به نام خدا

یادگیری ماشین

تمرین سوم

محمد ناصری

۸۱۰۱۰۰۴۸۶

بهار ۱۴۰۱

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

۱- سوال اول (۲۰ نمره)

توزیع نرمال $p(x) \sim N(\mu.\sigma^2)$ و تابع پنجره پارزن $\varphi(x) \sim N(0.1)$ را در نظر بگیرید. نشان دهید که تخمین پنجره پارزن

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

برای h_nهای کوچک دارای ویژگی‌های زیر است:

- $\tilde{p}_n(x) \sim N(\mu.\, h_n^2 + \sigma^2)$
- $p_n(x) - \tilde{p}_n(x) \cong \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2 \left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right] p(x)$
- $var[p_n(x)] \cong \frac{1}{2nh_n\sqrt{\pi}} p(x)$

الف

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

$$= \frac{1}{h_n} \int_{-\infty}^{\infty} \varphi\left(\frac{x-u}{h_n}\right) p(u)\, du$$

$$= \frac{1}{h_n} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{x-u}{h_n}\right)^2}{2}\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\frac{u-\mu}{\sigma}\right)^2}{2}\right) du$$

$$= \frac{1}{2\pi h_n \sigma} \exp\left(-\frac{\frac{x^2}{h_n^2} + \frac{\mu^2}{\sigma^2}}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\left(\frac{1}{h_n^2} + \frac{1}{\sigma^2}\right) - 2u\left(\frac{x}{h_n^2} + \frac{\mu}{\sigma^2}\right)\right) du$$

$$\begin{cases} \theta^2 = \dfrac{\sigma^2 h_n^2}{\sigma^2 + h_n^2} \\ \alpha = \theta^2\left(\dfrac{x}{h_n^2} + \dfrac{\mu}{\sigma^2}\right) \end{cases} \Rightarrow \bar{p}_n(x) = \frac{\sqrt{2\pi}\theta}{2\pi h_n \sigma} \exp\left[\frac{-1}{2}\left(\frac{x^2}{h_n^2} + \frac{\mu^2}{\sigma^2}\right) + \frac{1}{2}\frac{\alpha^2}{\theta^2}\right]$$

$$\bar{p}_n(x) = \frac{1}{\sqrt{2\pi} h_n \sigma} \times \frac{h_n \sigma}{\sqrt{h_n^2 + \sigma^2}} \exp\left[\frac{-1}{2}\left(\frac{x^2}{h_n^2} + \frac{\mu^2}{\sigma^2} - \frac{\alpha^2}{\theta^2}\right)\right]$$

$$\frac{\alpha^2}{\theta^2} = \frac{\theta^4}{\theta^2}\left(\frac{x}{h_n^2} + \frac{\mu}{\sigma^2}\right)^2$$

$$= \frac{(x h_n)^2}{(h_n^2+\sigma^2)h_n^2} + \frac{(\mu\sigma)^2}{(h_n^2+\sigma^2)\sigma^2} - \frac{2x\mu}{h_n^2+\sigma^2} = \frac{(x-\mu)^2}{h_n^2+\sigma^2}$$

$$\bar{p}_n(x) \sim N(\mu, h_n^2 + \sigma^2)$$

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

۲- سوال دوم (۱۰ نمره)

متریک فاصله اقلیدسی را در d بعد در نظر بگیرید:

$$D(a.b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2}$$

فرض کنید عناصر هر بعد را در یک مقدار حقیقی غیر صفر ضرب میکنیم. یعنی k = 1, 2, ..., d داریم:

$$x_k = \acute{}\alpha_k x_k$$

ثابت کنید پس از ضرب نیز این متریک فاصله همچنان یک فاصله استاندارد است یعنی ویژگی‌های گفته شده برای یک فاصله استاندارد را دارا می‌باشد. در مورد تاثیر این امر بر طبقه بند knn بحث کنید.

