1_ $\mathrm{Var}(\hat{P}_n(x)) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h}\right)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h}\right)\right) \overset{\bigstar,\ \text{iid } \{x_i\}}{=\!=\!=} \frac{1}{n}\mathrm{Var}\left(\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h}\right)\right) = \frac{1}{n}E\left[\frac{1}{V_n^2}\varphi^2\left(\frac{x-x_i}{h}\right)\right] - \frac{1}{n}\underbrace{E^2\left[\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h}\right)\right]}_{E^2[\hat{P}(x)]}$

$\frac{1}{V_n}\varphi(\square)\times\frac{1}{V_n}\varphi(\square)$

★ مستقل X_2, X_1 : $\mathrm{Var}(X_1+X_2) = \mathrm{Var}(X_1)+\mathrm{Var}(X_2)$

$$= \frac{1}{n}\int\frac{1}{V_n^2}\varphi^2\left(\frac{x-x_i}{h}\right)P(x_i)\,dx_i \underbrace{- \frac{1}{n}E^2[\hat{P}(x)]}_{\leq 0}$$

$$\leq \frac{1}{n}\frac{\mathrm{Sup}(\varphi)}{V_n}\underbrace{\int\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h}\right)P(x_i)\,dx_i}_{E[\hat{P}(x)]}$$

$$\boxed{\longrightarrow \mathrm{Var}[\hat{P}(x)] \leq \frac{\mathrm{Sup}(\varphi)\,E[\hat{P}(x)]}{nV_n}}$$

2_

1. $\varphi(x) = \begin{cases} e^{-x} & x>0 \\ 0 & x\leq 0\end{cases}$, $V_n = \int\varphi\left(\frac{x}{h_n}\right)dx = \int_0^\infty e^{-\frac{x}{h_n}}dx = -h_n e^{-\frac{x}{h_n}}\Big|_0^\infty = h_n$

$$P_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h_n}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x-x_i}{h_n}\right)$$

$$\longrightarrow \bar{P}_n(x) = E[P_n(x)] = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x-x_i}{h_n}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}E\left[\frac{1}{h_n}\varphi\left(\frac{x-x_i}{h_n}\right)\right]$$

$$\xrightarrow{\text{i.i.d. } \{x_i\}} = E\left[\frac{1}{h_n}\varphi\left(\frac{x-x_i}{h_n}\right)\right]$$

$$= \int\frac{1}{h_n}e^{-\frac{x-x_i}{h_n}}P(x_i)\,dx_i$$

$x\leq 0 \longrightarrow \varphi\left(\frac{x-x_i}{h_n}\right) = 0$; $1\leq i\leq n \longrightarrow E[P_n(x)] = 0$

$0<x<a \longrightarrow \varphi\left(\frac{x-x_i}{h_n}\right) = \begin{cases} 0 & x\leq x_i \\ e^{-\frac{x-x_i}{h_n}} & x>x_i\end{cases} \longrightarrow E[P_n(x)] = \int_0^x\frac{1}{h_n}e^{-\frac{x-x_i}{h_n}}\frac{1}{a}\,dx_i = \frac{1}{ah_n}e^{-\frac{x}{h_n}}\left(h_n e^{\frac{x_i}{h}}\Big|_0^x\right) = \frac{1}{a}\left(1-e^{-\frac{x}{h_n}}\right)$

$x\geq a$, $0\leq x_i\leq a$ $\longrightarrow \varphi\left(\frac{x-x_i}{h_n}\right) = e^{-\frac{x-x_i}{h_n}} \longrightarrow E[P_n(x)] = \int_0^a\frac{1}{h_n}e^{-\frac{x-x_i}{h_n}}\frac{1}{a}\,dx_i = \frac{1}{a}e^{-\frac{x}{h_n}}\left(e^{\frac{a}{h_n}}-1\right)$

