



آمار محاسباتی و آماری در مهندسی

بسمه تعالی

بازشناسی آماری الگو



دانشگاه صنعتی خواجه نصیرالدین طوسی

تمرین سری پنجم

تاریخ تحویل:

۱۴۰۴/۸/۲۴

۱- (اختیاری) داده‌های $x_1, x_2, \dots, x_n \sim p$ را در نظر بگیرید. \hat{f}_n تخمین پارزن آنها با استفاده از کرنل زیر است:

$$K(x) = \begin{cases} 1 & -\frac{1}{2} < x < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

دو مورد زیر را ثابت کنید:

$$E(\hat{f}(x)) = \frac{1}{h} \int_{x-(h/2)}^{x+(h/2)} f(y) dy \quad V(\hat{f}(x)) = \frac{1}{nh^2} \left[\int_{x-(h/2)}^{x+(h/2)} f(y) dy - \left(\int_{x-(h/2)}^{x+(h/2)} f(y) dy \right)^2 \right]$$

۲- (اختیاری) فرض کنید $p(x) = N(\mu, \sigma^2)$ و کرنل پارزن $k(x) = N(0, 1)$ نشان دهید که تخمینگر پارزن تابع چگالی احتمال:

$$\hat{p}_N(x) = \frac{1}{Nh_N} \sum_{i=1}^N k\left(\frac{x - x_i}{h_N}\right)$$

بازای مقادیر کوچک h_N دارای خواص زیر است:

$$E\{\hat{p}_N(x)\} = N(\mu, \sigma^2 + h_N^2) \quad \text{الف}$$

$$Var\{\hat{p}_N(x)\} \cong \frac{1}{2Nh_N\sqrt{\pi}} p(x) \quad \text{ب}$$

۳- (اختیاری) مساله امتحان میان ترم بازشناسی آماری الگو سال ۱۳۹۰

در یک مساله دو کلاسه فرض کنید یک مجموعه از بردارهای تصادفی از هر کلاس بصورت زیر در اختیار است:

$$X = \{\mathbf{x}_{11}, \dots, \mathbf{x}_{1N_1}, \mathbf{x}_{21}, \dots, \mathbf{x}_{2N_2}\}$$

که در آن \mathbf{x}_{jk} به معنی بردار (نمونه) k -ام از کلاس j -ام می‌باشد ($j=1, 2$). بردارهای تصادفی در فضای n بعدی تعریف شده‌اند و عناصر هر بردار فقط می‌توانند مقادیر $\{1, -1\}$ اختیار کنند. بعنوان مثال بازای $n=3$ یک نمونه تصادفی می‌تواند به فرم $\mathbf{x}_{jk}^t = [1 \quad 1 \quad -1]$ باشد.

الف- با انتخاب یک کرنل مناسب، رابطه‌ای بنویسید که بتوان با استفاده از آن چگالی احتمال هر کلاس را با استفاده از نمونه‌های آن و روش پارزن تخمین زد.

ب- چنانچه احتمال کلاسها برابر باشد ($P(\omega_1) = P(\omega_2)$)، قاعده دسته‌بندی بیز را برای بدست آوردن حداقل متوسط خطا و بر اساس تخمین توابع چگالی احتمال (با استفاده از روش پارزن) بیان کنید.

ج- چنانچه از کرنل زیر برای تخمین تابع چگالی احتمال به روش پارزن استفاده کنیم:

$$K(\mathbf{x}, \mathbf{x}_k) = \mathbf{x}^t \mathbf{x}_k$$

نشان دهید کلاسیفایر حاصل از بند (ب) از نوع کلاسیفایر نزدیکترین میانگین خواهد بود. (یعنی نمونه‌ها بر اساس فاصله آنها از میانگین کلاسها دسته‌بندی خواهند شد.)

۴- (اختیاری) فرض کنید $p(x)$ توزیع یکنواخت در فاصله $[0, a]$ باشد و کرنل پارزن بصورت زیر تعریف شود:

$$k(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

الف- نشان دهید:

$$E\{\hat{p}_N(x)\} = \begin{cases} 0 & x < 0 \\ \frac{1}{a}(1 - e^{-x/h_N}) & 0 \leq x \leq a \\ \frac{1}{a}(e^{a/h_N} - 1)e^{-x/h_N} & x \geq a \end{cases}$$