فاصله استاندارد ←
① $D(a,b) \geq 0$ ← جمع مربعات مثبت است ✓
② $D(a,b)=0 \Leftrightarrow a=b$ ← زیر رادیکال بخواهد صفر شود باید $a=b$ ✓
③ $D(a,b)=D(b,a)$ ← $(a_k-b_k)^2=(b_k-a_k)^2$ ✓
④ $D(a,b)+D(b,c) \geq D(a,c)$ ←

$D(a,b) = \|A(a-b)\|_2$, $A = \mathrm{diag}[\alpha_1, \dots, \alpha_d]$

$x = a-c$, $y = c-b$ ⟶ $D(a,b)+D(b,c) \geq D(a,c) \Rightarrow \|Ax\|_2+\|Ay\|_2 \geq \|A(x+y)\|_2$

$w = Ax$, $z = Ay$ ⟶ $\|w\|_2+\|z\|_2 \geq \|w+z\|_2$ $\xrightarrow{\wedge 2}$ $\|w\|_2^2+\|z\|_2^2+2\|w\|_2\|z\|_2$

$\geq \|w\|_2^2+\|z\|_2^2+2\sum_{i=1}^{d} wz$

$\Rightarrow |w|_2 |z|_2 \geq \sum_{i=1}^{d} w_i z_i$ ✓

فضای متریک knn همیشه یک نقطه اولیه انتخاب شده P با حداقل فاصله از یک نقطه دیگر x^* خواهد بود

$\delta = \min_x D(P,x)$: $n \to \infty \Rightarrow D(P,x_n) \to \delta$

$D(P,x^*) = \delta$ ⇐ $D(P,x^*) \leq \delta$, $D(P,x^*) \geq \delta$ ⇐ $D(P,x_n)+D(x_n,x^*) \geq D(P,x^*)$ ⇐ $x_n \to x^*$ ✓

$\Rightarrow P(w_i|P) \simeq P(w_i|x^*)$

۳- سوال سوم (۲۰ نمره)

یک مسئله طبقه بندی با روش knn را در نظر بگیرید. مجموعه داده دو کلاسه D را نیز به صورت $D= \{x^q, \omega_i{}^q\}$, q=1,..., Q داریم. این داده ها نتایج یک نظر سنجی بوده و دیتاپوینت ها به صورت پرچسب خورده، مستقل از هم هستند و فرض می کنیم تعداد داده های دو کلاس یکسان است. برای هر سمپل تست نزدیک ترین k دیتاپوینت را به صورت $\{x_i\}$ i=1, ..., k نمایش می دهیم. هر دوی p(x|1) و p(x|2) توزیع یکنواخت بر روی یک کره به شعاع واحد دارند و مرکز دو ابر کره نیز از هم ۱۰ واحد فاصله دارند.

الف) نشان دهید اگر k فرد باشد متوسط احتمال خطا از رابطه زیر به دست می‌آید:

$$p_Q(e) = \frac{1}{2^Q}\sum_{j=0}^{\frac{k-1}{2}} \binom{Q}{j}$$

ب) با توجه به بخش قبل نشان دهید که در این حالت خطای طبقه بند نزدیک ترین همسایه کمتر از حالت $k \geq 2$ است و دلیل مشاهده این موضوع را توضیح دهید.

$\|x\|<1$ ، 1 ، 2 ، o.w $\Rightarrow P(\omega_1|x)$

$|x-a|<1$ ، 1 ، 2 ، o.w $\Rightarrow P(\omega_2|x)$

توزیع یکنواخت

ج) نشان دهید: $\lim_{Q\to\infty} p_Q(e) = 0$

الف

$P_Q(e) = Pr[e\in\omega_1, \omega_2 \text{ is frequent}] + Pr[e\in\omega_2, \omega_1 \text{ is frequent}]$

$= 2Pr[e\in\omega_1, \omega_2 \text{ is freq}] = 2P(\omega_1)\, Pr(\text{count}(\omega_2) < \frac{k-1}{2})$

$= \sum_{j=0}^{(k-1)/2} \binom{Q}{j}\frac{1}{2^j} \times \frac{1}{2^{(Q-j)}} = \frac{1}{2^Q}\sum_{j=1}^{(k-1)/2}\binom{Q}{j}$

ب

$k=1 \longrightarrow P_n(e) > \frac{1}{2^Q} \longrightarrow \frac{1}{2^Q}\sum_{j=0}^{(k-1)/2}\binom{Q}{j} \longrightarrow k>1$

