باسمه تعالی

دانشکده مهندسی کامپیوتر

آمار و احتمال مهندسی

نیم‌سال اول ۹۵-۹۴

دکتر مطهری

تمرین هفتم	آمار	زمان تحویل بخش مسایل: ۹۴/۱۰/۲۱- بخش شبیه سازی: ۹۴/۱۱/۰۳

مسایل

مساله اول

فرض کنید مقدار y از متغیر تصادفی Y را مشاهده کرده‌ایم، که از توزیع $f(y;\theta)$ پیروی می‌کند. θ بردار تمام پارامترهای توزیع و y بردار تمامی مشاهدات است. منظور از likelihood یا درستی‌نمایی θ برحسب y،

$$L(\theta) = f(y;\theta)$$

است که L تابعی از θ برای y ثابت است. در حالتی که y برداری از مشاهدات مستقل باشد به وضوح داریم

$$L(\theta) = \prod f(y_j;\theta)$$

بخش ۱ *فرض کنید $y = (y_۱, \ldots, y_n)$ نمونه‌هایی تصادفی از توزیع نمایی $f(y;\theta) = \theta^{-۱}e^{-y/\theta}$ باشد. درستی‌نمایی y را حساب کنید. این مقدار به ازای چه مقداری از θ بیشینه می‌شود؟ آیا درستی‌نمایی شامل تک تک داده‌هاست یا فقط تابعی از آن‌ها را در دل خود دارد؟*

معمولاً درستی‌نمایی را در مقیاس لگاریتمی نمایش می‌دهند و به آن log-likelihood می‌گویند:

$$\ell(\theta) = \log L(\theta) = \sum \log f(y_j;\theta) = \sum \ell_j(\theta)$$

بخش ۲ *مقدار $\ell(\theta)$ را برای سوال قبل حساب کنید.*

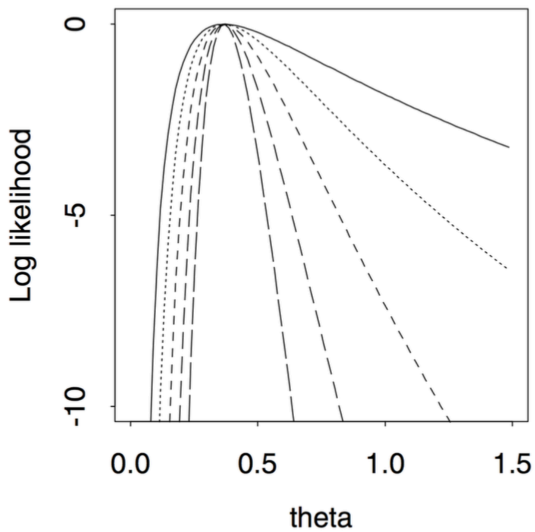
بخش ۳ *اگر در حال مقایسه‌ی دو مدل متفاوت برای یک‌سری داده باشیم، آیا صحیح است که مقدار درستی‌نمایی هر یک را حساب کنیم و بر اساس آن قضاوت کنیم که کدام مدل بر داده‌ها بیشتر منطبق است؟*

می‌توان به جای مقدار اصلی درستی نمایی، درستی‌نمایی نسبی (Relative Likelihood) را تعریف کرد. چون مقدار درستی‌نمایی تحت تبدیلات یک به یک عوض می‌شود، بنابراین منطقی است که نسبت را ملاک قرار دهیم:

$$RL(\theta) = \frac{L(\theta)}{\max_{\theta'} L(\theta')}$$

این نسبت عددی بین صفر و یک خواهد بود. بنابراین به نظر می‌آید که مقادیری از θ که به ازای آن‌ها مقدار RL زیاد است، بهتر داده‌های ما را توصیف کنند. مثلاً بگوییم اگر $\frac{۱}{۲} < RL(\theta) \leq ۱$، مقدار θ بسیار مناسب است.

وقتی که تعداد پارامترها زیاد باشد، معمولاً از خلاصه‌شده (Summariezed)ی درستی‌نمایی استفاده می‌کنند، به این شکل که حول نقطه‌ای که مقدار درستی‌نمایی بیشینه می‌شود $(\hat{\theta})$ بسط تیلور می‌نویسند و تابع را با تابعی درجه ۲ تقریب می‌زنند. به $\hat{\theta}$، MLE یا برآوردگر بیشینه درست‌نمایی می‌گویند. (برآوردگر از تمرین قبل یادتان هست؟)



شکل ۱: مقدار واقعی پارامتر توزیع نمایی $e^{-۱} \approx ۰٫۳۶$ بوده است.