2. $\mathrm{bias} = P(x) - E[\hat{P}_n(x)]$

$$\longrightarrow \mathrm{bias} = \frac{1}{a} - \frac{1}{a}\left(1-e^{-\frac{x}{h_n}}\right) < 0.01$$

$$\longrightarrow e^{-\frac{x}{h_n}} < \frac{a}{100} \xrightarrow{\ln} -\frac{x}{h_n} < \ln\frac{a}{100}$$

$$\longrightarrow h_n < x\ln\frac{100}{a}$$

3 –

1. $x_i = [x_1, \dots, x_d]$, $A = \begin{bmatrix} — x_1 — \\ \vdots \\ — x_n — \end{bmatrix}_{n \times d} \longrightarrow A\beta = \begin{bmatrix} x_1 \beta \\ \vdots \\ x_n \beta \end{bmatrix}$, $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$

$$\longrightarrow Y - A\beta = \begin{bmatrix} Y_1 - x_1\beta \\ \vdots \\ Y_n - x_n\beta \end{bmatrix} \longrightarrow \|Y - A\beta\|_2^2 = \sum_{i=1}^{n} (Y_i - x_i\beta)^2$$

$$\longrightarrow \hat{\beta} = \arg\min_{\beta} \overbrace{\underbrace{\|Y - A\beta\|_2^2}_{(Y-A\beta)^T(Y-A\beta)} + \lambda \|\beta\|_2^2}^{J(\beta)} =$$

$$= Y^TY - Y^TA\beta - \beta^TA^TY + \beta^TA^TA\beta$$

$$= Y^TY - 2\beta^TA^TY + \beta^TA^TA\beta$$

$$\longrightarrow \nabla_\beta J(\beta) = -2A^TY + 2A^TA\beta + 2\lambda\beta = 0 \longrightarrow (A^TA + \lambda I)\beta = A^TY$$

$$\longrightarrow \hat{\beta} = (A^TA + \lambda I)^{-1} A^TY$$

2. در حالت کلی regularization روشی برای جلوگیری از overfitting می‌باشد. این روش با افزودن یک ترم پنالتی به تابع خطا (Loss function) انجام می‌شود. در L_1 Regularization که به نام LASSO هم شناخته می‌شود، قدر مطلق ضرایب w_i (نرم 1) به عنوان ترم پنالتی به Loss function افزوده می‌شود. در L_2 Regularization که به نام Ridge Regularization هم شناخته می‌شود، نرم 2 ضرایب به تابع خطا افزوده می‌شود.

تفاوت‌ها:
- ترم پنالتی: $\|w\|_1$ و $\|w\|_2$
- استفاده از روش L_1 باعث می‌شود برخی از ضرایب دقیقاً صفر شوند و در نتیجه مدل Sparse می‌شود. اما در L_2 اینگونه نیست.
- روش L_2 به دلیل استفاده از توان دوم ضرایب w_i، حساسیت بیشتری نسبت به داده‌ها دارد و در مقابل L_1 حساسیت کمتری به داده‌های دور افتاده دارد و به اصطلاح robust تر است.

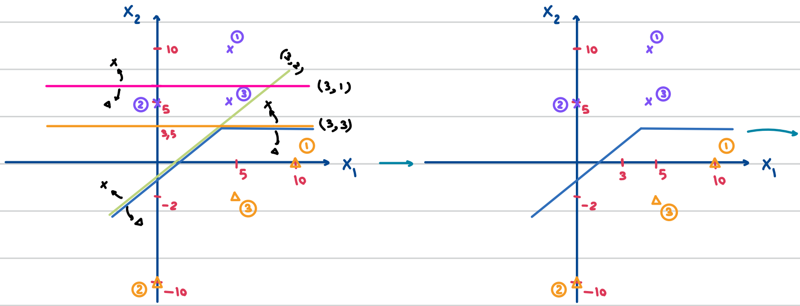
۴-

W1 △		W2 ×		W3 ●	
X1	X2	X1	X2	X1	X2
1 10	0	1 5	10	1 2	8
2 0	-10	2 0	5	2 -5	2
3 5	-2	3 5	5	3 10	-4

$\mu_1 = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$ $\mu_2 = \begin{bmatrix} 3,33 \\ 6,66 \end{bmatrix}$ $\mu_3 = \begin{bmatrix} 2,33 \\ 2 \end{bmatrix}$

