بسمه تعالی

بازشناسی آماری الگو

دانشگاه صنعتی خواجه نصیرالدین طوسی — تمرین سری ششم — تاریخ تحویل: ۱۴۰۴/۹/۱ — آزمایشگاه بینایی ماشینی و پردازش تصاویر پزشکی

۱- (اختیاری) دو کلاس ω_1 و ω_2 در فضای دو بعدی در نظر بگیرید. کلاس ω_1 دارای توزیع یکنواخت درون یک دایره به شعاع r است. کلاس ω_2 نیز دارای توزیع یکنواخت درون یک دایره به شعاع r است. فاصله بین مراکز دو دایره بزرگتر از $4r$ است. چنانچه N نمونه آموزشی در اختیار داشته باشیم نشان دهید بازای $k \geq 3$ خطای کلاسیفایر 1NN از kNN کوچکتر است.

۲- (اختیاری) مساله امتحان میان ترم بازشناسی آماری الگو سال ۱۳۸۲

مقایسه خطای متوسط (حدی) روش 1NN (P_e^1) با خطای متوسط روش بیز (P_e) نشان می‌دهد:

$$\begin{cases} P_e^1 = P_e & P_e = 0 \\ P_e^1 = P_e & P_e = \dfrac{M-1}{M} \\ P_e^1 > P_e & \text{در غیر اینصورت} \end{cases}$$

مساله زیر نشان می‌دهد که امکان تساوی دو خطای متوسط روش بیز و روش 1NN در حالتی که خطای بیز بین دو مقدار حدی ذکر شده در بالا قرار دارد وجود دارد.

الف- در فضای یک بعدی در صورتیکه داشته باشیم $P(\omega_i)=1/M$ $i=1,\dots,M$ و:

$$p(x|\omega_i) = \begin{cases} 1 & 0 \leq x \leq \dfrac{M}{M-1}r \\ 1 & i \leq x \leq i+1-\dfrac{M}{M-1}r \\ 0 & \text{در غیر اینصورت} \end{cases}$$

مطلوب است رسم منحنی‌های زیر در حالت یک مساله با سه کلاس و بازای $r=0.1$:

$$P(\omega_i|x) \quad i = 1,2,3$$

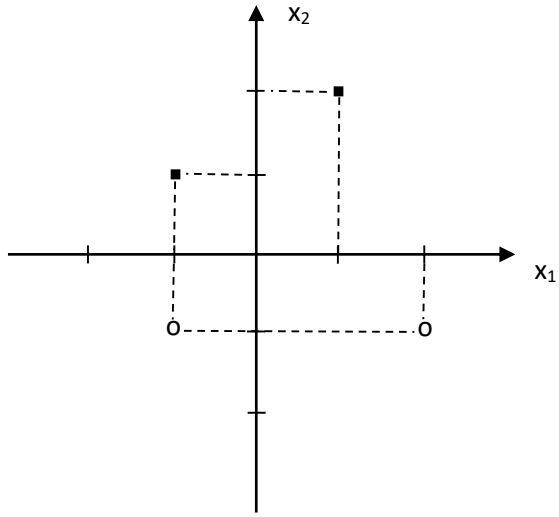
ب- نشان دهید که خطای متوسط بیز در این حالت از رابطه زیر محاسبه می گردد:

$$P_e=r$$

ج- ثابت کنید در این حالت خطای متوسط روش 1NN و خطای بیز برابرند:

$$P_e=P_e^1$$

۳- (اختیاری) در شکل زیر تعدادی نمونه در فضای دوبعدی از دو کلاس ▪ و o داده شده است. مطلوب است رسم کلاسیفایر 1NN.



۴- (اختیاری) سوال امتحانی میان ترم ۱۳۹۱

یکی از مشکلات روش kNN ضرورت محاسبه و مرتب کردن N فاصله بر اساس یک معیار فاصله دقیق مثل فاصله اقلیدسی (d_{E}) و سپس انتخاب k نزدیکترین همسایه‌ی نمونه‌ی مجهول $\mathbf{x}$ است. دیدیم که کیتلر (Kittler) برای کاهش این هزینه، روشی را با استفاده از فاصله مانهاتان (d_{C}) به صورت زیر پیشنهاد کرد:

- ابتدا k-امین نزدیکترین همسایه‌ی $\mathbf{x}$ را بر حسب فاصله مانهاتان پیدا کنید ($\mathbf{x}_k$).
- فاصله اقلیدسی آن را از $\mathbf{x}$ محاسبه کنید ($d_{\mathrm{E}}(\mathbf{x},\mathbf{x}_k)$).
- در مجموعه Y که معمولا تعداد عناصر آن به مراتب از N کوچکتر است، k نزدیکترین همسایه‌ی $\mathbf{x}$ را بر حسب فاصله اقلیدسی جستجو کنید:

$$Y = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbf{X}, d_{\mathrm{C}}(\mathbf{x},\mathbf{x}_i) \le \sqrt{n} d_{\mathrm{E}}(\mathbf{x},\mathbf{x}_k)\}$$

ملاحظه می‌شود که روش فوق می تواند تعداد محاسبات دقیق مورد نیاز را به مراتب کاهش دهد.

الف- بدیهی است برای تضمین درستی عملکرد روش کیتلر، مجموعه Y باید لزوما در برگیرنده k نزدیکترین همسایه‌ی $\mathbf{x}$ بر حسب فاصله اقلیدسی باشد. با رسم یک شکل در فضای دوبعدی و ارائه توضیحات نشان دهید که مجموعه Y لزوما در برگیرنده k نزدیکترین همسایه‌ی $\mathbf{x}$ بر حسب فاصله اقلیدسی هست.

ب- روشی مشابه روش کیتلر ولی اینبار بر اساس فاصله ماکزیمم (d_{M}) بنا کنید و مراحل آن را مشابه با الگوریتم کیتلر بیان نمایید.

$$d_C(\mathbf{y},\mathbf{z}) = \sum_{j=1}^{n} |\mathrm{y_j} - z_j|$$
$$d_M(\mathbf{y},\mathbf{z}) = \max_{j=1,\dots,n} |\mathrm{y_j} - z_j|$$

ج- بدیهی است روش پیشنهادی شما مجموعه‌ی دیگری مثل Y' را برای جستجوی k نزدیکترین همسایه‌ی $\mathbf{x}$ بر حسب فاصله اقلیدسی پیشنهاد می‌کند که هم اندازه مجموعه‌ی X است. با رسم شکل در فضای دوبعدی و ارائه توضیحات نشان دهید که مجموعه Y' محدوده ای هم اندازه ولی متفاوت از فضای ویژگی را در مقایسه با مجموعه Y اشغال می‌کند.

د- با ترکیب روش کیتلر و روش خودتان، روش سومی را پیشنهاد کنید که با استفاده ترکیبی از دو فاصله مانهاتان و ماکزیمم، مجموعه Y'' را که از هر دو Y و Y' کوچکتر است برای جستجوی k نزدیکترین همسایه‌ی $\mathbf{x}$ بر حسب فاصله اقلیدسی در اختار قرار می دهد. با رسم شکل در فضای دوبعدی نشان دهید که مجموعه Y'' سطح اشغال کوچکتری در فضای ویژگی نسبت Y و Y' دارد.

۵- (اختیاری) در این مساله می‌خواهیم مساله آفت ابعاد (curse of dimensionality) یعنی مشکلات ناشی از فضاهای ویژگی با ابعاد بالا را مثلا در یک دسته‌بند نزدیکترین همسایه بررسی کنیم. فرض کنید می‌خواهیم تابع چگالی احتمال $p(\mathbf{x})$ را در یک فوق مکعب به اضلاع واحد در فضای R^n بر اساس N نمونه تخمین بزنیم. اگر $p(\mathbf{x})$ شکلی پیچیده داشته باشد، ناچاریم تعداد زیادی نمونه را برای رسیدن به یک تخمین قابل قبول بکار ببریم.

الف- اگر $\mathrm{N_1}$ تعداد نمونه‌های قابل قبول برای تخمین چگالی در فضای R^1 باشد، تعداد نمونه‌های لازم برای تخمین همان چگالی در فضای R^n چقدر است؟ اگر $\mathrm{N_1}=100$ باشد، چه تعداد نمونه در فضای ۲۰ بعدی مورد نیاز خواهد بود؟

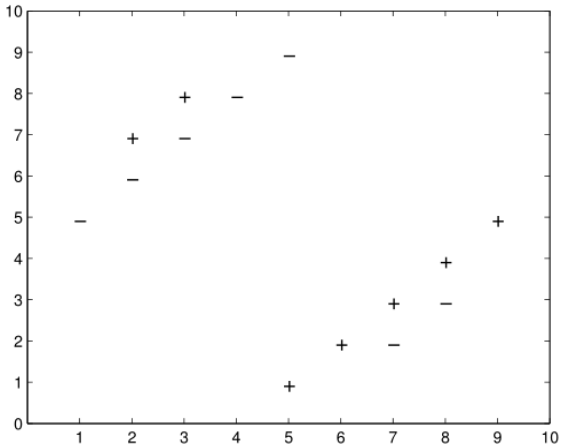
ب- نشان دهید فاصله بین نقاط در فضای R^n زیاد و تقریبا برابر است و برای یافتن حتی تعداد محدودی همسایه یک نقطه، شعاع همسایگی زیادی مورد نیاز است. (فرض کنید N محدود است و نمونه‌ها بصورت یکنواخت در داخل فوق مکعب به ضلع واحد توزیع شده‌اند.)