بخش ۴ *این کار را برای سوال ۱ انجام دهید. یعنی مشتق دوم را در نقطه‌ی $\hat{\theta}$ محاسبه کنید، و سعی کنید تقریبی درجه ۲ از مقدار $\log RL(\theta)$ ارائه دهید.*

نمودار زیر لگاریتم تابع درستی‌نمایی را به ازای $n = ۵, ۱۰, ۲۰, ۴۰, ۸۰$ نمونه از توزیع نمایی کشیده است (کدام نمودار برای کدام n است؟). آیا این موضوع (بسته‌تر شدن دهانه‌ی سهمی در نزدیکی $\hat{\theta}$) با نتیجه‌ی مسئله‌ی ۴ که شما به دست آوردید، مطابقت دارد؟

مشخص است که هرچه دهانه‌ی سهمی بسته‌تر باشد، با صراحت بیشتری می‌توان گفت که پارامتر اصلی، نزدیک به $\hat{\theta}$ است. بیایید بسط تیلور را بنویسیم:

$$\log RL(\theta) = (\theta - \hat{\theta})\ell'(\hat{\theta}) + \frac{۱}{۲}(\theta - \hat{\theta})^{۲}\ell''(\theta_{۱}) = \frac{۱}{۲}(\theta - \hat{\theta})^{۲}\ell''(\theta_{۱})$$

که در اینجا $\theta_۱$ مقداری بین $\theta, \hat{\theta}$ است و تساوی آخر هم به این دلیل است که ℓ در $\hat{\theta}$ بیشینه است. دقت کنید که دهانه‌ی سهمی را مقدار $\ell''(\theta)$ تعیین می‌کند. این مقدار اینقدر مهم است که به آن «اطلاعات مشاهده شده» (یا observed information) می‌گویند:

$$J(\theta) = -\frac{\mathrm{d}^{۲}\ell(\theta)}{\mathrm{d}\,\theta^{۲}} = \sum_{j=۱}^{n} -\frac{\mathrm{d}^{۲}\log f(y_j;\theta)}{\mathrm{d}\,\theta^{۲}}$$

بخش ۵ *به طور کلی، انتظار ما این است که هر چه داده‌ها بیشتر شوند، $n \to \infty$، با قطعیت بیشتری می‌توان در مورد پارامترها حرف زد. بیان این حرف بر اساس گزاره‌های گذشته چیست؟ آیا این حرف در مورد توزیع نمایی درست است؟*

حال یک سوال جالب: اگر ما داده‌ها را داشته باشیم، خواهیم فهمید که اطلاعات به دست آمده از پارامترهای ما چقدر است ($J(\theta)$)، آیا قبل از انجام آزمایش نیز می‌توان شهودی نسبت به این مقدار داشت؟ یعنی بفهمیم که بعد از انجام این آزمایش، حول و حوش چقدر اطلاعات کسب خواهیم کرد! این کار انجام پذیر است و به آن اطلاعات فیشر (Fisher Information) می‌گویند و به این صورت تعریف می‌شود:

$$I(\theta) = \mathbb{E}\left(-\frac{\mathrm{d}^{۲}\ell(\theta)}{\mathrm{d}\,\theta^{۲}}\right)$$

اگر داده‌های ما نمونه‌هایی تصادفی باشند، داریم

$$I(\theta) = n \cdot i(\theta) = n \cdot \mathbb{E}\left(-\frac{\mathrm{d}^{۲}\log f(Y_j;\theta)}{\mathrm{d}\,\theta^{۲}}\right)$$

بخش ۶ *برای توزیع دوجمله‌ای، با مخرج m و احتمال موفقیت p، مقدار $I(p)$ را حساب کنید. رابطه‌ی اطلاعات با m چگونه است؟*

بنابراین طبق این تعریف، می‌توان آزمایش‌ها را از جهت داده‌های مورد نیاز برای قطعیت بیشتر در مورد یک پارامتر با هم مقایسه کرد. فرض کنید آزمایش A، اطلاعات $I_A(\theta)$ و آزمایش B، $I_B(\theta)$ را می‌دهد. اگر اطلاعات این دو بخواهد یکسان باشد، $I_A(\theta) = I_B(\theta)$ داریم،

$$n_A i_A(\theta) = n_B i_B(\theta) \Rightarrow \frac{n_B}{n_A} = \frac{i_A(\theta)}{i_B(\theta)}$$

یعنی تعداد آزمایش‌های موردنیاز به نسبت عکس اطلاعات به دست می‌آیند.

