

آمار و احتمال مهندسی، مجلد 26 : 21 خرداد 1401

* ادله بحث مقدماتی حدی

• مقیاس حد مرکزی (Central Limit theorem)

* مباحثی از آمار و استنباط آماری

• راجع به نظایر احتمالات و استنباط آماری

• درگاه (پارچوب) مشهور در آمار

• تغییر پیشینه درست نمایی (Maximum Likelihood Estimator)

• تغییر MAP (Maximum A posteriori Estimator)

• مقیاس مرکزی :

اگر $x_1, \dots, x_n$ مقیاس های مستقل که توزیع مشابه دارند باشند (iid)

و هر کدام از این ها دارای میانگین μ و واریانس σ^2 باشند.

$$\mathbb{E}[X_i] = \mu, \quad \text{Var}[X_i] = \sigma^2, \quad X_i \perp X_j$$

$$S_n = X_1 + \dots + X_n \quad \text{مجموع کسب متغیرات:}$$

$$\Rightarrow \mathbb{E}[S_n] = n\mu$$

$$\Rightarrow \text{Var}[S_n] = n\sigma^2$$

حال متغیر استاندارد Z را به صورت زیر تعریف کنیم:

$$Z_n = \frac{S_n - n\mu}{\sqrt{n} \sigma}$$

$$\Rightarrow \mathbb{E}[Z_n] = 0$$

$$\Rightarrow \text{Var}[Z_n] = 1$$

متغیر استاندارد Z بیان می‌کند که برای n های به اندازه کافی بزرگ، توزیع Z_n به یک توزیع گاوسی $N(0,1)$ میل می‌کند.

$$Z_n \xrightarrow[n \rightarrow \infty]{D} \mathcal{N}(0,1)$$

به عبارت دقیق تر :

$$\mathbb{P}[Z_n \leq \alpha] \xrightarrow[n \rightarrow \infty]{} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{x^2}{2}} dx$$

• تقریب گاوسی برای توزیع دو جمله‌ای :

فرض کنید $X \sim \text{Binomial}(n, p)$. قبلاً دیدیم که X را به صورت جمع n متغیر تصادفی مستقل نوشت.

$$X = I_1 + \dots + I_n, \quad I_i \sim \text{Ber}(p), \quad I_i \in \{0, 1\}$$

$$, \quad I_i \perp I_j$$

می‌توانیم از قضیه مرکزی راسخانه کنیم :

$$\frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}[X]}} = \frac{X - np}{\sqrt{np(1-p)}} \xrightarrow[n \rightarrow \infty]{D} \mathcal{N}(0,1)$$

نکته: این تفریب برای زمانی که $np(1-p)$ تقریب قوی است.

مثال: یک مسافه قبایلی بهرتاب دارت داریم که احتمال تروتن با به هدف $p=0.01$

است. این دارت را به صورت مستقل ۱۰۰۰۰ بار بهرتاب می‌کنیم.

سوال: احتمال اینکه حداقل ۴۰ بار به هدف افتد چیست؟

راه حل دقیق:

$$P_{\leq 40} = \sum_{i=0}^{40} \binom{10000}{i} (0.01)^i (0.99)^{10000-i}$$

این عبارت را می‌تواند به دست آورد.

راه حل تقریبی با استفاده از تقیه مرکزی:

$$P[X \leq 40] = P\left[\frac{X - 100}{\sqrt{99}} \leq \frac{40 - 100}{\sqrt{99}}\right]$$

$Z \sim N(0,1)$ و معیار $\frac{X-100}{\sqrt{99}}$ تقریباً به توزیع Z می‌باشد.

CDF یک گامی $N(0,1)$ →

$$\approx P[Z \leq -6.03] = \frac{1}{F_Z(-6.03)}$$

$\approx 8.2 \times 10^{-10}$ ← با دست‌ده از جدول یا
کتابخانه‌های کامپیوتری

حال اگر بپرسیم که احتمال اینکه امتیاز دانشجوی 8 و 120 باشد چقدر است، به صورت
عکس می‌کنیم:

$$P[80 \leq X \leq 120] = P\left[\frac{80-100}{\sqrt{99}} \leq \underbrace{\left(\frac{X-100}{\sqrt{99}}\right)}_{\sim N(0,1)} \leq \frac{120-100}{\sqrt{99}}\right]$$

$$\approx P[-2.01 \leq Z \leq 2.01]$$

$$= F_Z(2.01) - F_Z(-2.01) \approx 0.956$$

با ۵۰٪ بیش از ۵۰٪ استیازها در یازده [۸۰، ۱۲۰] مشاهده بود.

