احتمالات در علم داده

تهیه کننده:

رویا احمددوست رزداری

عضو تیم «علم داده»

royaahmaddoost.rz@gmail.com

اولین دوره

آموزشی رایگان

«علم داده»

Data Science

Course

Data Science Group

Dr. M. Fozouni

احتمال برای آمار و علم داده

احتمال: رخدادن یک واقعه است که با مقادیر عددی بین ۰و۱ اندازگیری می‌شود.

مثال: در پرتاب سکه پنجاه درصد احتمال شیر بودن و پنجاه درصد احتمال خط بودن داریم که احتمال هر کدام مقداری بین ۰ و۱ است.

$$P(x) = \frac{\text{نتایج مورد نظر}}{\text{فضای نمونه}}$$

نتایج مورد نظر[۱] نتایجی هستند که می‌خواهیم اتفاق بیوفتد و فضای نمونه[۲] همه‌ی نتایج احتمالی که می‌تواند رخ دهد.

اگر دو رخداد مستقل باشند:

احتمال رخ دادن آنها به طور همزمان برابر است با حاصل ضرب احتمال رخ دادن تک تک آنها.

$$P(AB) = P(A) . P(B)$$

امید ریاضی[۳] وقتی استفاده می‌شود که بخواهیم اتفاقات آینده را پیش بینی کنیم به عبارتی امید ریاضی X یک میانگین وزنی از مقادیر ممکن است که X می‌تواند اختیار کند و وزن هر مقدار، احتمالی است که X می‌تواند آن مقدار را اختیار کند.

گسسته	$E(x)=\sum_{i=1}^{n} x_i \times p_i$
پیوسته	$E(x) = \int_{-x}^{x} xf(x)\, dx$

توزیع فراوانی احتمال[۴] مجموعه‌ای از تمام احتمالات برای هر پیش آمد ممکن است وقتی که امید ریاضی غیر قابل دستیابی باشد برای پیش بینی به توزیع فراوانی احتمال نیاز داریم.

[1] *Preferred outcom*
[2] *SampleSpace*
[3] *Expected value*
[4] *probability frequency distribution*

توزیع فراوانی احتمال با تقسیم هر فراوانی به اندازه فضای نمونه و مطابق جدول زیر به دست می‌آید.

Sum	*frequency*	*Probability*
2	*1*	*1/36*
3	*2*	*1/18*
4	*3*	*1/12*
5	*4*	*1/9*
6	*5*	*5/36*
7	*6*	*1/6*
8	*5*	*5/36*
9	*4*	*1/9*
10	*3*	*1/12*
11	*2*	*1/18*
12	*1*	*1/36*

جایگشت[۵] به معنی پیدا کردن تمامی روش‌های ممکن برای انجام یک کار است.

$$p(n) = n \times (n-1) \times (n-2) \times \dots \times 1$$

مثال:

چند ترتیب متفاوت از پرتاب توپ در بازی بیسبال که متشکل از ۹ بازیکن است وجود دارد؟

حالتهای ممکن پرتاب توپ برابر با $9! = 362880$ است.

تغییرات[۶] نشان دهنده تعداد روش‌های مختلف است که می‌توان تعدادی از عناصر را انتخاب کرد.
تغییرات با تکرار :

$$\bar{V}(n,p) = n^p$$

مثال: اگر بخواهیم یک رمز دو رقمی را از بین سه عدد پیدا کنیم :

---

5 *Permutations*
6 *Variations*

$$n^p = 3^2 = 9$$

یعنی ۹ حالت ممکن وجود دارد.

تغییرات بدون تکرار:

$$V(n,p) = \frac{n!}{(n-p)!}$$

مثال: اگر بخواهیم سه میز را کنار هم قرار دهیم در انتخاب اول سه حالت و در انتخاب دوم دو حالت و در انتخاب سوم یک حالت یعنی$3 \times 2 \times 1$

$$V(n,p) = \frac{n!}{(n-p)!} = \frac{3!}{0!} = 6$$

ترکیب[7] تعداد حالات انتخاب تعدادی معین از اعضای یک مجموعه است

$$C(n,p) = \frac{n!}{(n-p)!\,p!}$$

مثال: فرض کنید می خواهیم از بین حروف a,b,c,d دو حرف انتخاب کنیم روش‌های انتخاب ما می‌تواند یکی از شکل‌های زیر باشد.

$\{c, d\}$
$\{a, b\}$
$\{a, c\}$
$\{a, d\}$
$\{b, c\}$
$\{b, d\}$
$\{c, d\}$

بنابراین $C = \frac{4!}{2!2!} = 6$ می شود.

