درس : یادگیری ماشین
دانشجو : امیرمحمد خرازی
شماره دانشجویی : ۴۰۱۵۲۵۲۱۰۰۲
استاد درس : دکتر منصور رزقی آهق

تمرین کلاسی سری اول | دانشکده علوم ریاضی ، گروه علوم کامپیوتر، گرایش داده‌کاوی

گیت‌هاب درس (لینک) | گیت‌هاب این تمرین (لینک)

پرسش ۱ :

فرض کردیم مربعی به اندازه ضلع a و دایره این با قطر a یا شعاع $\frac{a}{2}$ ، محاط در این دایره ، در اختیار داریم. مساحت و حجم این مربع و دایره بصورت زیر خواهد بود :

$$
\begin{aligned}
&V_{circle} = S = \pi R^2 \longleftrightarrow \frac{\pi}{4}a^2 \\
&V_{Square} = S = a^2 \\
&V_{occupied} = V_{circle} = \frac{\pi}{4}a^2 \\
&V_{empty} = V_{Square} - V_{occupied} = a^2 - \pi\frac{a^2}{4} = a^2(1-\frac{\pi}{4})
\end{aligned}
$$

این برای زمانی است که در دو بعد هستیم. حالا ابعاد را یک واحد افزایش می‌دهیم. دایره به کره و مربع به مکعب تبدیل می‌شوند.

$$
\begin{aligned}
&V_{Sphere} = S = \frac{4}{3}\pi R^3 \longleftrightarrow \frac{4}{3}\pi\frac{a^3}{8} = \frac{\pi}{6}a^3 \\
&V_{Cube} = S = a^3 \\
&V_{occupied} = V_{Sphere} = \frac{\pi}{6}a^3 \\
&V_{empty} = V_{Cube} - V_{occupied} = a^3 - \frac{\pi}{6}a^3 = a^3(1-\frac{\pi}{6})
\end{aligned}
$$

با همین دو حالت می‌توان حدس زد که اگر ابعاد را اضافه کنیم به چه عددی همگرا می‌شوند. یعنی می‌بینیم که $V_{occupied}$ به سفر متمایل می‌شود و در نتیجه V_{empty} ، که فضای خالی است، به $V_{nD-Cube}$ متمایل می‌شود. پس با افزایش ابعداد، فاضی اشغال شده به صفر و فضای خالی به فضای مربع در ابعاد بالاتر همگرا می‌شود.
اما برای اثبات این موضوع باید، حجم دایره یا کره در ابعاد بالاتر را بدست آوریم. برای این منظور از فرمول‌های مربوط به n-ball استفاده می‌کنیم. بنابر ویکی‌پدیا ،

حجم یک کره n بعدی بصورت زیر محاسبه می‌شود :

$$V_{2K}(R) = \frac{\pi^k}{k!}R^{2k}$$
$$V_{2k+1}(R) = \frac{2(k!)(4\pi)^k}{(2k+1)!}R^{2k+1}$$

و می‌دانیم که برای عدد حقیقی n ، رابطه $n! \geq c^n$ که c یک عدد ثابت است ، برقرار است.
با این تفاسیر اگر از روابط بالا حد بگیریم :

$$\lim_{2k\to\infty}(\frac{a}{2}) = \lim_{2k\to\infty}\frac{\pi^k}{k!}(\frac{a}{2})^{2k}$$
$$\Rightarrow \pi^k \leq k!$$
$$\Rightarrow \lim_{2k\to\infty}\frac{\pi^k}{k!} = 0$$
$$\Rightarrow \lim_{2k\to\infty}(\frac{a}{2}) \approx 0 \times (\frac{a}{2})^{2k} \approx 0$$
$$\Rightarrow V_{occupied}^{(n\to\infty)} \approx 0$$

همچنین برای زمانی که n فرد باشد، داریم :

$$\lim_{2k+1\to\infty}(\frac{a}{2}) = \lim_{2k+1\to\infty}\frac{2(k!)(4\pi)^k}{(2k+1)!}(\frac{a}{2})^{2k+1}$$
$$\Rightarrow 2(k!)(4\pi)^k \leq (2k+1)!$$
$$\Rightarrow \lim_{2k+1\to\infty}\frac{2(k!)(4\pi)^k}{(2k+1)!} = 0$$
$$\Rightarrow \lim_{2k+1\to\infty}(\frac{a}{2}) \approx 0 \times (\frac{a}{2})^{2k+1} \approx 0$$
$$\Rightarrow V_{occupied}^{(n\to\infty)} \approx 0$$

در نتیجه می‌بینیم که این حجم با افزایش ابعاد به صفر همگرا می‌شود.
در دو حالت دو بعدی و سه بعدی که بصورت دستی حساب کردیم، نیز این عبارت قابل حدس زدن بود یعنی ابتدا $(1-\frac{\pi}{4})$ داشتیم، سپس $(1-\frac{\pi}{6})$ شد و به نظر می‌آمد این عبارت $\frac{\pi^{\square}}{\square}$ در نهایت صفر شود و لذا حجم خالی بصورت $V_{empty} = V_{nD-Cube} \times (1-0)$ همگرا شود (که یعنی حجم اشغال شده به صفر همگرا شده است و حجم خالی تمام حجم مکعب n بعدی است).