ج

$P_n(e) = Pr[Bin(n, \tfrac{1}{2}) \leq (k-1)/2] = Pr[Y_1+\dots+Y_n \leq \frac{k-1}{2}]$

$Q\to\infty \Rightarrow P_n\to 0 \longleftarrow \frac{k-1}{2} < \frac{\sqrt{Q}}{2} - 1 < 0 \longleftarrow$ ($Q\to\infty$)

سوال ۴

الف)

در ابتدا برای این سوال لازم است که مفهم Bias variance trade offدر یادگیری ماشین بیان گردد. در ابتدا مفهوم بایاس یا سوگیری بیان میشود که این مفهوم ناشی از فرضیات غلط و اشتباه در یادگیری است به عبارتی یک مدل با مقدار بایاس بالا توجه بسیار کمی به داده‌های آموزشی دارد و مدل را بیش از حد ساده در نظر میگیرد. اما واریانس به این معناست که یک خطا از حساسیت به نوسانات کوچک در مجموعه داده‌ی آموزش است، به این معنا که واریانس بالا سبب میشود که یک الگوریتم به جای خروجی‌های مورد نظر نویزهای تصادفی را در داده های آموزشی مدل کند. به عبارت دیگر یک مدل با میزان واریانس بالا توجه زیادی به داده‌های آموزشی دارد و قابل تعمیم به داده های دیده نشده نیست. یک عبارت ریاضی برای این مفاهیم به شکل زیر است:

$$\text{Bias}(X) = E[\hat{f}(X)] - f(X)$$

$$\text{Var}(X) = E\left[\left(\hat{f}(X) - E[\hat{f}(X)]\right)^2\right]$$

بایاس در واقع تفاوت بین برچسب واقعی و مقدار برچسب پیش‌بینی شده می‌باشد اما واریانس در آمار تعریف میشود و امید ریاضی مربع انحراف یک متغیر تصادفی از میانگین آن است که در اینجا fنشان دهنده‌ی مدل ما در دنیای واقعی است. همچنین باید توجه کرد که ما دارای یک نویز تصادفی هستیم، که از آن نمیتوانیم اجتناب کنیم و آن را با اپسیلون نشان میدهند همچنین توجه شود که true labelرا به شکل زیر نشان میدهند

$$y = f(X) + \epsilon.$$

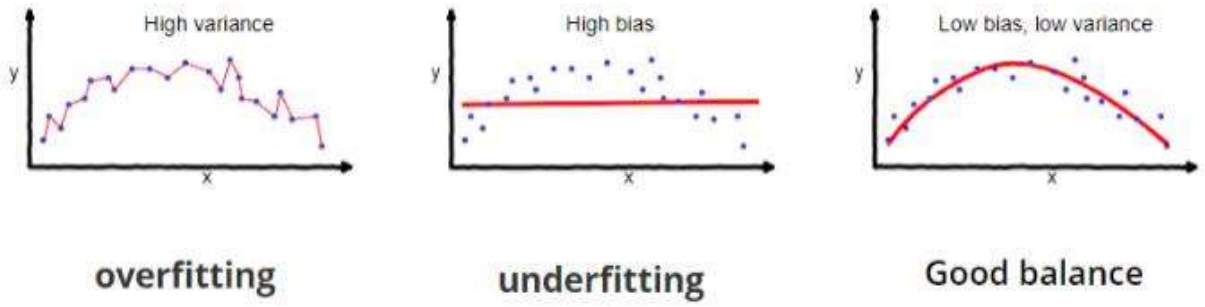
و ما میتوانیم خطا را به شکل زیر محاسبه کنیم

$$\begin{aligned}
\text{Err}(X) &= E\left[\left(y - \hat{f}(X)\right)^2\right] \\
&= E\left[\left(f(X) + \epsilon - \hat{f}(X)\right)^2\right] \\
&= \left(E[\hat{f}(X)] - f(X)\right)^2 + E\left[\left(\hat{f}(X) - E[\hat{f}(X)]\right)^2\right] + \sigma_\epsilon^2 \\
&= \text{Bias}^2 + \text{Variance} + \text{Random Error}.
\end{aligned}$$