بخش ۷ *می‌خواهیم ببینیم اگر اعداد را رند کنیم، چقدر اطلاعات از دست می‌رود. فرض کنید Y متغیری با توزیع $N(0, \sigma)$ باشد. در ذخیره‌سازی داده‌ها، مقدار Y به مقدار X رند شده، که X نزدیک‌ترین مضرب δ به Y است. یعنی اگر Y در بازه‌ی $[(k-\frac{1}{2})\delta, (k+\frac{1}{2})\delta)$ باشد، $X = k\delta$ خواهد بود.*

نسبت مقدار اطلاعات را برای X و Y در مورد σ به دست آورید.

مساله دوم

فرض کنید $Y_1, \ldots, Y_n$ نمونه‌ای تصادفی و نرمال باشند. $\bar{Y}$ را میانگین نمونه‌ای و S^2 را واریانس نمونه‌ای می‌گوییم و برابر مقادیر زیر قرار می‌دهیم:

$$\bar{Y} = \frac{Y_1 + \cdots + Y_n}{n}, \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

همانطور که می‌دانید این‌ها برآوردگرهایی از مقادیر μ و σ^2 هستند. سعی کنید گزاره‌های زیر را ثابت کنید:

$$\bar{Y} \sim N(\mu, n^{-1}\sigma^2)$$

$$(n-1)S^2 \sim \sigma^2 \chi^2_{n-1}$$

مساله سوم

فرض کنید داده‌هایی در اختیار دارید و می‌دانید که داده‌های شما از k دسته‌ی متفاوت $1, \ldots, k$ آمده است. برای هر مشاهده‌ی x احتمال اینکه x از دسته‌ی iـام آمده باشد را π_i فرض کرده‌اید. به کمک آماره‌ی پیرسون می‌توانید درستی فرض چنین احتمال‌هایی برای دسته‌هایتان را آزمون کنید. ثابت کنید اگر تعداد n مشاهده داشته باشیم و O_i را تعداد مشاهدات دستهی iـام بگیریم خواهیم داشت،

$$\sum_{i=1}^{k} \frac{(O_i - n\pi_i)^2}{n\pi_i} \to \chi^2_{k-1}$$

شبیه سازی

سوال اول

به دانشجویان یک کلاس ۵۰ نفره یک تمرین داده شده است. در این تمرین افراد باید از زمان صفر شروع به روشن کردن لامپ کنند و پس از سوختن هر لامپ، بلافاصله لامپ بعدی را روشن کنند و در آخر زمان سوختن لامپ ۱۰۰ را گزارش کنند. اعداد گزارش شده‌ی این افراد در فایل "d1.txt" به شما داده شده است. اگر بدانیم که عمر لامپ‌ها یک متغیر تصادفی توانی با پارامتر ۱ می‌باشد، با استفاده از قضیه حد مرکزی به سوالات زیر پاسخ دهید:

۱. فرض کنید که شما برای تصحیح تمرین این افراد باید فیلم عملکرد آنها را تماشا کنید اما چون این کار وقت گیر است تصمیم می‌گیرید که از دانش آماری خود بهره بگیرید. روشی ارایه کنید که بتوان گفت اگر کسی آزمایش را به درستی انجام داده باشد، با احتمال ۹۵ درصد نمره‌ی قبولی می‌گیرد و بر این اساس به هر کدام از افراد یک نمره اختصاص دهید. اگر پاسخ درست است ۱ و اگر نادرست است ۰ را به عنوان نمره، در سطرهای فایل "p11.txt" یادداشت فرمایید.

۲. حال شما دوست دارید که به طور میانگین تنها یکی از افراد کلاس به شما در مورد نمره‌ی خود اعتراض کند (فرض کنید فقط افرادی که آزمایش را به درستی انجام داده‌اند و نمره‌ی ۱ نگرفته‌اند اعتراض می‌کنند). برای این کار یک روش ارائه کنید و بر اساس آن نمره‌های افراد را در فایل "p12.txt" یادداشت کنید. یادداشت کنید. آیا استفاده از این روش‌ها به نفع دانش آموزان تنبل است یا به ضرر آن‌ها؟

۳. اگر برای تمام افراد فیلم‌ها نگاه شود، خطای تصحیح ۰ خواهد بود، اما زمان زیادی صرف خواهد شد. یک روش مناسب می‌تواند آن باشد که جواب‌هایی که خیلی خوب هستند را از بررسی خارج کنیم و جواب‌هایی که مشکوک هستند را بررسی کنیم. برای این کار یک روش مناسب ارائه کنید و نمراتی را که از بررسی خارج می‌کنید را با عدد ۱ در فایل "p13.txt" مشخص کنید و به جای دیگر اعداد گزارش شده ۰ بگذارید. اتخاذ این روش به نفع افراد تنبل کلاس است یا به ضرر آن‌ها؟

