به نام خدا

یادگیری ماشین

نیمسال دوّم ۰۴-۰۳

استاد: جناب آقای دکتر امیرخانی

دستیار استاد: علی اثنی عشری

دانشکده مهندسی کامپیوتر

تمرین دوّم | تخمین | مهلت ارسال: ۸ فروردین ماه

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ مشخص شده است.
- هم‌کاری و هم‌فکری در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر فرد باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری یا استفاده از منابع خارج از درس، نام هم‌فکران و آدرس منابع مورد استفاده را ذکر کنید.
- برای دریافت نمره کامل هر سوال نیاز است تمامی روابط و فرمول ها نوشته شده و توضیحات تشریحی کامل داده شود.
- لطفاً تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد. (برگه پاسخ سوالات نظری به صورت حضوری دریافت می گردد.)
- پاسخ تمام سوالات را در یک فایل فشرده به صورت ML_Hw2_[firstName]_[lastName]_[StudentId] نامگذاری کرده و ارسال کنید.

سوالات نظری

سوال یک (۸ نمره)

فرض کنید تابع چگالی احتمال متغیر تصادفی Y به شکل زیر است.

$$f_Y(y|\theta) = \begin{cases} \frac{1}{\theta} r y^{r-1} e^{-\frac{y^r}{\theta}}, & \theta > 0, y > 0 \\ 0, & elsewhere \end{cases}$$

که r یک ثابت مثبت است.

الف) تابع $log\ likelihood$ را به دست بیاورید.

ب) توضیح دهید در چه حالتی و چرا تخمین گر مقدار برآورد بیشینه احتمال پسین [1] به تخمین گر درست‌نمایی بیشینه [2] میل می کند.

سوال دو (۱۲ نمره)

توزیع نرمال $p(x) \sim N(\mu, \sigma^2)$ و پنجره پارزن $\varphi(x) \sim N(0,1)$ را در نظر بگیرید. نشان دهید که تخمین پنجره پارزن [3] زیر

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

برای h_n های کوچک دارای ویژگی های زیر است :

الف) $p_n^{\sim}(x) \sim N(\mu, h_n^2 + \sigma)$

ب) $p_n(x) - p_n^{\sim}(x) \approx \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2 \left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right] p(x)$

پ) $var[p_n(x)] \approx \frac{1}{2nh_n\sqrt{\pi}} p(x)$

سوال سه (۱۰ نمره)

توزیع یکنواخت $p(x)$ و پنجره پارزن $\varphi(x)$ به صورت زیر تعریف شده است:

$$p(x) \sim U(0, a)$$

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

الف) نشان دهید که میانگین چنین تخمینی از پنجره پارزن به صورت زیر می‌شود.

$$\bar{p}_n(x) = \begin{cases} 0; & x < 0 \\ \frac{1}{a}\left(1 - e^{-\frac{x}{h_n}}\right); & 0 \leq x \leq a \\ \frac{1}{a}\left(e^{\frac{a}{h_n}} - 1\right) e^{-\frac{x}{h_n}}; & a \leq x \end{cases}$$

ب) $\bar{p}_n(x)$ را بر حسب x برای $h_n = \{1, \frac{1}{4}, \frac{1}{16}\}$ و $a = 1$ رسم کنید.

پ) h_n چه قدر باشد تا در بازه $0 \leq x < a$ مقدار بایاس کمتر از ۱ درصد باشد.

[1] $Maximum\ A\ Posteriori\ Estimation$
[2] $Maximum\ Likelihood\ Estimation$
[3] $Parzen\ Window - based\ Estimation$

سوال چهار (۱۸ نمره)

ساختار الگوریتم k–نزدیک‌ترین همسایه را در نظر بگیرید. سپس به سوالات زیر پاسخ بدهید.

الف) چرا در این روش، مقدار فرد k بر مقادیر زوج ترجیح داده می شود؟

ب) آیا در این روش، هنگام آموزش و آزمون انجام $feature\ scaling$ لازم است؟

پ) در این روش، مقادیر خیلی کوچک یا خیلی بزرگ K چه تاثیری در نتایج پیش بینی دارد؟

ث) توضیح دهید که چرا این الگوریتم محاسبات بیشتری را در مجموعه آزمون نسبت به بخش آموزش انجام می‌دهد.