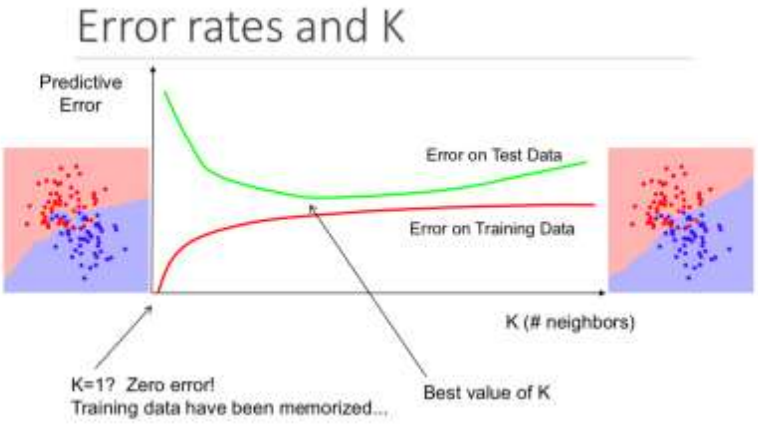
ضمنا توجه شود که دو اصطلاح دیگر در مورد بایاس و واریانس وجود دارد به نامهای Overfittingو Underfitting. منظور از : Overfittingبه این معناست که مدل داده های آموزش را به خوبی پیش‌بینی میکند.اما در مورد داده های دیده نشده عملکرد خوبی ندارد

منظور از Underfitting :این است که مدل با داده های آموزش به خوبی مطابقت ندارد و به عبارت دیگر، داده های آموزش را به خوبی پیش‌بینی نمیکند.

به طور معمول Underfittingنشان دهنده ی بایاس زیاد و واریانس کم و Overfittingنشان دهنده‌ی بایاس کم و واریانس زیاد است



با توجه به میزان Kباید توجه داشت که KNNدر ابتدا میزان یادگیری پایینی دارد، در K=1داده‌های یادگیری به خوبی پیشبینی میشوند و به این معناست که مقدار بایاس باید برابر با صفر باشد، اما در K=1زمانی که داده ی جدید دیده نشده داریم(داده‌ی تست) شانس خطا بالاتر میرود به این معنا که واریانس بالایی داریم. اما زمانی که میزان Kبیشتر میشود، خطای یادگیری(افزایش بایاس) افزایش می یابد اما خطای تست ممکن است با کاهش روبرو شود(کاهش واریانس.) ما میتوانیم فکر کنیم که با افزایش Kتعداد همسایه ها بیشتر میشوند و مدل پیچیده تر میگردد، با مدل زیر به راحتی میتوان میزان بایاس و واریانس را با افزایش سایز kبیان کنیم.



همانطور که در شکل بالا مشخص است در K=1میزان خطای داده های تست بسیار بالا و میزان خطای یادگیری در حدود صفر است و مطالب گفته شده مطابق نمودار بالا نشان داده شده است

برای پنجره پارزن ما با مفهومی به نام Undersmoothingو Oversmoothingروبرو هستیم زمانی که hکوچک باشد، مقدار واریانس زیاد است و مقدار بایاس کوچک است. زمانی که hبزرگ است، مقدار واریانس کوچک است اما بایاس زیاد است.

ب)

در ابتدا لازم است که توضیح کوتاهی در مورد روشهای پارامتریک بیان گردد، مفروضات در مورد شکل یک تابع میتواند روند یادگیری را ساده نماید و مدلهای پارامتریک با ساده سازی تابع به شکل شناخته شده مشخص میشود و مدل پارامتریک یادگیرنده‌ای است که داده‌ها را از طریق مجموعه‌ای از پارامترها خلاصه میکند. پارامترها دارای اندازه‌ی ثابتی هستند یعنی اینکه مدل از قبل تعداد پارامترهای مورد نیاز خود را بدون توجه به داده‌های آنها میداند. در واقع پارامترها نیز مستقل از تعداد نمونه‌های آموزشی هستند. در مدلهای پارامتریک، دو مرحله وجود دارد، اولین مرحله انتخاب فرم تابع است و مرحلهی دوم یادگیری ضرایب تابع از داده‌های آموزشی است. نمونه‌های از مدلهای پارامتری شامل Logistic regressionو SVMخطی است.