ب- مطلوب است رسم $E\{\hat{p}_N(x)\}$ بر حسب x بازای $a=1$ و $h_N=1, 1/4$

ج- برای آنکه بایاس در ۹۹٪ فاصله $[0, a]$ کمتر از یک درصد باشد مقدار h_N چقدر اختیار شود؟

۵- (اختیاری) حسگری در اختیار داریم که سطح روشنایی x_i را گزارش می‌کند. می‌خواهیم با استفاده از اطلاعات این حسگر تشخیص دهیم که آیا امروز هوا آفتابی (کلاس ω_1) است یا ابری (کلاس ω_2). برای این منظور از تخمین پارزن با کرنل مستطیلی شکل به صورت زیر استفاده می‌نماییم.

$$\phi\left(\frac{x}{h}\right) = \begin{cases} 1 & |x| < \frac{h}{2} \\ 0 & \text{otherwise} \end{cases} \quad \hat{p}(x) = \frac{1}{nV} \sum_{i=1}^n \phi\left(\frac{x-x_i}{h}\right)$$

اطلاعات پنج روز به ما داده شده است. سه روز با سطوح روشنایی ۴، ۱ و ۵ آفتابی بوده و دو روز با سطوح روشنایی ۳ و ۲ ابری گزارش شده‌اند.

الف. توابع تخمین پارزن را برای $\hat{p}(x|\omega_1)$ و $\hat{p}(x|\omega_2)$ با در نظر گرفتن $h = 1$ رسم کنید.

ب. توابع پسین $\hat{P}(\omega_1|x)$ و $\hat{P}(\omega_2|x)$ را برای همه مقادیر تعریف شده در x رسم کنید. برای تخمین احتمال کلاس‌ها از تخمین بیشینه درست‌نمایی استفاده کنید.

۶- (اختیاری) سوال امتحانی میان ترم ۱۳۹۷

الف- در فضای یک بعدی دو نمونه تصادفی $x_1=1$ و $x_2=2$ از یک توزیع مشاهده شده است. اگر $K(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$ و h_N را برابر با ۲ در نظر بگیریم، $p(1.6)$ چقدر است؟

ب- اگر توزیع نمونه‌ها از نوع یکنواخت در فاصله $x \in [0, 4]$ باشد مطلوب است رسم تقریبی امید ریاضی تخمین پارزن توزیع $\hat{p}(x)$ بر اساس کرنل معرفی شده در بند الف.

ج- افزایش h_N چه تاثیری در تخمین چگالی احتمال دارد؟ کاهش h_N چه تاثیری در تخمین چگالی احتمال دارد؟

۷- یک مجموعه از نمونه‌ها در فضای یک بعدی در اختیار است:

$$X = \{0, 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 4, 5\}$$

با استفاده از مباحث مربوط به تخمین پارامتری و غیرپارامتری (پارزن) یک چگالی احتمال به سوالات زیر پاسخ دهید:

الف- هیستوگرام داده‌ها را بر اساس پهنای بین ۱ و فرض اینکه مراکز بینها روی مجموعه نقاط $\{0, 1, 2, 3, 4, 5\}$ قرار داشته باشند رسم کنید.

ب- (اختیاری) رابطه محاسبه تخمین چگالی احتمال را بر حسب یک کرنل دلخواه K بنویسید.

ج- یک کرنل مثلثی بصورت زیر انتخاب کنید:

$$K(u) = (1 - |u|)\delta(|u| \leq 1)$$

که در آن u تابعی از فاصله نمونه x_i تا نمونه مورد سوال x تقسیم بر پهنای h است یعنی:

$$u = \frac{x - x_i}{h}$$

تخمین چگالی احتمال را برای مقادیر x که در زیر مشخص شده بدست آورید.

$$x = \{0, 1, 2, 3, 4, 5\}$$

د- چنانچه بتوان چگالی نمونه‌های مجموعه X را گوسی فرض کرد، تخمین بیشینه شباهت پارامترهای گوسی را بدست آورید.