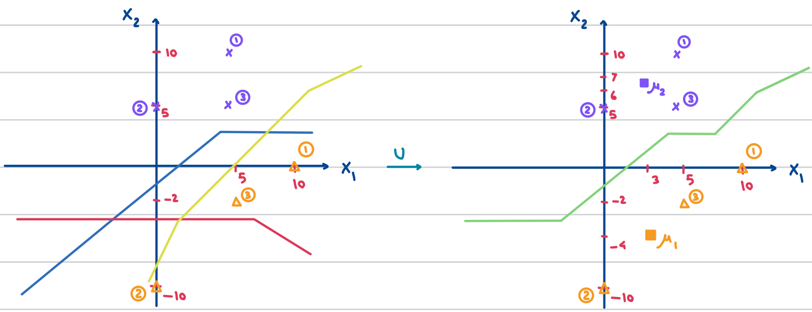
۱. فواصل هر یک از داده‌های کلاس w_1 را با تمام داده‌های w_2 پیدا می‌کنیم و عمودمنصف‌ها را رسم می‌کنیم و سپس با اشتراک گرفتن ناحیه مربوط به هر داده را بدست می‌آوریم. درنهایت با اجتماع گرفتن نواحی، مرز تصمیم دو کلاس بدست می‌آید.

برای داده سوم از w_1 داریم.

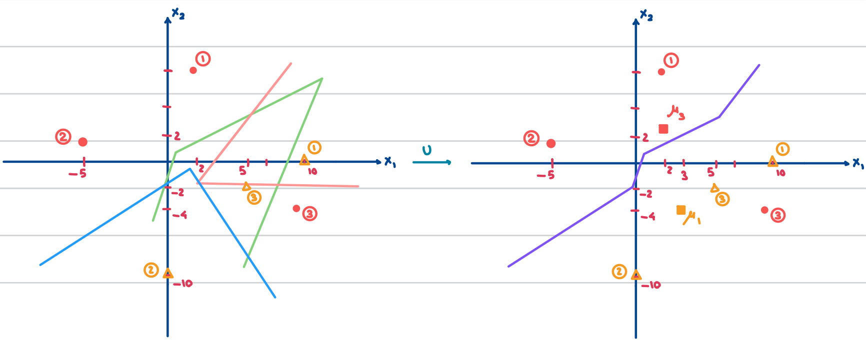


ناحیه مربوط به داده سوم w_1

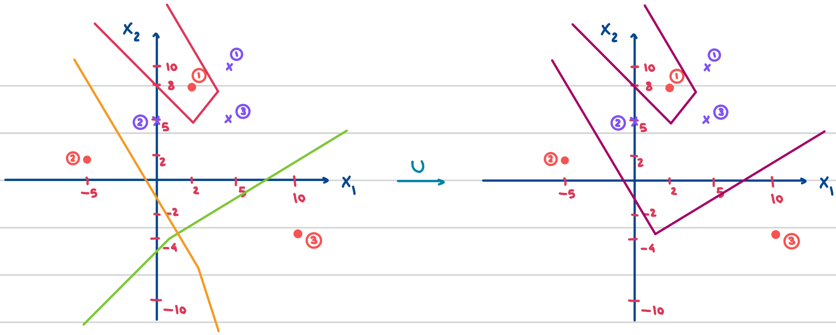
حال نواحی مربوط به داده ۱ و 2 را پیدا می‌کنیم و بین این سه ناحیه اجتماع می‌گیریم:



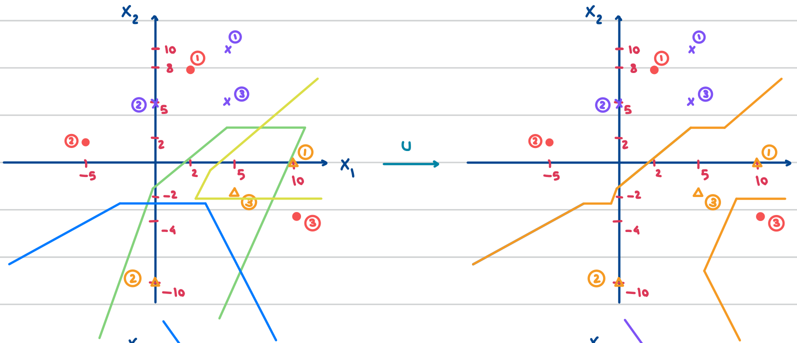
.2



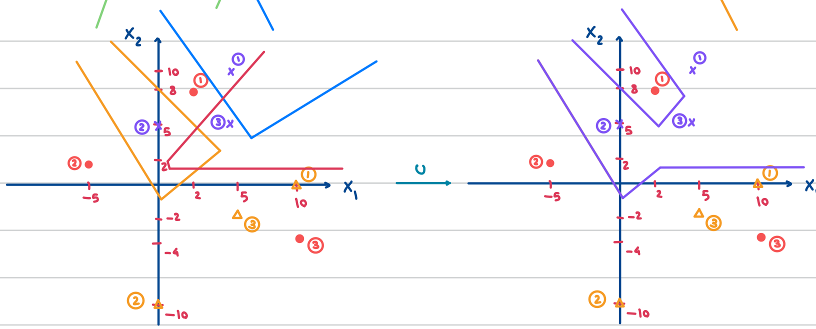
.3



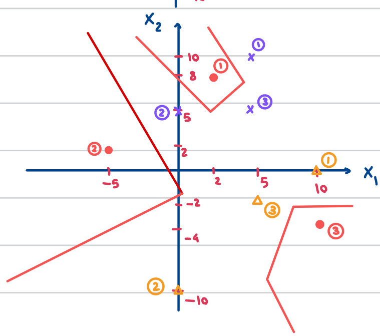
.4 ناحیه مربوط به کلاس w_1



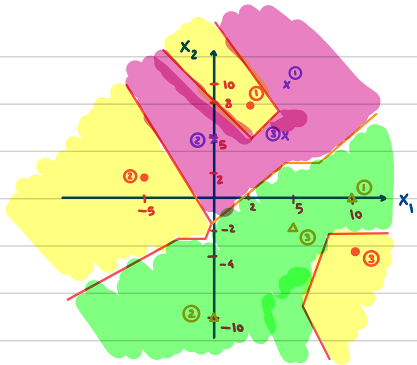
ناحیه مربوط به کلاس w_2



ناحیه مربوط به کلاس w_3



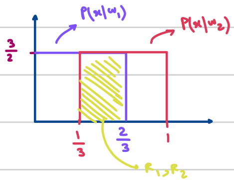
در نهایت با تجمیع این نواحی داریم :



5-

الف) قانون بیز : $P(\omega_i|x) = \dfrac{P(x|\omega_i)P(\omega_i)}{P(x)} \longrightarrow P(\omega_1|x) = \dfrac{P(x|\omega_1)P(\omega_1)}{P(x)} \xrightarrow{\text{فرض } P(\omega_1)=P(\omega_2)} P(\omega_1|x) = \dfrac{3}{4}$

به طریق مشابه : $P(\omega_2|x) = \dfrac{3}{4}$



$$P(e) = P(\underbrace{y=\omega_1, \hat{y}=\omega_2}_{x\in R_2}) + P(\underbrace{y=\omega_2, \hat{y}=\omega_1}_{x\in R_1})$$

$$= \int_{R_2} P(x|\omega_1)P(\omega_1)\,dx + \int_{R_1} P(x|\omega_2)P(\omega_2)\,dx$$