پرسش ۲ :

در فایل کد ارائه شده بطور کامل شرح داده شده است.
تحلیل خروجی سه پرسش : با افزایش ابعاد، درصد نقاطی که در هر سلول قرار می‌گیرند کاهش می‌یابد. دلیل آن شاید این

می‌تواند باشد که با افزایش بعد، هر بار ویژگی جدیدی اضافه می‌شود که می‌تواند نمونه را از نمونه دیگر جدا کند (که این جدا سازی را با سلول ها انجام دادیم). مثلا زمانی که یک بعد داشتیم، تنها مقدار x برای مشخص شدن تعلق نمونه به سلول مورد نظر، کافی بود. با اضافه شدن بعد دوم، مقادیر x و y را برای مشخص کردن جای نمونه، نیاز داریم. با افزایش بیشتر ابعاد نیز، به همین شکل نیاز به استفاده از مقادیر بیشتری برای مشخص کردن محل نمونه خواهیم داشت که در نتیجه با افزایش بیشتر ابعاد، هر سلول احتمالا می‌تواند شامل حداکثر یک نمونه باشد (۱ یا ۰ نمونه).

بطور کلی در حالت ۱ بعدی ۳ سلول و ۵۰ داده داریم لذا حدودا ۱۶ داده در هر سلول قرار می‌گیرد. در حالت ۲ بعدی ۲۵ سلول داریم لذا ۲ داده در هر سلول قرار می‌گیرد و در حالت ۳ بعدی ۱۲۵ سلول داریم که یعنی ۵۰ سلول، شامل ۱ داده هستند و ۷۵ سلول دیگر خالی‌اند.

این مشکل را می‌توان از روش‌هایی چون PCA ی غیره و همچنین اضافه کردن داده، تا حدی برطرف کرد.

پرسش ۳ :

در فایل کد ارائه شده، بطور کامل شرح داده شده است.

برای بدست آوردن نقاط دایره، از مختصات قطبی استفاده می‌کنیم. که بطور خلاصه داریم :

$$x = r\cos\theta \qquad y = r\sin\theta$$

که در اینجا داریم:

$$r = \sqrt{x^2 + y^2} \qquad \theta = \tan^- 1(\frac{y}{x})$$

و لذا می‌توان با داشتن یکی از مختصات، مختصات دیگر را بدست آورد. در کد هم همین کار را انجام داده‌ایم. یعنی مختصات قطبی رندم را بدست آوردیم و سپس آن‌را به مختصات معمولی بردیم.

ادامه سوال، بسیار شبیه به سوال ۱ است. در سوال ۱ این موارد را بررسی کردیم و دیدیم که به صفر میل می‌کرد.

منتها در اینجا، شعاع دایره برابر ۱ در نظر گرفته شده است که در نتیجه، ضلع مربع ۲ واحد خواهد بود. همچنین در کره یا سایر ابعاد، این رابطه برقرار است.

کاری که در نهایت برای این سوال انجام شده است، بدین صورت است که N داده بصورت تصادفی در بعدی مشخص، تولید شده است، چنانچه داخل دایره قرار داشت، سبز و در غیر اینصورت قرمز نمایش داده شده است. برای بررسی نسبت حجم مربع به دایره، تعداد این نقط شمرده شده است. این کار را برای همه ابعادی که گفته شده است، انجام داده‌ایم.

بعدا مشاهده کردیم که (در جایی که نمودار فراوانی به نسبت ابعاد کشیدیم) با افزایش ابعاد، نقاط کمتری داخل این دایره یا کره قرار می‌گیرند و کم کم به صفر میل می‌کند.

پرسش ۴ :

در فایل کد ارائه شده، بطور کامل شرح داده شده است.

تغییر روش محاسبه فصله و شباهت، می‌تواند روی موضوع ابعاد بزرگ تاثیر داشته باشد. بدیهی است که با افزایش ابعاد، مثلا $x = (x_1, x_2, \ldots, x_D)$ ، برای محاسبه هر فاصله در هر نقطه، باید تعداد زیادی عملیات انجام دهیم که چنانچه متر مورد

استفاده بگونه ای باشد که عملیات را کاهش دهد، درمانی موقت برای معضل ابعاد خواهد بود.
و به نظر می‌رسد که متر cos از بقیه بهتر عمل می‌کند. چون اگر به فاصله های بدست مده نگاه کنیم، مخصوص در زمانی که ابعاد خیلی بالا هستند، این معیار بهتر عمل کرده است.
قبلا در این مورد که چه راه حل‌هایی برای حل معضل ابعاد داریم، صحبت کرده‌ام.

پرسش ۵ :

مثال اصلی این بخش ، در مورد مسئله ابعاد است که در سوالات قبل به آن اشاره کرده‌ام و واقعا چیز خاصی بخاطر ندارم. تنها در همین مقدار که با افزایش ابعاد ، سلول‌های ما افزایش پیدا می‌کند (مانند سوال ۲ که کد آن را نوشتم) و این افزایش بصورت نمایی است.
همان نمودار، مربع و مکعب ها در کلاس بحث شده بود.
پیش از آن در مورد سلول‌ها صحبت کرده، بعد از یک رگرسیون زده برای D متغییر ورودی (تا درجه ۳ رفته جلو).
سپس مثالی از یک کره در فضای D بعدی زده و گفته که مثلا اگر شعاع آن را r بگیریم و ϵ یک عدد کوچک باشد، چند درصد داده بین r و $r-\epsilon$ قرار دارند (که در اینجا r ر ۱ گرفته است).
فرمول آن را محاسبه کرده و سپس برای D های متفاوت این تناسب را بدست آورده و نتیجه گرفته که در D های بالاتر، این مقدار به ۱ نزدیک می‌شود، یعنی همه در روی پوسته کره قرار دارند (حتی اگر ϵ را نیز خیلی کوچک کنیم).

پرسش ۶ :

تمرین 1.15 و 1.16 را به این لینک ارجاع می‌دهم :

bishop_solutions.

صفحات ۱۰ و ۱۱ و ۱۲ مربوط به سوال ۱۵ و صفحات ۱۳ و ۱۴ مربوط به سوال ۱۶ است.