[7] *Combinations*

مجموعه: تعدادی از اشیا است که ویژگی مشترکی دارند مجموعه‌ایی که هیچ مقداری نمی‌گیرد و خالی است را با $\emptyset$ نشان می دهیم .

متمم یک واقعه هر چیزی است که یک واقعه نیست.

$$A' = Not\ A$$

مشخصه‌های متمم:

متمم هرگز نمی‌تواند به طور همزمان رخ دهد.

$(A + A' = Sample\ space)$

$(P(A) + P(A') = 1)$

$((A')' = A)$ متمم یک متمم ، رخداد اصلی است.

مثال:

اگر $P(A') = 0.75$ و $P(A') = 1 – P(A)$ بنابراین $P(A) = 0.2$

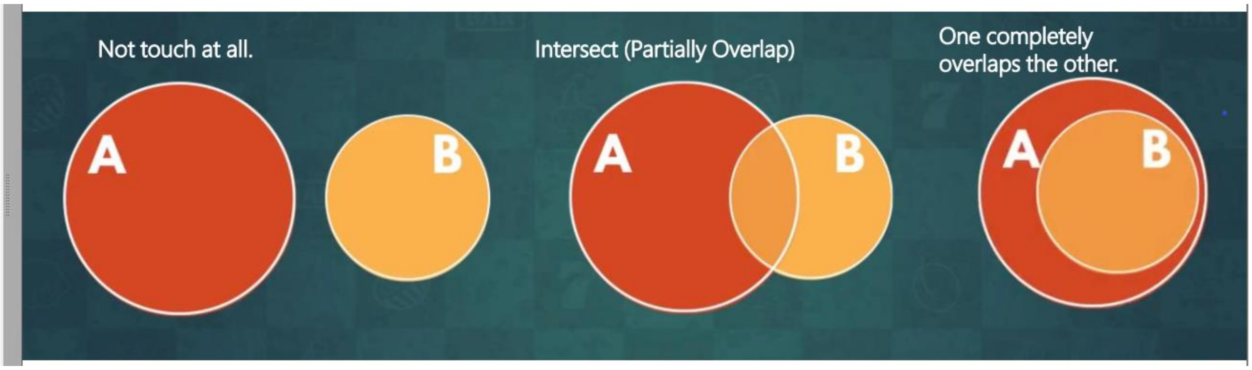
دو مجموعه A و B را در نظر می‌گیریم وقتی زیر مجموعه‌های دو مجموعه با هم اشتراک[8] داشته باشند می‌گویم هر دو پیشامد می‌توانند هم زمان رخ دهند.

نمادهای کاربردی:

x عضوی از مجموعه A است	$x \in A$
هر x	$\forall x:$
A زیرمجموعه B	$A \subseteq B$

[8] *intersection*

مجموعه‌ایی از نتایجی که رخ می‌دهد به یکی از سه مدل زیراست :



برای مثال اگر گروهA را تیم فوتبال و گروه B را تیم والیبال در نظر بگیریم ، زهرا و فاطمه عضو تیم فوتبال و زینب عضو تیم والیبال باشند در این صورت هیچ اشتراکی ندارند ولی اگر زهرا و زینب در هر دو تیم باشند با هم اشتراک دارند.

احتمال شرطی[۹] ، اگر احتمال رخداد A به احتمال رخداد B بستگی داشته باشد می‌گوییم رخداد آنها وابسته است در غیر اینصورت غیر وابسته است .

در صورت غیروابسته بودن $$P(A|B) = P(A)$$

و در صورت وابسته بودن $$P(A|B) = \frac{P(A\cap B)}{P(B)}$$

مثال: فرض کنید دو تاس نااریب را پرتاب می‌کنیم احتمال اینکه هر دو عدد ۶ باشد فرض کنید که A پیشامد مشاهده عدد ۶ در یکی از تاس‌ها باشد همچنین پیشامد مشاهده جفت را هم با B نشان می‌دهیم:

$A=\{(1,6),(6,1),(2,6),(6,2),(3,6),(6,3),(4,6),(6,4),(5,6),(6,5),(6,6)\}$

$B=\{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$

پس $A\cap B=\{(6,6)\}$ در نتیجه مقدار احتمال شرطی $P(A|B)$ به صورت زیر *محاسبه می‌شود*

[9] *conditional probability*

$$P(A\Big|B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{1}{36}}{\frac{6}{36}}$$