مدلهای غیرپارامتریک

الگوریتمهایی که مفروضات خاصی در مورد نوع تابع نگاشت ندارند به عنوان الگوریتمهای غیرپارامتریک شناخته میشوند. این روشها آزادی را دارند که هر فرم را از داده‌های آموزشی انتخاب کنند. در نتیجه مدل‌های پارامتریک برای تخمین تابع نگاشت به داده‌های بیشتری نسبت به مدل‌های پارامتریک نیاز دارند. ممکن است تصور شود که روشهای غیرپارامتریک به این معنی است که هیچ پارامتری وجود ندارد. اما این موضوع صحیح نیست بلکه به این معنی است که پارامترها نه تنها قابل تنظیم هستند بلکه میتوانند تغییر کنند. این منجر به تمایز کلیدی بین الگوریتم‌های پارامتریک و غیرپارامتریک است، همانطور که پیش از این گفتیم الگوریتم‌های پارامتریک بدون توجه به میزان داده‌های آموزشی تعداد پارامترهای ثابتی دارند. با این حال در مدل‌های غیرپارامتریک تعداد پارامترها به مقدار داده‌های آموزشی بستگی دارد و هر چه داده‌های آموزشی بیشتر باشد تعداد پارامترها بیشتر میشود و نتیجه این موضوع این است که در الگوریتم‌های غیرپارامتریک ممکن است فرایند آموزش بسیار طول بکشد. به عنوان یک مثال الگوریتم KNNنمونه‌ای از الگوریتم غیرپارامتریک است. توجه شود که هیچ فرضی در مورد شکل تابع نگاشت به جز یک فرض وجود ندارد. فرض بر این است که مشابه‌ترین الگوهای آموزشی احتمال بیشتری برای تولید خروجی مشابه دارند. در این قسمت در مورد فواید و محدودیت‌های هر کدام از این مدلها صحبت میکنیم

سادگی :روش‌های الگوریتم‌های پارامتریک برای درک آسانتر است. تفسیر پذیری نتایج نیز در مقایسه با مدل‌های غیرپارامتریک آسانتر است.

داده های آموزشی :مدل‌های پارامتریک به نسبت مدل‌های غیرپارامتریک به داده‌های کمتری نیاز دارند.

سرعت آموزش :روش‌های پارامتریک از نظر محاسباتی سریعتر از روش‌های غیرپارامتریک هستند، آنها را میتوان سریعتر از روش‌های غیرپارامتریک آموزش داد زیرا معمولا پارامترهای کمتری برای آموزش نیاز دارند

در مورد مدلهای **غیر پارامتریک:**

کارایی :مدل‌های غیرپارامتریک ممکن است، پیش‌بینی‌های دقیقتری ارائه دهند، زیرا تناسب بهتری با داده‌ها نسبت به مدل‌های پارامتریک ارائه میدهند.

انعطاف پذیری :این الگوریتم های تناسب خوبی برای داده‌ها فراهم میکنند و میتوانند بسیاری از اشکال یک تابع را جا بدهند.

فرضیات کمی دارند :در مقایسه با الگوریتم‌های پارامتریک الگوریتم‌های غیرپارامتریک بیشتر از داده‌ها یاد میگیرند. این به این دلیل است که یادگیری الگوریتم‌های پارامتری ممکن است با مفروضاتی که آنها ایجاد میکنند محدود شوند

محدودیتها:

در مدل های پارامتریک:

محدودیت فرم :محدودیت مدل‌های پارامتریک در تعیین فرم تابعی است.

محدودیت در *Fit* :این روش‌ها بهترین تناسب را با داده‌ها ارائه نمیدهند.

پیچیدگی :الگوریتم‌های پارامتریک پیچیدگی محدودی ارائه میدهند، به این معنی است که آنها برای مسئله‌های با مقدار کمتری پیچیدگی مناسب هستند.

مدلهای پارامتریک:

بیش برازش:

به همان اندازه که این الگوریتم‌ها تمایل دارند، داده‌هارا بهتر از الگوریتم‌های پارامتریک برازش دهند بیشتر مستعد برازش بیش از حد هستند

داده های آموزش :این روشها به داده‌های بسیار بیشتری نسبت به الگوریتم‌های پارامتریک نیاز دارد.

