تکلیف ششم درس شناسایی الگو

وحید ملکی
شماره دانشجویی: ۴۰۳۱۳۰۰۴

۷ نوامبر ۲۰۲۵

سوال ۷: نفرین ابعاد (Curse of Dimensionality)

الف- نسبت مساحت دایره به مربع در فضای دو بعدی

ما یک دایره به شعاع r محاط در یک مربع به ضلع $2r$ داریم. مساحت دایره (V_c) و مساحت مربع (V_s) به صورت زیر محاسبه می‌شوند:

$$V_c = \pi r^2$$

$$V_s = (2r)^2 = 4r^2$$

بنابراین نسبت مساحت‌ها برابر است با:

$$\frac{V_c}{V_s} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} \approx 0.785$$

این بدان معناست که در فضای دوبعدی، حدود 78.5% از نمونه‌ها (با توزیع یکنواخت) درون دایره قرار می‌گیرند.

ب- نسبت فاصله میانگین تا ضلع و گوشه مربع

فاصله مرکز (مبدا) تا نزدیکترین نقطه روی ضلع مربع برابر با شعاع دایره محاطی است:

$$d_{side} = r$$

فاصله مرکز تا گوشه مربع (وتر مثلث قائم‌الزاویه با اضلاع r) برابر است با:

$$d_{corner} = \sqrt{r^2 + r^2} = r\sqrt{2}$$

نسبت فاصله مرکز از ضلع به فاصله آن از گوشه برابر است با:

$$\frac{d_{side}}{d_{corner}} = \frac{r}{r\sqrt{2}} = \frac{1}{\sqrt{2}} \approx 0.707$$

ج- نسبت حجم فوق کره به فوق‌مکعب در فضای n بعدی

رابطه حجم فوق کره به شعاع r در فضای n بعدی داده شده است:

$$V_c(n) = \frac{\pi^{n/2} r^n}{\Gamma(\frac{n}{2}+1)}$$

حجم فوق‌مکعب به ضلع $2r$ در فضای n بعدی برابر است با:

$$V_s(n) = (2r)^n = 2^n r^n$$

نسبت این دو حجم برابر است با:

$$\frac{V_c(n)}{V_s(n)} = \frac{\frac{\pi^{n/2} r^n}{\Gamma(\frac{n}{2}+1)}}{2^n r^n} = \frac{\pi^{n/2}}{2^n \Gamma(\frac{n}{2}+1)} = \frac{(\frac{\pi}{4})^{n/2}}{\Gamma(\frac{n}{2}+1)}$$

د- بررسی حد نسبت حجم‌ها وقتی $n \to \infty$

با استفاده از تقریب استرلینگ داده شده برای تابع گاما:

$$\Gamma(x+1) \approx (2\pi)^{0.5} e^{-x} x^{x+0.5}$$

با فرض $x = \frac{n}{2}$، مخرج کسر قسمت (ج) را تقریب می‌زنیم:

$$\Gamma(\frac{n}{2}+1) \approx \sqrt{2\pi} e^{-n/2} (\frac{n}{2})^{\frac{n+1}{2}}$$

حال نسبت حجم‌ها را بازنویسی می کنیم:

$$\text{Ratio} \approx \frac{\pi^{n/2}}{2^n \left[\sqrt{2\pi} e^{-n/2} (\frac{n}{2})^{\frac{n}{2}} (\frac{n}{2})^{0.5}\right]}$$

با ساده‌سازی عبارت:

$$\text{Ratio} \approx \frac{1}{\sqrt{\pi n}} \left(\frac{\pi e}{2n}\right)^{n/2}$$

همانطور که مشاهده می‌شود، در عبارت داخل پرانتز، مخرج شامل ترم n است. وقتی $n \to \infty$، عبارت $\frac{\pi e}{2n}$ به سمت صفر میل می کند و به توان $n/2$ نیز می‌رسد که با سرعت بسیار زیادی به صفر نزدیک می‌شود.

$$\lim_{n\to\infty} \frac{V_c}{V_s} = 0$$

این نشان می‌دهد که در ابعاد بالا، حجم فوق کره (ناحیه مرکزی) نسبت به حجم فوق‌مکعب ناچیز است.