قانون احتمال کل[۱۰]این قانون روشی برای دسته بندی اتفاق‌ها بر پایه احتمال رخداد یا عدم رخداد یک اتفاق است.

$$P(A)=P(A|B_1)\times P(B_1)+\cdots+P(A|B_n)\times P(B_n)$$

قانون جمعی[۱۱]

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

مثال: در شهری تعدادی دزدی صورت گرفته براساس تحقیقات میدانی مشخص شده است که ۳۵٪ دزدها از سوپر مارکت دزدی کرده‌اند و ۴۰٪ دزدها از منازل دزدی کرده‌اند همچنین ۲۵٪ باقیمانده دزدها از هر دو اماکن یعنی هم سوپر مارکت هم منازل دزدی کرده‌اند نشان دهید که در آینده چقدر امکان این می‌باشد که حداقل در یکی از اماکن ذکر شده دزدی صورت گیرد؟

$$P(S\cup H)=P(S)+P(H)-P(S\cap H)=\frac{35}{100}+\frac{40}{100}-\frac{25}{100}=50\%$$

قانون ضرب احتمال[۱۲]

$$P(A\cap B)=P(A|B)\times P(B)$$

قانون بیز[۱۳]فرض کنید پیشامدA اتفاق افتاده و می‌خواهیم احتمال وقوع یکی از پیشامدهای B_k را حساب کنیم. برای اینکار ازقانون بیز که ترکیبی از قانون ضرب و قانون احتمال کل است، استفاده می‌کنیم:

---

[10] *law of total probability*

[11] *additive law*

[12] *multiplication rule*

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

مثال: براساس تحقیقی بر روی مراجعان بیمارستانی دریافته ایم که ۲/۳ افرادی که تست کرونا مثبت دارند آقا هستند و همچنین ۱/۳ دیگر افراد که تست انها مثبت شده خانم هستند با توجه به اینکه نسبت اقایان به خانوم‌ها در جامعه ۱/۲ می باشد مشخص کنید که اگر تستی را انجام داده‌ایم و مثبت شده ولی جنسیت صاحب تست پاک شده باشد با چه احتمالی امکان دارد که این شخص مذکر باشد؟

$$P(P|M) = 2/3 , P(P|F) = 1/3 , P(M)=1/2 , P(F)=1/2$$

$$P(M \Big| P) = \frac{P(M)P(P|M)}{P(P)} = \frac{P(M)P(P|M)}{P(M)P(P|M) + P(F)P(P|F)}$$

$$= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{2}{3} + \frac{1}{2} \times \frac{1}{3}} = 2/3$$

توزیع احتمال

در نظریه احتمال وآمار تابع توزیع احتمال بیانگر احتمال هر یک از مقادیر متغیر تصادفی (در مورد متغیر گسسته) و احتمال قرار گرفتن متغیر در بازه مشخص (در مورد متغیر پیوسته) می‌باشد.

توزیع ها دو مشخصه مهم دارند به نام‌های میانگین و واریانس دارند

[13] *Bayes' Rule*

واریانس: جهت مشخص کردن پراکندگی داده ها به کار می‌رود که هرچه واریانس بیشتر پراکندگی بیشتر است به طور کلی اگر X یک متغیر تصادفی با میانگین μ باشد ، آنگاه واریانس X که آن را با $Var(x)$ نشان می‌دهیم به صورت زیر تعریف می‌شود:

$$Var(x) = E[(X-\mu)^2]$$

انحراف استاندارد جذر واریانس است که با نماد σ نشان داده می شود و اگر برای نمونه باشد با نماد s نشان داده می شود.

رابطه بین تابع توزیع و تابع چگالی:

$$f(x) = \frac{dx}{dy}F(X)$$

توزیع های خاص گسسته

توزیع برنولی

آزمایشی که نتیجه آن یکی از حالت‌های شکست یا موفقیت باشد توزیع برنولی نام دارد اگر موفقت حاصل شود متغیر تصادفی X برابر ۱ و وقتی شکست حاصل شود آن را برابر۰ قرار می-دهیم. آنگاه تابع احتمال X برابر است با

$$p(0) = q$$
$$p(1) = p$$

واریانس	امید ریاضی	تابع چگالی احتمال
$Var(x)=pq$	$E(x)=p$	$P(X=x) = p^x q^{1-x}$

توزیع دوجمله‌ایی

اگر یک آزمایش برنولی با احتمال موفقیت p را n بار تکرار کنیم و X تعداد موفقیت‌ها در این n آزمایش برنولی باشد، می‌گوییم متغیر تصادفی X دارای توزیع "دوجمله‌ایی" با پارامتر های nوp می باشد.