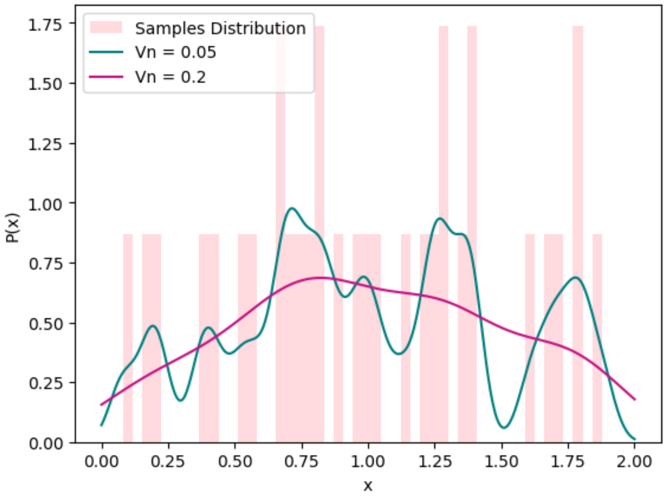
$$= \int_{\frac{1}{3}}^{\frac{2}{3}} \frac{3}{2}\times\frac{1}{2}\,dx + \int_{\frac{1}{3}}^{\frac{2}{3}} \frac{3}{2}\times\frac{1}{2}\,dx = \frac{1}{2}$$

با توجه به سطر ۲ و محل اشتراک آن x مورد نظر در آن محدوده یا به ترتیب صحیح داده یا غلط پس خطای آن $\frac{1}{2}$ است.

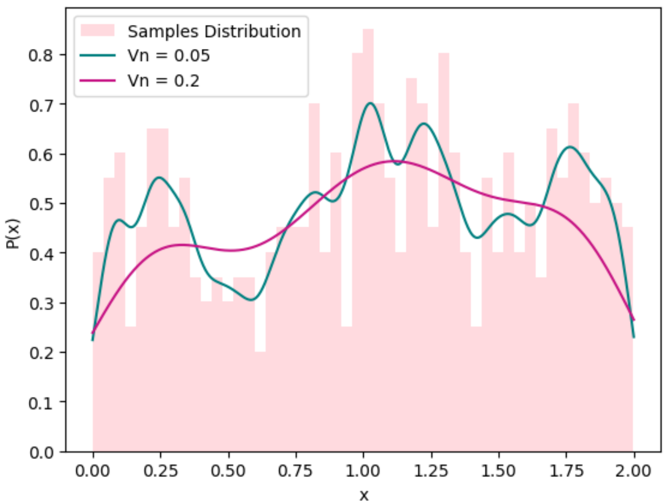
ب) احتمال خطا برابر است با احتمال اینکه نقطه تست به ω_2 نزدیکتر باشد، درنتیجه مساحتی که مشترکاً ω_1 و ω_2 دارند، احتمال خطا را تشکیل می‌دهند.

$$P(e_1) = \int_{R_1} P(x|\omega_2) = \int_{\frac{1}{3}}^{\frac{2}{3}} \frac{3}{2}\,dx = \frac{1}{2}$$

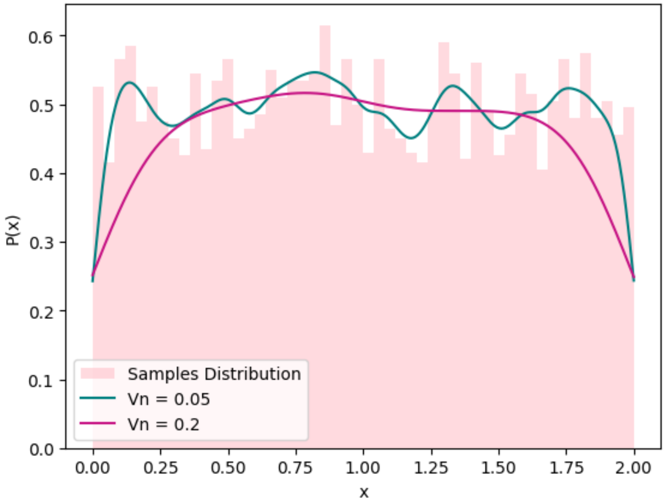
٦. نتایج را به شکل زیر خواهیم داشت:



شکل ۱: نتایج برای ۳۲ داده آموزشی



شکل ۲: نتایج برای ۲۵۶ داده آموزشی

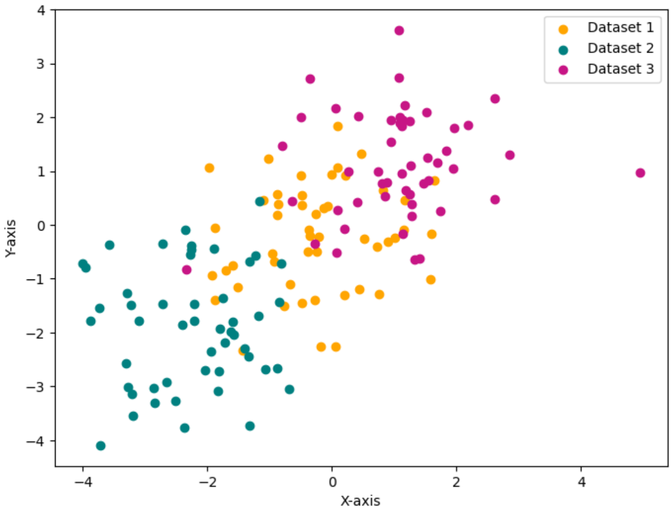


شکل ۳: نتایج برای ۵۰۰۰ داده آموزشی

با توجه به نمودارها مشاهده می‌کنیم به ازای N یکسان برای V_n بیشتر تقریب نرم‌تری خواهیم داشت. در واقع به نسبت نویز کمتری داریم.

همچنین با افزایش تعداد داده‌های آموزشی به ازای V_n یکسان، تقریب به نمودار واقعی نزدیک‌تر می‌شود.

۷. پراکندگی نمونه‌ها به شکل زیر خواهند بود:



شکل ۱: نمودار پراکندگی داده‌ها

٤ نقطه به شکل زیر در نظر می‌گیریم:

$$P_1 = \begin{pmatrix} -4 \\ -4 \end{pmatrix}, \quad P_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad P_3 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad P_4 = \begin{pmatrix} 1.5 \\ 0 \end{pmatrix}$$

با توجه به نمودار پراکندگی داده‌ها انتظار داریم برچسب داده‌ها به ترتیب زیر باشد:

$$\hat{y}_1 = 2, \quad \hat{y}_2 = 1, \quad \hat{y}_3 = 3, \quad \hat{y}_4 = 3$$

برای بررسی دو حالت خواهیم داشت:

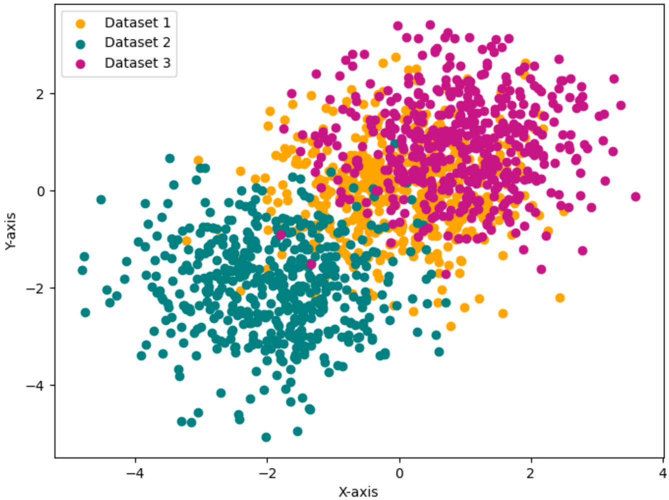
(آ) $N = 50, \quad V_n = 1$

(ب) $N = 50, \quad V_n = 0.1$

در حالت اول حجم همسایگی بیشتر است و تعداد بیشتری از نمونه‌ها داخل همسایگی قرار می‌گیرند. در نتیجه دقت تصمیم‌گیری در کل بالاتر است.

با تست نقاط، حالت (آ) در بیشتر مواقع پیش‌بینی درست است اما در حالت (ب) به دلیل کمتر بودن نمونه‌ها در یک همسایگی (به دلیل کاهش V_n) جواب بسته به پراکندگی‌های داده‌های آموزشی متفاوت خواهد بود. به خصوص نقطه P_4 که برچسب آن بین کلاس ۲ و ۳ تغییر می‌کند.