ج) نشان دهید که برای یک مجموعه داده آموزشی به اندازه N، خطای طبقه‌بندیِ k–نزدیک‌ترین همسایه، P_{kNN}، کران بالای زیر را نتیجه می دهد:

$$P_B \leq P_{kNN} \leq P_B \left(2 - \frac{2}{M}\right)$$

که در آن، P_B خطای طبقه‌بندی بیز و M تعداد کلاس‌ها است. توضیح دهید که چرا این نتیجه نشان می‌دهد که k–نزدیک‌ترین همسایه با افزایش $N \to \infty$ به‌صورت تقریباً بهینه عمل می‌کند.

ح) نشان دهید که مدل k–نزدیک‌ترین همسایه برای $(K \neq 1)$ تابع توزیع نامناسبی را تعریف می‌کند که انتگرال آن در تمام فضا واگرا می‌باشد.

ه) براساس باقی مانده شماره دانشجویی شما بر عدد ۲، یکی از الگوریتم های زیر را انتخاب کنید و توضیح دهید و مثال بزنید که چگونه این الگوریتم می تواند در تسریع الگوریتم k–نزدیک‌ترین همسایه کمک کند

۰) الگوریتم $K - D\ Tress$

۱) الگوریتم $Inverted\ List$

سوال پنج (۱۲ نمره)

ساختار الگوریتم هیستوگرام را در نظر بگیرید. سپس به سوالات زیر پاسخ بدهید.

الف) کران پایین کرامر–رائو[۴] بیان می‌کند که واریانس هر برآوردگر نااریب[۵] دارای یک کران پایین است. نشان دهید که واریانس برآوردگر هیستوگرام مقدار زیر را نتیجه می دهد:

$$VAR(\hat{p}(x)) \geq \frac{p(x)(1 - p(x)V)}{NV}$$

[۴] $Cramer - Rao\ lower\ bound$

[۵] $unbiased$

و توضیح دهید که این نتیجه چگونه ضعف هیستوگرام‌ها را در فضاهای با ابعاد بالا نشان می‌دهد.

ب) تعداد بازه‌های M در یک هیستوگرام به عنوان یک پارامتر هموارسازی[۶] عمل می‌کند. با استفاده از بهینه‌سازی میانگین مربعات خطا [۷]، نشان دهید که انتخاب بهینه M برای یک مجموعه داده تک‌بعدی با چگالی واقعی $p(x)$ و N نمونه به‌صورت زیر نتیجه می‌شود:

$$M^* = \mathcal{O}(N^{1/3})$$

پ) فرض کنید داده‌های ما در یک فضای d-بعدی قرار دارند و ما m بازه در هر بعد انتخاب می‌کنیم. نشان دهید که تعداد کل بازه‌ها به صورت نمایی رشد می‌کند:

$$M = m^d$$

و توضیح دهید که چرا این موضوع باعث می‌شود که برآورد چگالی مبتنی بر هیستوگرام در ابعاد بالا غیرعملی شود.

سوال شش (۴۰ نمره)

در یک آزمایشگاه تحقیقاتی دانشگاه علم و صنعت، تیمی از دانشجویان در تلاش بودند تا مدل پیش‌بینی‌گر قدرتمندی برای تحلیل داده‌های ناشناخته توسعه دهند. برای این منظور، آن‌ها تصمیم گرفتند از روش k-نزدیک‌ترین همسایه استفاده کنند.

یکی از اعضای تیم، مأموریت ویژه‌ای بر عهده داشت: او باید نوتبوک k-نزدیک‌ترین همسایه را از این لینک باز کرده و مراحل تعیین‌شده را با دقت اجرا می‌کرد. چالش اصلی این بود که پس از پردازش داده‌ها، مدل باید روی مجموعه‌ای از داده‌های آزمایشی که در فایل $test.csv$ قرار داشت، پیش‌بینی‌هایی انجام می‌داد.

نتایج این پیش‌بینی‌ها باید در یک فایل خروجی به نام $result.csv$ ذخیره می‌شد؛ فایلی که تنها یک ستون با نام $target$ داشت. لطفاً به این فرد در به انجام رساندن این ماموریت کمک کنید.

---

[۶] *smoothing*

[۷] *Mean Squared Error*