واریانس	امید	تابع چگالی احتمال
$Var(x)=npq$	$E(x)=np$	$P(X=x)=\binom{n}{x}p^x q^{n-x}$

مثال: ۵ سکه سالم را پرتاب کرده‌ایم اگر نتایج حاصل از هر پرتاب مستقل باشند، تابع احتمال تعداد شیرهای ظاهر شده را بدست آورید.

اگر X برابر با تعداد شیرها (موفقیت) در نظر گرفته شود، آنگاه X یک متغیر تصادفی دوجمله‌ای با پارامترهای $n=5$, $p=\frac{1}{2}$ است و بنابراین با استفاده از تابع چگالی داریم:

$$P\{X=0\}=\binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5=\frac{1}{32}$$

توزیع پواسن

اگر متغیر تصادفی X به وسیله پارامتر α به مکان یا زمان مربوط شود، می‌گوییم این متغیر تصادفی دارای توزیع پواسن با پارامتر α است

واریانس	امید ریاضی	تابع چگالی احتمال
$Var(x)=\alpha$	$E(x)=\alpha$	$P(X=x)=\frac{\alpha^x e^{-\alpha}}{x!}$

مثال: یک اپراتور تلفن در هر سه دقیقه به طور متوسط پنج تلفن دارد. چقدر احتمال دارد در یک دقیقه به تصادف انتخاب شده ، هیچ تماس تلفنی نداشته باشد؟

اگر در هر سه دقیقه پنج تلفن داشته باشد $X\sim P(5)$ ، بنابراین برای یک دقیقه خواهیم داشت

$$X \sim P(\frac{5}{3})$$

حال احتمال اینکه در یک دقیقه هیچ تلفنی نداشته باشیم:

$$P(X=0)=\frac{e^{-\frac{5}{3}}(\frac{5}{3})^{0}}{0!}=e^{-\frac{5}{3}}$$

توزیع خاص پیوسته

توزیع نرمال

توزیع نرمال برای تقریب احتمال متغیرهای تصادفی دو جمله‌ای وقتی که پارامتر دوجمله‌ای n بزرگ باشد به کار می‌رود.

واریانس	امیدریاضی	تابع چگالی احتمال
$Var(x)=\sigma^2$	$E(x)=\mu$	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

توزیع تی استیودنت

برای بررسی این نکته که میانگین نمونه‌های برداشته شده از یک متغیر تصادفی تا چه حد به میزان «واقعی» (که آزمایشگر نمی‌داند) نزدیک است از تست تی-استیودنت استفاده می‌شود.

واریانس	امیدریاضی	تابع چگالی احتمال
$Var(x)=\frac{n}{n-2}$	$E(x)=0$	$f(x)=\frac{1}{\sqrt{n}\beta(\frac{n}{2},\frac{1}{2})}(1+\frac{t^2}{n})^{-\frac{n+1}{2}}$

توزیع کای دو

توزیعی است که از مجموعه‌ای از داده‌های نرمال به تصادف متغیرهایی را انتخاب می‌کنیم و به توان دو می‌رسانیم .

تابع چگالی احتمال	امیدریاضی	واریانس
$f(x) = \frac{1}{(\frac{v}{2})2^{\frac{v}{2}}} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}$	$E(x)=v$	$V(x)=2v$

توزیع لجستیک

توزیع متغیر تصادفی لجستیک، شبیه توزیع نرمال است ولی در دم‌ها، میزان احتمال بیشتر از توزیع نرمال است. بنابراین می‌توان توزیع لجستیک را در گروه توزیع‌های دم-زخیم در نظر گرفت یکی از رایج‌ترین کاربردهای توزیع لجستیک و متغیر تصادفی آن، رگرسیون لجستیک است که برای مدل‌سازی متغیرهای وابسته از نوع طبقه‌ای (مثلاً گزینه‌های بله-خیر) مورد استفاده قرار می‌گیرد.

تابع چگالی احتمال	امید ریاضی	واریانس
$f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s\left(1+e^{-\frac{(x-\mu)}{s}}\right)^2}$	$E(x) = \mu$	$Var(x) = \frac{s^2\pi^2}{3}$

ویدیوهای کامل دوره را در کانال آپارات علم داده در آدرس زیر دنبال کنید:

https://www.aparat.com/elmedade