ه- تخمین چگالی بدست آمده از روش هیستوگرام، روش پارزن با کرنل مثلثی و تخمین پارامتری بیشینه شباهت را مقایسه کنید. کدامیک بهتر توزیع داده‌ها را تخمین می‌زند؟ ضعف هر کدام از تخمین‌ها چیست؟ بیان کنید هر کدام از تخمین‌های فوق را در چه شرایطی ترجیح می‌دهید؟

۸- (اختیاری) (میان ترم بازشناسی آماری الگو سال ۱۳۹۹)

۳ نمونه از یک توزیع یک بعدی در اختیار داریم و می‌خواهیم با کمک روش پارزن تخمین چگالی احتمال آن را بدست آوریم. این سه نمونه در نقاط $x = 1, 2, 100$ قرار گرفته است. اگر $h_N = 2$ و از کرنل پارزن مستطیلی استفاده کنیم:
الف - تخمین تابع چگالی احتمال این توزیع را بدست آورید. ب- شکل توزیع آن را رسم کنید. ج- میانگین این توزیع چقدر است؟ د- مقدار واریانس این توزیع چقدر است؟

۹- از شما خواسته شده است که با پیشنهاد یک دسته‌بند، جنسیت افراد (G) را مبتنی بر وزن فرد (W) و وضعیت فارغ التحصیلی او (S) مشخص کنید. فرض کنید وزن و وضعیت فارغ التحصیلی از هم مستقل باشند و واریانس توزیع احتمال $P(W|G=female)$ با واریانس $P(W|G=male)$ برابر باشد. دقت داشته باشید که G فقط دو مقدار male و female، W مقادیر پیوسته و S دو مقدار 0 و 1 را می‌پذیرد.

الف) آیا استفاده از طبقه‌بند Naïve Bayes برای این مسئله منطقی است؟

ب) اگر پاسخ شما به این سوال منفی است، دلیل بیاورید و توضیح دهید که چگونه می‌توان با تغییر روابط در این مسئله، امکان استفاده از کلاسبند Naïve Bayes را فراهم ساخت؟ در غیر این صورت، اگر پاسخ شما به سوال مطرح شده مثبت می‌باشد، تمام توزیع‌های احتمالاتی که کلاسبند شما باید آموزش ببیند را بنویسید و تعداد کل پارامترهایی که کلاسبند شما باید از مجموعه داده آموزش تخمین بزند را بیان کنید.

۱۰- (اختیاری) (میان ترم بازشناسی آماری الگو ۱۴۰۱)

سازگاری (consistency) یک تخمینگر را تعریف کنید و نشان دهید تخمینگر پارزن یک تخمینگر سازگار است.

۱۱- (میان ترم بازشناسی آماری الگو ۱۴۰۲)

فرض کنید نمونه‌های $1, -2, 3, 3, 7, 7$ از کلاس اول و نمونه‌های $8, 5, 3, -2, -3$ را از کلاس دوم مشاهده کرده‌ایم. با استفاده از تخمینگر پارزن با کرنل زیر و $h=2$ توزیع هر کلاس را تخمین زده‌ایم. نشان دهید به ازای $a=2$ و با استفاده از کلاسیفایر بیشینه شباهت، نمونه جدید در نقطه 4 برای گروه 1 دسته‌بندی می‌شود.

$$\varphi(x) = \begin{cases} \frac{a}{2}; & |x| < a \\ 0; & O.W. \end{cases}$$

۱۲- یک مجموعه از نمونه‌ها در فضای یک بعدی از یک توزیع ناشناخته $p(x)$ در اختیار است:

$x=1, 1.5, 1.75, 2, 2.5, 2.75, 3, 5, 6, 6.25, 6.5, 7, 7.5$

مطلوب است تخمین چگالی احتمال $\hat{p}(x)$ در نقاط زیر:

$x=0, 1, 3, 5, 7, 9$

با استفاده از یک پنجره پارزن با پهنای ۱

پاسخ:

$$\hat{p}(0) = 0, \hat{p}(1) = 2/13, \hat{p}(3) = 3/13, \hat{p}(5) = 1/13, \hat{p}(7) = 3/13, \hat{p}(9) = 0$$

۱۳- (میان ترم ۱۴۰۳) کدام گزینه‌ها درباره تخمین پارزن درست هستند؟

- الف- دسته‌بند پارزن (با انتخاب بهینه پارامتر هموارسازی) برای مجموعه‌های آموزشی بسیار بزرگ، دسته‌بند بیز را تقریب می‌زند.
- ب- تخمینگر پارزن یک تخمین بدون بایاس از توزیع ارائه می‌دهد.
- ج- معمولاً از تخمینگر پارزن برای ایجاد یک دسته‌بند غیرپارامتری و از تخمینگر kNN برای تخمین تابع چگالی احتمال استفاده می‌شود.
- د- هر چقدر تعداد نمونه‌ها بیشتر باشد، از کرنل هموارتری برای تخمین توزیع با روش پارزن استفاده می‌شود.

موفق باشید