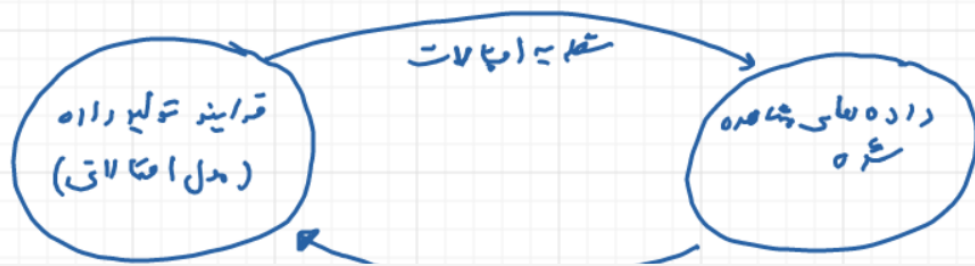
- مثال بالا بیان می‌کند بیشتر افعال با حول می‌نگین و در یازده کوچکی قرار می‌گیرد.
- یک نکته قابل توجه در مثال بالا هست. اگر آزمائشی برای مقایسه مان را به مقدار زیاد بارنگار کنیم و بهیضی کیفیت با به مقایسه فامی می‌کنند و دیگر زیاد مقایسه می‌کنند.

Con Con tration



* آمار و استنباط آماری (Statistical Inference):

- ۵۰٪ استنباطی در بیشتر نظریه‌ها است و راجع به ۵۰٪ بود.
- راجع به ۵۰٪ استنباطی و آمار در شکل زیر بیان می‌شود:



کتابخانه آماري - داده کاوي

• مسئله پایه ای که در امتحان بررسی می کنیم :

با داشتن یک قرائد تولید داده (یک مدل احتمالات) چه حرف هایی در مورد خودی و مشاهده
با از خواند تولید داده می توان زد؟

• مسئله پایه ای که در آمار و کتابخانه آماري بررسی می شود:

با داشتن مشاهدات و خودی های عمیق تولید داده ، چه حرف هایی در مورد خود قرائد
(مدل احتمالات) می توان زد؟

• نکته: مسائل پیشینی (prediction) ، دسته بندی (classification) ، خوشه بندی
(clustering) ، تخمین (estimation) ، حالت های خاصی از کتابخانه آماري
هستند.

• به طور کلی در نگاه (بازچوب یا Framework) برای بررسی مسائل وجود دارد.

① نگاه کلاسیک یا نگاه فرادان گرایانه (Frequentist):

در این نگاه مدل احتمالات با پارامتری مانند θ توصیف و کشود که نمی‌تواند تغییر کند و ثابت است.

یک خانواده از مدل‌های احتمالاتی که توسط پارامتر θ توصیف و کشود (θ و توانمند بر داده‌ها با هم با هم)

$f(x; \theta)$

پارامتر نامعلوم که باید از روی داده‌ها استخراج شود

② نگاه بیزین (Bayesian):

در این روش این مدل با یک پارامتر نامعلوم θ مشخص و کشودلی θ خودی یک متغیر تصادفی است.

این نگاه امتیازهای روشی که آزمایند یک اعتماد بیشین (دانشی اولیه یا نظر آزمایند)

در مورد θ را با استفاده از یک توزیع اولیه $\pi(\theta)$ بر روی θ وارد مدل کند.

توزیع پیشین یا prior dist.

بروز مشاهدات $\leftarrow$ توزیع پیشین روی θ را با قاعده بیز به روزی کنیم:

$$\pi(\theta | \underline{x}) = \frac{f(\underline{x} | \theta) \cdot \pi(\theta)}{m_{\underline{x}}(\underline{x})} \rightarrow \underline{x} \text{ توزیع ماکسیمای}$$

$\downarrow$
توزیع پسین θ
posterior dist.

$$\downarrow$$

$$m_{\underline{x}}(\underline{x}) = \int f(\underline{x}, \theta) d\theta$$

$$= \int f(\underline{x} | \theta) \pi(\theta) d\theta$$

حال یادگرفتن توزیع پسین $\pi(\theta | \underline{x})$ سعی می‌کنیم متناهی در مورد θ بکشد می‌آوریم.