ج- رابطه محاسبه $l_n(P)$ یعنی طول ضلع یک فوق مکعب در فضای n بعدی که حاوی کسر P از تمام نقاط باشد را بدست آورید ($0<P<1$). مثلا طول یک ضلع فوق مکعب که حاوی 0.01 یا 0.1 کل نقاط در فضای 5 و 20 بعدی است چقدر می‌باشد. (فرض کنید نمونه‌ها بصورت یکنواخت در داخل فوق مکعب به ضلع واحد توزیع شده‌اند.)

د- نشان دهید که تقریبا همه نقاط به یکی از اضلاع فوق مکعب به ضلع واحد در فضای n بعدی نزدیک هستند. به این منظور فاصله ماکزیمم (d_M) یک نقطه را به نزدیکترین همسایه آن محاسبه کنید. این موضوع نشان می‌دهد که تقریبا فاصله همه نقاط از یکی از اضلاع فوق مکعب واحد کمتر از فاصله بین آنها است.

ه- از قوه تصور خود کمک بگیرید و با کمک نتایج فرضهای (ج) و (د) ذهن خود را به فضای ویژگی با ابعاد بالا ببرید. چه نتیجه‌ای می‌توانید در خصوص توزیع نمونه‌های یک توزیع (حتی غیر یکنواخت) در حجم داخل یک فوق مکعب (در گوشه‌ها و در مرکز آن) بگیرید؟ این توزیع حجم داخل یک مکعب در فضای با ابعاد بالا چرا به عنوان یک آفت در مسایل بازشناسی الگو مطرح می‌شود و به چه دلیل می‌تواند موجب افت کارایی یک دسته‌بند شود؟

۶- (اختیاری) مجموعه داده دو بعدی زیر را در نظر بگیرید:



الف. داده جدید $\begin{bmatrix}2\\8\end{bmatrix}$ از راه می رسد. کلاس آن را به روش 3NN مشخص نمایید.

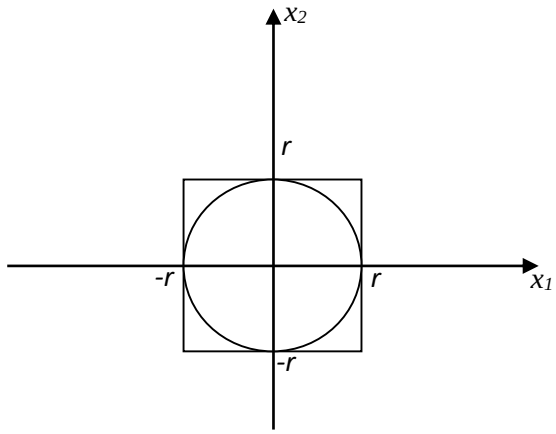
ب. اگر برای دسته‌بندی از 7NN استفاده شود کلاس آن کدام خواهد بود؟

ج. مرز 1NN را برای دو کلاس رسم کنید.

د. توضیح دهید انتخاب مقادیر خیلی کوچک و خیلی بزرگ برای k چه مشکلی در دسته بندی با kNN ایجاد میکند. شما چه مقداری را برای k به منظور دسته‌بندی داده‌های فوق با روش kNN مناسب می‌دانید؟

۷- مساله امتحان میان ترم ۱۳۹۶

در این مساله می‌خواهیم مساله آفت ابعاد (curse of dimensionality) را تشریح نماییم و نشان دهیم چرا با افزایش ابعاد فضا، نمونه‌ها در فضای ویژگی بطور عمده گوشه‌ها را اشغال می‌کنند. شکل زیر در فضای دو بعدی یک دایره به شعاع r و یک مربع به ضلع $2r$ را بصورت هم مرکز نشان می‌دهد.