ه- نسبت فاصله مرکز تا وجه و گوشه در فضای n بعدی

فاصله مرکز تا نزدیکترین وجه فوق‌مکعب همچنان برابر است با:

$$d_{face} = r$$

فاصله مرکز تا دورترین نقطه (گوشه فوق‌مکعب) با استفاده از تعمیم قضیه فیثاغورس در فضای n بعدی برابر است با:

$$d_{corner} = \sqrt{\sum_{i=1}^{n} r^2} = \sqrt{nr^2} = r\sqrt{n}$$

نسبت این فواصل:

$$\frac{d_{face}}{d_{corner}} = \frac{r}{r\sqrt{n}} = \frac{1}{\sqrt{n}}$$

حد این نسبت وقتی $n \to \infty$:

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0$$

و- نتیجه‌گیری و مفهوم نفرین ابعاد

نتیجه‌گیری هندسی: بر اساس بند (د)، حجم در دسترس در فضای ابعاد بالا به سرعت به سمت صفر میل می‌کند، به این معنی که تقریباً تمام حجم یک فوق‌مکعب در "گوشه‌ها" (Corners) و پوسته‌های بیرونی آن متمرکز شده است و مرکز آن تقریباً خالی است. بر اساس بند (ه)، فاصله گوشه‌ها از مرکز بسیار بیشتر از فاصله وجوه از مرکز می‌شود، به طوری که شکل به نوعی "تیغ‌دار" می‌شود.

چرا این یک آفت (Curse) است؟

۱. **پراکندگی داده‌ها (Data Sparsity):** برای اینکه بتوانیم فضای ویژگی را به درستی با نمونه‌های آموزشی پوشش دهیم، با افزایش ابعاد، تعداد نمونه‌های مورد نیاز به صورت نمایی افزایش می‌یابد. از آنجا که تعداد داده‌ها محدود است، فضا بسیار خالی می‌ماند و تخمین چگالی دشوار می‌شود.

۲. **بی‌معنی شدن فاصله اقلیدسی:** همانطور که نشان داده شد، فاصله دورترین و نزدیک‌ترین نقاط به هم نزدیک می‌شوند (به صورت نسبی). این باعث می‌شود که معیارهای شباهت مبتنی بر فاصله (مانند الگوریتم KNN) کارایی خود را از دست بدهند، زیرا همه نمونه‌ها تقریباً در فاصله یکسانی از هم قرار می‌گیرند و مفهوم "همسایگی" از بین می‌رود.

۳. **خطر بیش‌برازش (Overfitting):** به دلیل خالی بودن فضا، دسته‌بندها می‌توانند به راحتی مرزهای تصمیم‌گیری پیچیده‌ای پیدا کنند که نویزها را مدل کرده و روی داده‌های جدید تعمیم‌پذیری (Generalization) ضعیفی داشته باشند.

سوال ۱۱: دسته‌بند KNN با وزن‌دهی ویژگی و فاصله

الف- انتخاب وزن مناسب برای ویژگی‌ها

هدف از وزن‌دهی به ویژگی‌ها در اینجا، کاهش اثر مقیاس یا افزایش اهمیت ویژگی‌های تفکیک‌کننده است. با نگاه به نمودار داده‌ها، مشاهده می‌کنیم:

- کلاس‌ها (علامت‌های $+$ و $-$) عمدتاً در امتداد **محور افقی** (x_1) از هم جدا شده‌اند (کلاس مثبت در سمت چپ و کلاس منفی در سمت راست).
- در امتداد **محور عمودی** (x_2)، هر دو کلاس پراکندگی زیادی دارند و همپوشانی کامل وجود دارد. این یعنی محور عمودی اطلاعات تفکیک‌کننده کمتری دارد و بیشتر شبیه به نویز عمل می‌کند.

بنابراین، برای افزایش دقت دسته‌بندی و برجسته کردن ویژگی مهم‌تر، باید وزن بیشتری به x_1 و وزن کمتری به x_2 بدهیم تا فاصله‌ها در راستای افقی (که نشان‌دهنده تغییر کلاس است) تاثیر بیشتری در محاسبه فاصله اقلیدسی داشته باشند.
نتیجه:

$$w_1 = \frac{2}{3} \quad (\text{برای محور } x_1), \quad w_2 = \frac{1}{3} \quad (\text{برای محور } x_2)$$

ب- تعیین کلاس نقطه مجهول با KNN ($k = 5$)

نقطه مجهول (تست) را با x_u نشان می‌دهیم. با مشاهده چشمی نمودار، مختصات تقریبی نقطه مجهول برابر است با:

$$x_u \approx (5, 6.2)$$

فرمول فاصله اقلیدسی وزن‌دار برای دو نقطه x و y عبارت است از:

$$d(x,y) = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2} = \sqrt{\frac{2}{3}(x_1 - y_1)^2 + \frac{1}{3}(x_2 - y_2)^2}$$