در این تمرین فرض شده است که افراد تقلب نمی‌کنند و اعداد گزارش شده‌ی آن‌ها تنها نشان دهنده‌ی مهارت آن‌ها در روشن کردن لامپ‌ها و اندازه‌گیری زمان و نیز سرعت عمل آن‌ها است! مثلا اگر شخصی لامپ را خوب نبندد عمر آن کمتر می‌شود و اگر سرعت عمل خوبی نداشته باشد زمان بیشتر از مقدار واقعی را گزارش خواهد کرد و اگر در اندازه‌گیری زمان مشکل داشته باشد ممکن است زمان را بیشتر یا کمتر از مقدار واقعی گزارش کند. همچنین نفع یا ضرر روش‌ها را به نسبت تصحیح مورد به مورد تمرین‌ها با مشاهده‌ی فیلم تمرین‌ها بسنجید.

سوال دوم

اعداد داده شده در فایل "d2.txt" هر کدام میانگین ۲۰۰ نمونه‌ی تصادفی از یکی از توزیع‌های زیر هستند.

۱. $Exponential(\lambda = ۰/۱۱)\ ,\ [mean = ۱/\lambda]$

۲. $Normal(\mu = ۱۰/۵, sigma = ۳)\ ,\ [variance = \sigma^۲]$

۳. $Poisson(\lambda = ۱۰)\ ,\ [mean = \lambda]$

در فایل "p2.txt" مشخص کنید که محتمل‌ترین توزیع متناظربرای عدد گزارش شده کدام است. این مشخص کردن با یکی از اعداد ۱، ۲ و یا ۳ صورت می‌پذیرد. واضح است که نمی‌توان انتظار داشت که تمام حدس‌های ما در این بخش درست باشد. انتظار دارید که توزیع حدود چند تا از اعداد با آنچه به عنوان محتمل ترین توزیع به دست آورده اید یکی باشد؟

سوال سوم

در این سوال از کتابخانه‌ی "MASS" و مجموعه داده‌ی "Boston" که در همان کتابخانه وجود دارد، استفاده کنید. پس از بارگذاری کتابخانه، این مجموعه داده به راحتی و با نوشتن اسم آن قابل دسترسی است. با نوشتن `Boston?` می توانید به توضیحاتی در مورد این دادگان دست پیدا کنید.

در این تمرین هدف آن است که با استفاده از یافتن یک خط که مجذور خطا را کمینه کند، به پیش بینی مقادیر ستون آخر این دادگان (medv) بپردازیم. می‌خواهیم مقادیر ستون آخر را به عنوان تابعی از دیگر ستون‌ها در نظر بگیریم و از بین ستون‌های دیگر آن را انتخاب کنیم

که بهترین قدرت پیش بینی در مورد مقادیر ستون آخر را به ما می‌دهد. پس چون در کل ۱۴ ستون داریم شما باید ۱۳ خط پیدا کنید و بررسی کنید که کدام یکی از این خط‌ها، خطای کمتری در تخمین مقادیر ستون آخر دارد. مقادیر خطای هر کدام از موارد را به ترتیب در یک خط از فایل "p3.txt" بنویسید. همچنین نمودار بهترین خط را در گزارش خود رسم کنید.

سوال چهارم

فرض کنید جدول زیر بیانگر تعداد کلاس‌های یک دانشگاه با تعداد مشخصی از دانشجویان باشد. ابتدا تعداد مناسبی دانشجو در نظر بگیرید و به هر دانشجو ۵ یا ۶ کلاس اختصاص دهید به گونه‌ای که مطابق جدول زیر باشد. حال میانگین اندازه‌ی کلاس‌ها را با استفاده از جدول به دست بیاورید. یک روش برای تخمین زدن اندازه‌ی کلاس‌های این دانشگاه آن است که تعدادی دانشجو را به صورت تصادفی انتخاب کنیم و از آن ها میانگین اندازه‌ی کلاس‌هایی را که در آن حضور دارند را بپرسیم و بین این اعداد به دست آمده میانگین بگیریم. این عدد را با میانگین واقعی مقایسه کنید و نتیجه گیری (اخلاقی!) کنید. همچنین نمودار تخمین به دست آمده را بر حسب اندازه‌ی نمونه رسم کنید.

تعداد دانشجویان در کلاس	تعداد کلاس‌ها
۵–۹	۸
۱۰–۱۴	۸
۱۵–۱۹	۱۴
۲۰–۲۴	۴
۲۵–۲۹	۶
۳۰–۳۴	۱۲
۳۵–۳۹	۸
۴۰–۴۴	۳
۴۵–۴۹	۲
۵۰–۵۴	۱