داده‌های توزیع نمائی یک‌بعدی با استفاده تخمین K-NN و مقدار K برابر ۶؛ سمت راست: تخمین توزیع داده‌های توزیع نمائی دوبعدی با استفاده تخمین K-NN و مقدار K برابر ۶.

همان‌طور که از نتایج مربوطه به تخمین K-NN هم برای داده‌های یک‌بعدی و هم دوبعدی پیداست، به ازای افزایش مقدار پارامتر همسایگی K شاهد تخمین هموارتر و دقیق‌تری از توزیع نمائی خواهیم بود و البته می‌توان از تصاویر نیز مشاهده نمود که با افزایش مقدار K مساحت زیر نمودار نیز کمتر شده و به مقدار مطلوب ۱ نزدیک‌تر می‌گردد.

در پایان باید خاطرنشان کرد که در مورد تمامی تخمین‌گرهای مورد بررسی، با افزایش تعداد نمونه‌های تصادفی موجود شاهد تخمین دقیق‌تری بوده و البته تخمین‌های فعلی نیز تا حد زیادی به توزیع واقعی شباهت دارند، به طوری که تنها با یک نگاه گذرا می‌توان به نوع توزیع مدنظر پی برد، که در این‌جا یک توزیع نمائی می‌باشد.