الف- با فرض اینکه نمونه‌ها در فضای ویژگی دارای توزیع یکنواخت باشند، نسبت مساحت دایره (V_c) به مساحت مربع (V_s) می‌تواند به عنوان احتمال قرارگیری نمونه‌ها در محدوده به فاصله r از میانگین (مبدا در شکل مقابل) در نظر گرفته شود. این نسبت را محاسبه نمایید.

$$\frac{V_c}{V_s} = ?$$

ب- نسبت فاصله میانگین (مبدا در شکل مقابل) از یک ضلع مربع به فاصله آن از گوشه مربع چقدر است؟

ج- می‌خواهیم سوال فرض الف را در فضای ویژگی n بعدی مطرح کنیم. چنانچه رابطه محاسبه حجم یک فوق کره به شعاع r در فضای n بعدی بصورت زیر باشد:

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

مطلوب است محاسبه نسبت حجم فوق کره به شعاع r به حجم فوق مکعب به ضلع $2r$ (V_s) هر دو به مرکز مبدا مختصات.

د- از تقریب زیر استفاده کنید و نشان دهید هنگامیکه ابعاد فضای ویژگی به سمت بینهایت میل کند، نسبت حجم فوق کره به شعاع r به فوق مکعب به ضلع $2r$ بدست آمده در بند قبل به سمت صفر میل خواهد کرد.

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

ه- نسبت فاصله میانگین از یک وجه فوق مکعب به فاصله آن از یک گوشه فوق مکعب را بدست آورید و نشان دهید با افزایش n این نسبت به سمت صفر میل می‌کند.

و- از قوه تصور خود کمک بگیرید و با کمک نتایج فرضهای (د) و (ه) ذهن خود را به فضای ویژگی با ابعاد بالا ببرید. چه نتیجه‌ای می‌توانید در خصوص توزیع نمونه‌های یک توزیع (حتی غیر یکنواخت) در حجم داخل یک فوق مکعب (در گوشه‌ها و در مرکز آن) بگیرید؟ این توزیع حجم داخل یک مکعب در فضای با ابعاد بالا چرا به عنوان یک آفت در مسایل بازشناسی الگو مطرح می‌شود و به چه دلیل می‌تواند موجب افت کارایی یک دسته‌بند شود؟

۸- (اختیاری) سوال امتحانی میان ترم ۹۸

در فضای دو بعدی و در حالت مساله دسته‌بندی دو کلاسه فرض کنید ۵۰۰۰ نمونه آموزشی از هر یک از دو کلاس در اختیار است.

الف- چنانچه هیچگونه اطلاعی از توزیع کلاسها در اختیار نباشد از میان کلاسیفایرهای بیز، پارزن و kNN کدامیک را ترجیح می دهید و چرا؟

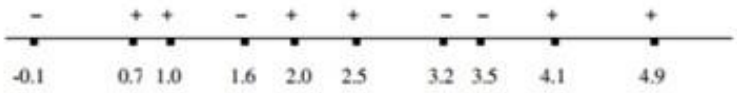
ب- چنانچه در فضای ۵۰ بعدی و در حالت دوکلاسه فقط ۴۰ نمونه آموزشی از هر کلاس داشته باشیم از میان کلاسیفایرهای بیز، kNN و پارزن کدام را ترجیح می دهید و چرا؟

ج- کدام روش دسته‌بندی غیرپارامتری (یعنی پارزن یا kNN) مستقیما احتمالات پسین کلاسها را تخمین می‌زند؟ چرا؟

۹- (اختیاری) جدول زیر مقادیر نمونه‌ها در فضای یک بعدی و برچسب هریک (+/-) را نشان می‌دهد. در سوالهای زیر فرض می شود که روش kNN با فاصله اقلیدسی برای پیش‌بینی برچسب نمونه x استفاده می شود.

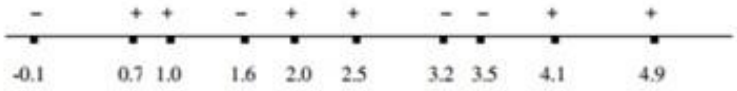
X	Y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+

الف- چنانچه از روش 1NN برای دسته‌بندی نمونه‌ها استفاده شود، و بخواهیم خود نمونه‌ها را دسته‌بندی کنیم، در نمودار زیر نمونه‌هایی که به اشتباه دسته‌بندی می‌شوند را با رسم دایره ای دور آنها مشخص کنید.



ب- مطلوب است محاسبه خطای دسته‌بندی. این روش ارزیابی دسته بندی 1NN را Leave-one-out cross validation method نامند.