سرعت :الگوریتم‌های غیرپارامتریک آهسته تر آموزش داده میشوند زیرا معمولا پارامترهای بیشتر برای آموزش در نظر میگیرند

ج)

تعدادی از معایب و مشکلات این روش‌ها را بیان میکنیم:

اولین مشکل این روش‌ها غیر قابل درک بودن آنهاست: درک آن چیزی که روشهای کرنل آموزش دیده‌اند دشوار است. دومین مشکل، انتخاب کردن یک کرنل مناسب است. امکان تعمیم این مسائل به چندکلاس وجود ندارد، در مقایسه با روش‌هایی مثل درخت تصمیم که میتواند به راحتی با چند کلاس سازگار گردد اما تعمیم به مسائل چند کلاسه با روش‌های مبتنی بر کرنل سخت است. مشکل بعدی این است ک این روش‌ها به خوبی روش‌های یادگیری عمیق عمل نمیکنند. یک مشکل این است که هزینه محاسباتی روشهای کرنل معقول است اما ناچیز نیست و زمانی که تعداد ابعاد

زیاد باشد ولی تعداد نمونه‌ها نسبتا کم باشد(کمتر از یک میلیون) روش‌های مبتنی بر کرنل از نظر محاسباتی مناسب هستند.

د)

مقدار P(X) طریق رابطه $P(X) = \frac{K}{nV}$محاسبه میشود که در دو روش پارزن و KNNتفاوت در ثابت نگه داشتن مقادیر Kیا Vمیباشد. در روش پارزن مقدار حجمV ثابت نگه داشته میشود ومشخص میشود چه مقدار K درون آن حجم قرار میگیرد که این مقدار حجم دارای رابطه‌ی $V_n = 1/\sqrt{n}$با تعداد نمونه ها دارد.
اما در روش KNNپارامتر Kثابت نگه داشته میشود و حجم مربوط از دادهها که این مقدار Kرا در خود جا میدهد مشخص میشود. جزو شرایط این روشهای این است که مقدار V_nدر بی نهایت به صفر میل نماید و همچنین یعنی اگرچه V_nدر بینهایت به صفر میل میکند اما سرعت رشد آن باید از سرعت رشد داده ها کمتر باشد.

سوال ۴-۲

ج)

طبقه بندی KNNدارای این فرض است که نقاط مشابه یکدیگر دارای برچسب مشابه هم هستند. اما در فضاهای با ابعاد بالا نقاطی که براساس یک توزیع احتمالی هستند، به یکدیگر نزدیک نخواهند بود. به طور کلی طبقه بندهایی مثل KNNکه به فاصله‌ی زوجی بین نقاط متکی هستند به شدت تحت تاثیر مشکل نحسی ابعاد قرار میگیرند. در مسئله‌ی KNNما یک نقطه تست داریم و میخواهیم Kتا از نزدیکترین همسایه‌ها به این نقطه‌ی تست را بیابیم. به وضوح میتوانیم مشاهده کنیم با افزایش بعد مسئله، در فضاهای با تعداد بعد بالاتر داده‌های آموزشی چگالی کمتری خواهند داشت و برای یافتن همسایگان نقطه‌ی تست باید حجم زیادی از فضا جستجو شود. پس فاصله بین جفت نقاط با افزایش بعدها افزایش مییابد و در این صورت همسایه‌ها ممکن است از نقطه‌ی تست آنقدر دور باشند که دیگر اشتراک زیادی با آن نقطه تست نداشته باشند. به طور کلی طول کوچکترین Hyper Cubeکه شامل همه Kنزدیکترین همسایه یک نقطه تست میباشد برابر است با $(k/N)^{1/d}$که در اینجا Nتعداد نمونه‌ها و dهم اندازه ابعاد ماست.

از این عبارت میتوان به این نتیجه رسید که با افزایش خطی تعداد ابعاد، تعداد نمونه‌های آموزشی باید به صورت تصاعدی افزایش یابد تا با نحسی ابعاد مقابله کند. همچنین روش دیگر برای مقابله با این موضوع این است که بعد را کاهش بدهیم یا داده‌ها را به ابعادی با فضایی پایینتر تبدیل کنیم.