حال ۵ نزدیک‌ترین همسایه را به صورت چشمی انتخاب کرده و فاصله آن‌ها را محاسبه می‌کنیم. (فرض می‌کنیم نزدیک‌ترین همسایه‌ها شامل ۴ نمونه از کلاس منفی و ۱ نمونه از کلاس مثبت که نزدیک‌تر است باشند):

همسایه	کلاس	مختصات تقریبی	اختلاف مختصات $(\Delta x_1, \Delta x_2)$	فاصله وزن‌دار (d)	وزن رأی $(w_k = \frac{1}{d})$
۱	$-$	$(5, 5)$	$(0, 1.2)$	$\sqrt{0 + \frac{1}{3}(1.44)} \approx \mathbf{0.69}$	**1.45**
۲	$-$	$(6, 6.2)$	$(1, 0)$	$\sqrt{\frac{2}{3}(1) + 0} \approx \mathbf{0.81}$	**1.23**
۳	$-$	$(5, 7.2)$	$(0, 1)$	$\sqrt{0 + \frac{1}{3}(1)} \approx \mathbf{0.58}$	**1.72**
۴	$+$	$(4, 6)$	$(1, 0.2)$	$\sqrt{0.66 + 0.013} \approx \mathbf{0.82}$	**1.21**
۵	$-$	$(6, 7.2)$	$(1, 1)$	$\sqrt{0.66 + 0.33} \approx \mathbf{1.00}$	**1.00**

جدول ۱: محاسبه فاصله و وزن برای ۵ همسایه نزدیک

محاسبه مجموع امتیازات (Weights) برای هر کلاس:

- **کلاس منفی ($-$):** مجموع وزن‌ها = $1.45 + 1.23 + 1.72 + 1.00 = \mathbf{5.40}$
- **کلاس مثبت ($+$):** مجموع وزن‌ها = $\mathbf{1.21}$

نتیجه‌گیری: با توجه به اینکه مجموع وزن‌های کلاس منفی بسیار بیشتر از کلاس مثبت است، نقطه x_u به **کلاس منفی ($-$)** تعلق می‌گیرد.

سوال ۱۲: رفتار دسته‌بند ۱NN

پاسخ

بله، این امکان وجود دارد که تمامی نمونه‌های تست به کلاس مثبت اختصاص یابند، حتی اگر در داده‌های آموزشی کلاس منفی نیز وجود داشته باشد.

تشریک و مثال

در دسته‌بند «یک نزدیک‌ترین همسایه» ،(۱NN) فضای ویژگی بر اساس موقعیت داده‌های آموزشی به نواحی مختلفی تقسیم می‌شود که به آن‌ها «سلول‌های ورونوی» (Voronoi Cells) می‌گویند. هر نقطه در فضا برچسب نمونه‌ای را می‌گیرد که به آن نزدیک‌تر است. اگر موقعیت نمونه‌های تست به گونه‌ای باشد که همگی در سلول‌های مربوط به نمونه‌های کلاس «مثبت» قرار گیرند (یعنی فاصله اقلیدسی آن‌ها تا یک نمونه مثبت کمتر از فاصله تا هر نمونه منفی باشد)، همگی مثبت برچسب گذاری می‌شوند. این حالت معمولاً زمانی رخ می‌دهد که نمونه‌های کلاس منفی «داده‌های پرت» (Outlier) باشند یا از ناحیه تمرکز داده‌های تست بسیار دور باشند.

مثال عددی (در فضای یک بعدی):

فرض کنید داده‌های آموزشی ($\mathcal{D}_{train}$) شامل دو نمونه باشد:

- یک نمونه از کلاس مثبت (+) در موقعیت $x = 2$
- یک نمونه از کلاس منفی (−) در موقعیت $x = 10$

مرز تصمیم‌گیری دقیقاً وسط این دو نقطه، یعنی در $x = 6$ قرار دارد. هر عددی کمتر از ۶ به کلاس مثبت و هر عددی بیشتر از ۶ به کلاس منفی تعلق می‌گیرد.