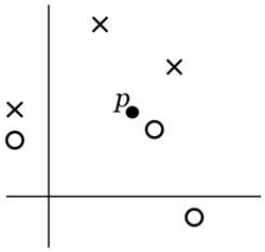
ج- بند الف را با روش 3NN انجام دهید و نتیجه را روی نمودار زیر نشان دهید.



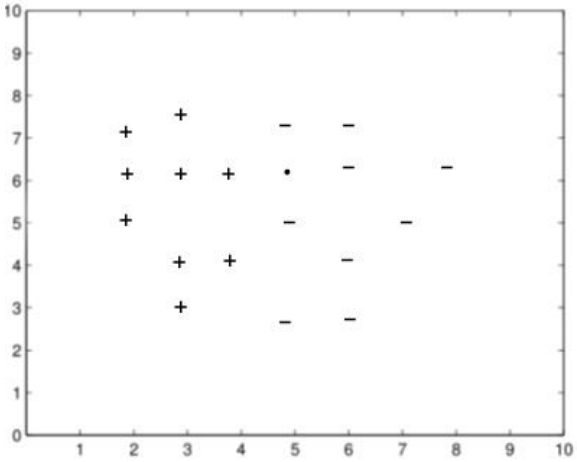
د- بند ب را با روش 3NN انجام دهید

۱۰- (اختیاری) (سوال امتحانی میان ترم ۹۹) می خواهیم با استفاده از 1NN نقطه p را در یکی از دسته های X یا O قرار دهیم. اگر از فاصله کسینوسی استفاده کنیم پیش بینی چه خواهد بود؟ (فاصله کسینوسی به شکل کسینوس زاویه بین دو بردار تعریف می شود و نشان می دهد که تا چه میزان دو بردار هم جهتند)

اگر از فاصله اقلیدسی استفاده کنیم چطور؟



۱۱- (سوال امتحانی میان ترم ۱۴۰۰) شکل رو به رو را در نظر بگیرید:



برای شکل رو به رو می‌خواهیم برای این که اثر مقیاس را از بین ببریم به هر یک از محورها که یک ویژگی را برای هر نمونه برای ما مشخص می‌کند وزنی بدهیم که وزن‌های ما $w_1=2/3$، $w_2=1/3$ است .

الف) با بیان دلیل خود بگوید که کدام وزن را برای کدام ویژگی مناسب می بینید.

ب) در نقطه ای که با علامت نقطه مشخص شده است مشخص کنید که طبق دسته بند KNN به ازای k=5 به کدام کلاس تعلق می‌گیرد برای این کار باید هم وزن‌دهی ویژگی که در بخش الف مطرح شده را در نظر بگیرید و هم وزن‌دهی فاصله‌ای به فرمول وزن‌دهی به فاصله به صورت زیر که در آن x_u نقطه مد نظر ما و x_k نزدیک‌ترین سمبل‌های ما هستند.

$$w_k = \frac{1}{d(x_u - x_k)}$$

۱۲- (سوال امتحانی میان ترم ۱۴۰۲) در یک دسته‌بند دو کلاسه 1NN علی‌رغم وجود نمونه‌ها از دو کلاس مثبت و منفی، آیا این امکان وجود دارد که همه نمونه‌های تست به دسته مثبت اختصاص یابند؟ اگر پاسخ مثبت است، با یک مثال این امکان را نشان دهید.

X_1	X_2	Y
1	7	+1
3	3	+1
1	1	-1
5	4	-1
2	5	-1

۱۳- (میان ترم ۱۴۰۳) جدول مقابل ۵ نمونه آموزشی در فضای دوبعدی به همراه برچسب هر کدام را نشان می دهد.

الف- نمونه (3,6) را با دسته بند 1NN و فاصله منهتن (City Block) دسته بندی کنید.

یادآوری: dC((u,v),(p,q))=|u-p|+|v-q|

ب- (درست یا نادرست) مرز تصمیم‌گیری با قاعده 1NN بر اساس ۵ نمونه داده شده در جدول مقابل دنباله ای از پاره خط‌های موازی با محورهای X1 یا X2 است.

ج- (درست یا نادرست) نتیجه یک دسته بند عمومی kNN که از فاصله اقلیدسی استفاده می کند، هنگامیکه مختصات نمونه های آموزشی در ۰,۵ ضرب شوند تغییر می کند.

د- (درست یا نادرست) با افزایش مقدار k در دسته بند kNN از ۱ به تعداد کل نمونه ها یعنی N، صحت دسته بندی همواره افزایش می یابد.

موفق باشید