با افزایش تعداد داده‌های هر کلاس خواهیم داشت:



شکل ۲: نمودار پراکندگی داده‌ها برای $N = 500$

مشابه قبل دو حالت خواهیم داشت:

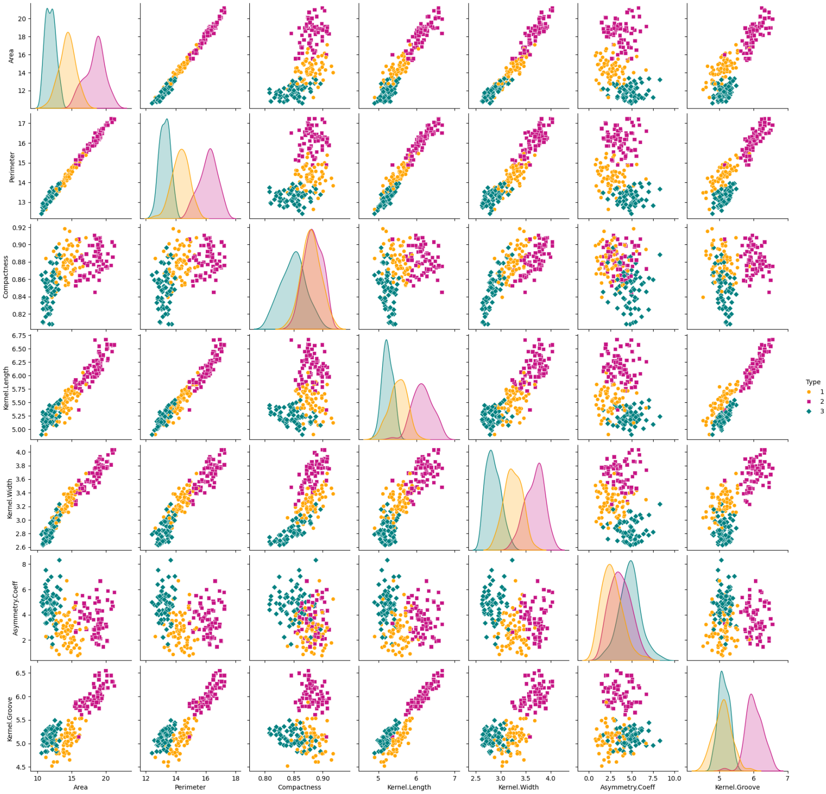
(آ) $N = 500, \quad V_n = 1$

(ب) $N = 500, \quad V_n = 0.1$

اینبار تعداد داده‌ها بیشتر شده و در یک حجم یکسان تعداد بیشتری نمونه داخل همسایگی قرار می‌گیرند. برای حالت (آ) نتایج دقت بالاتری خواهند داشت. در حالت (ب) در مقایسه با حالت قبل برچسب نقطه P_4 به درستی تعیین می‌شود اما در برخی تست‌ها برچسب نقطه P_2 بین کلاس ۱ و ۲ تغییر می‌کند.

در مجموع هر قدر N, V_n بیشتر باشند دقت بالاتری خواهیم داشت.