به عنوان مثال $\mathbb{E}[\theta | \underline{x}]$ می‌توانیم تغییرات مستطای برای θ بکشد.
 $\downarrow$
میانگین

به عنوان مثال دیگر، $\hat{\theta}$ ای که $\pi(\theta | \underline{x})$ بیشینه می‌کند به عنوان ماکسیمای تغییرات مستطای

دگر بنای θ قواصیر ← تفتیه MAP

تفتیه بیشینه درست بنای (Maximum Likelihood Estimator):

فرض کنیم یک مدل احتمالاتی داریم که با پارامتر θ توصیف می‌شود (θ و متغیرها را با کلاه

$$X \sim \left\{ \begin{array}{l} f(x|\theta) \\ f(x;\theta) \\ f_{\theta}(x) \end{array} \right\} \quad \begin{array}{l} f \text{ و متغیرها} \\ \text{pmf, pdf} \\ \text{باشد} \end{array}$$

حال فرض کنیم مشاهدات iid $x_1, \dots, x_n$ را از این مدل داشته باشیم:

$$X_i \sim f(x|\theta)$$

با توجه به این مدل احتمال مشاهده یک مشاهده خاص $x_1, \dots, x_n$ چقدر است؟

$$= \prod_{i=1}^n f(x_i|\theta) \quad \text{بهرین}$$

$$\prod_{i=1}^n \dots \quad \text{گستره}$$

$$= \prod_{i=1}^n p_X(x_i | \theta)$$

تابع درست‌نمایی به صورت زیر تعریف می‌گردد:

$$L(\theta | \underline{x}) = \prod_{i=1}^n f(x_i | \theta) \quad : \quad \text{به عنوان تابعی از } \theta$$

نگاه می‌کنیم

↑
بر حسب θ تابع
اعتباریست.

نقطه بیشینه درست‌نمایی:

برای یک مشاهده $\underline{x} = (x_1, \dots, x_n)$ محتمل‌ترین θ ، θ ای است که L را بیشینه کند:

$$\hat{\theta}_{ML}(\underline{x}) = \arg \max_{\theta} L(\theta | \underline{x})$$

مثال: فرض کنید یک سکه داریم که احتمال گیر آمدن آن p است و بقیه $1-p$ است.

این سکه را به طور مستقل n بار پرتاب می‌کنیم و مشاهده می‌کنیم که k بار شیر ظاهر می‌شود. در این شرایط تخمین ML را برای p بیابید.

حل: $X_i \sim \text{Ber}(p)$, $X_i \in \{0, 1\}$

$$P[X_i = 1] = p$$

$$X_i \perp X_j$$

مشاهدات
↑

$$L(p | \underline{x}) = P[\underline{x} \text{ مشاهده} | p]$$

$$= P[X_1 = x_1, \dots, X_n = x_n | p]$$

$$= \prod_{i=1}^n \underbrace{P_X(x_i | p)}_{\text{Ber}(p)} = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

$$\sum x_i \quad n - \sum x_i$$

$$= p (1-p)^{n-k}, \quad k = \sum x_i$$

$$= p^k (1-p)^{n-k}$$

باید تابع درست‌نمایی L را بر حسب p در بازه $[0, 1]$ بیشینه کنیم.

$$\frac{\partial L(p|x)}{\partial p} = 0 \rightarrow \text{کدام‌نوعی معتدل‌های پای L را بیشینه می‌کنیم}$$

$$\Leftrightarrow \frac{\partial \ln L(p|x)}{\partial p} = 0$$

$$\Rightarrow \ln L = k \ln p + (n-k) \ln(1-p)$$

$$\Rightarrow \frac{\partial \ln L}{\partial p} = k \cdot \frac{1}{p} + (n-k) \frac{-1}{1-p} = 0$$

$$\Rightarrow \frac{k}{p} = \frac{n-k}{1-p} \Rightarrow k - \cancel{k}p = n - \cancel{n}p$$

$$\Rightarrow \hat{p}_{ML} = \frac{k}{n}$$