حال فرض کنید داده‌های تست ($\mathcal{D}_{test}$) شامل نقاط زیر باشند:

$$\{1, 3, 4, 5.5\}$$

از آنجا که تمام این اعداد از ۶ کوچک‌تر هستند، فاصله همه آن‌ها تا نمونه مثبت (2) کمتر از فاصله‌شان تا نمونه منفی (10) است. بنابراین خروجی دسته‌بند برای **تمام** نمونه‌های تست، کلاس «مثبت» خواهد بود.

سوال ۱۳: محاسبات و مفاهیم KNN

الف- دسته‌بندی نمونه $(3,6)$ با فاصله منهتن (۱NN)

نمونه تست را $T=(3,6)$ در نظر می‌گیریم. فاصله منهتن (City Block) با فرمول $d=|x_1-p_1|+|x_2-p_2|$ محاسبه می‌شود. فاصله نمونه تست را تا ۵ نمونه آموزشی موجود در جدول محاسبه می‌کنیم:

نمونه آموزشی (P_i)	برچسب (Y)	محاسبه فاصله $(\lvert 3-x_1\rvert+\lvert 6-x_2\rvert)$	فاصله نهایی (d)
$(1,7)$	$+1$	$\lvert 3-1\rvert+\lvert 6-7\rvert=2+1$	3
$(3,3)$	$+1$	$\lvert 3-3\rvert+\lvert 6-3\rvert=0+3$	3
$(1,1)$	-1	$\lvert 3-1\rvert+\lvert 6-1\rvert=2+5$	7
$(5,4)$	-1	$\lvert 3-5\rvert+\lvert 6-4\rvert=2+2$	4
$(2,5)$	-1	$\lvert 3-2\rvert+\lvert 6-5\rvert=1+1$	**2**

جدول ۲: محاسبه فاصله منهتن نمونه تست تا نمونه‌های آموزشی

نتیجه: کمترین فاصله برابر با **2** است که مربوط به نمونه $(2,5)$ می‌باشد. چون برچسب این نمونه -1 است، طبق قاعده ۱NN، نمونه تست به کلاس منفی (-1) تعلق می‌گیرد.

ب- بررسی شکل مرز تصمیم‌گیری در فاصله منهتن

پاسخ: نادرست

دلیل: در فاصله منهتن (L_1)، مکان هندسی نقاطی که از دو نقطه خاص فاصله یکسانی دارند، لزوماً خطوط موازی با محورهای مختصات نیستند. این مرزها می‌توانند شامل پاره‌خط‌هایی با زاویه ۴۵ درجه (شیب ± 1) نیز باشند. بنابراین مرز تصمیم‌گیری مجموعه‌ای از قطعات خطی است که می‌تواند شامل خطوط افقی، عمودی و یا مورب باشد.

ج- تاثیر ضرب مختصات در اسکالر بر خروجی KNN

پاسخ: نادرست

دلیل: فاصله اقلیدسی نسبت به ضرب در یک عدد مثبت (Scaling) همگن است. اگر تمام مختصات در عدد 0.5 ضرب شوند، تمام فاصله‌های بین نقاط نیز دقیقاً در 0.5 ضرب می‌شوند:

$$d_{new}(x,y)=\sqrt{\sum(0.5x_i-0.5y_i)^2}=0.5\times d_{old}(x,y)$$

چون تمام فاصله‌ها به یک نسبت کوچک می‌شوند، «ترتیب» (Ordering) همسایه‌ها تغییر نمی‌کند. یعنی نزدیک‌ترین همسایه قبل از تغییر مقیاس، همچنان نزدیک‌ترین همسایه باقی می‌ماند و نتیجه دسته‌بندی تغییر نخواهد کرد.

د- رابطه افزایش k و صحت دسته‌بندی

پاسخ: نادرست

دلیل: افزایش k همیشه باعث بهبود دقت نمی‌شود.

- وقتی k خیلی کوچک است (مثلاً $k=1$)، مدل دچار بیش‌برازش (Overfitting) می‌شود و نسبت به نویز حساس است.

- وقتی k خیلی بزرگ می‌شود (به سمت N میل می‌کند)، مدل دچار کم‌برازش (Underfitting) می‌شود و مرزهای تصمیم‌گیری بیش از حد هموار می‌شوند. در حالت $k = N$، مدل صرفاً کلاسی را انتخاب می‌کند که بیشترین تعداد کل نمونه‌ها را دارد (کلاس اکثریت)، صرف‌نظر از اینکه ورودی تست کجاست. بنابراین معمولاً یک k بهینه در میانه وجود دارد و نمودار خطا U شکل است.