۸. (آ) EDA

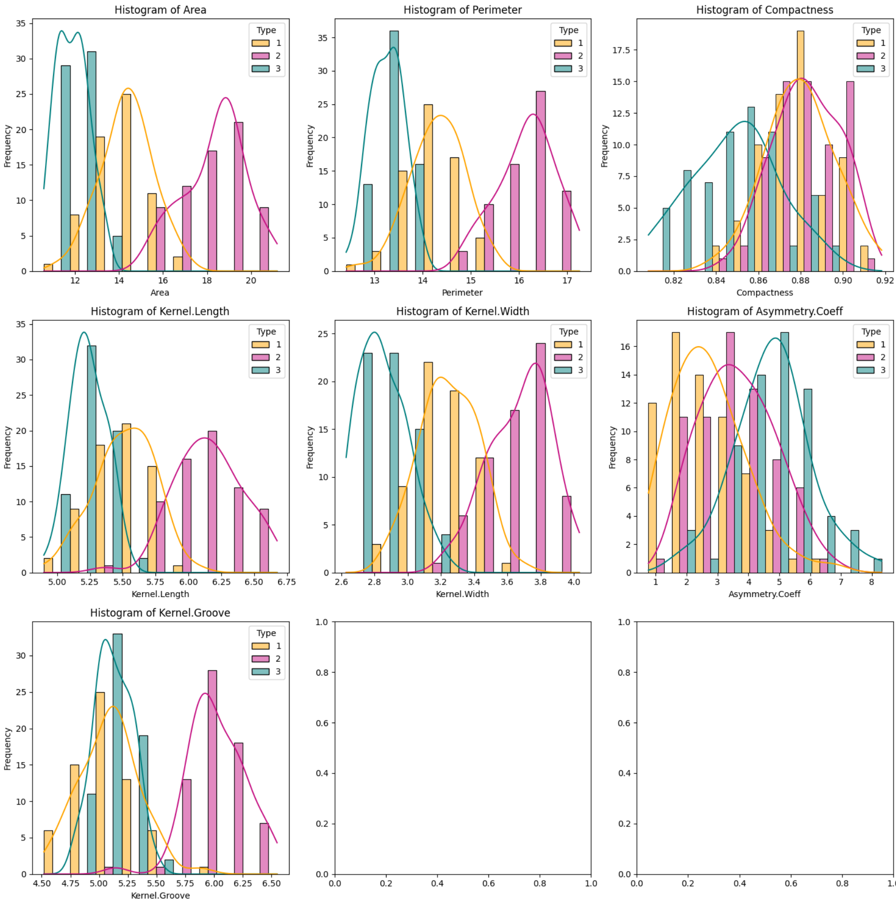


شکل ۱: نمودار ویژگی‌ها

با توجه به شکل ویژگی‌هایی که بهتر سه کلاس را جدا کردند:

(Area/Kernel:Groove),(Perimeter/Kernel:Groove),(Kernel:Length/Kernel:Groove)

حال هیستوگرام ویژگی‌ها را رسم می‌کنیم:



شکل ۲: نمودار هیستوگرام کلاس‌های هر ویژگی

(ب) برای پیش‌پردازش داده‌ها از z-score استفاده می‌کنیم.

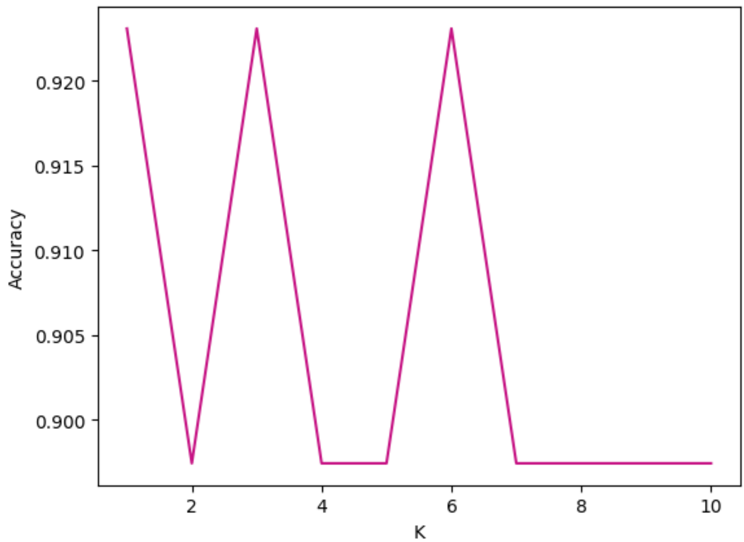
$$Z_{score} = \frac{x - \mu}{\sigma}$$

هر قدر این مقدار بیشتر باشد یعنی داده نویزی‌تر است و بهتر است حذف شود.

برای نرمال سازی داده‌ها برای هر ویژگی رابطه زیر را اعمال می‌کنیم:

$$x_{Normalized} = \frac{x - x_{min}}{x_{max} - x}$$

(د) با پیاده‌سازی الگوریتم KNN به نتیجه زیر خواهیم رسید:



شکل ۳: نمودار دقت به ازای مقادیر مختلف k

با توجه به نمودار بهترین مقدار k می‌تواند ۱، ۳ و ٦ باشد.