Istatistik - Ders 1

Bu notlar makine ogrenimi, veri madenciligi gibi konularda gerekli olasilik ve istatistik bilgisini paylasmak icin hazirlaniyor. Notlarda olasilik ve istatistik ayni anda anlatilacak, ve uygulamalara agirlik verilecek.

Orneklem Uzayi (Sample Space)

Orneklem uzayi Ω bir deneyin mumkun tum olasiliksal sonuclarin (outcome) kumesidir. Eger deneyimiz ardi ardina iki kere yazi (T) tura (H) atip sonucu kaydetmek ise, bu deneyin mumkun tum sonuclari soyledir

$\Omega = \{HH, HT, TH, TT\}$

Sonuclar ve Olaylar (Outcomes and Events)

Ω icindeki her nokta bir sonuctur (outcome). Olaylar Ω'nin herhangi bir alt kumesidir ve sonuclardan olusurlar. Mesela ustteki yazi-tura deneyinde "iki atisin icinden ilk atisin her zaman H gelmesi olayi" boyle bir alt kumedir, bu olaya A diyelim, $A = \{HH, HT\}$.

Ya da bir deneyin sonucu ω fiziksel bir olcum , diyelin ki sicaklik olcumu. Sicaklik $\pm$, reel bir sayi olduguna gore, $\Omega = (-\infty, +\infty)$, ve sicaklik olcumunun 10'dan buyuk ama 23'ten kucuk ya da esit olma "olayi" $A = (10, 23]$. Koseli parantez kullanildi cunku sinir degerini dahil ediyoruz.

Ornek

10 kere yazi-tura at. $A =$ "en az bir tura gelme" olayi olsun. T_j ise j'inci yazi-tura atisinda yazi gelme olayi olsun. $P(A)$ nedir?

Bunun hesabi icin en kolayi, hic tura gelmeme, yani tamamen yazi gelme olasiligini, A^c'yi hesaplamak, ve onu 1'den cikartmaktir. ^{c} sembolu "tamamlayici (complement)" kelimesinden geliyor.

$P(A) = 1 - P(A^c)$

$= 1 - P(hepsi\ yazi)$

$= 1 - P(T_1)P(T_2)...P(T_{10})$

$= 1 - \left(\frac{1}{2}\right)^{10} \approx .999$

Rasgele Degiskenler (Random Variables)

Bir rasgele degisken X bir eslemedir, ki bu esleme $X : \Omega \to \Re$ her sonuc ile bir reel sayi arasindaki eslemedir.

Olasilik derslerinde bir noktadan sonra artik ornekleme uzayindan bahsedilmez, ama bu kavramin arkalarda bir yerde her zaman devrede oldugunu hic aklimizdan cikartmayalim.

Ornek

10 kere yazi-tura attik diyelim. VE yine diyelim ki $X(\omega)$ rasgele degiskeni her ω siralamasinda (sequence) olan tura sayisi. Iste bir esleme. Mesela eger $\omega = HHTHHTHHTT$ ise $X(\omega) = 6$. Tura sayisi eslemesi ω sonucunu 6 sayisina esledi.

Ornek

$\Omega = \{(x, y); x^2 + y^2 \leq 1\}$, yani kume birim cember ve icindeki reel sayilar (unit disc). Diyelim ki bu kumeden rasgele secim yapiyoruz. Tipik bir sonuc $\omega = (x, y)$'dir. Tipik rasgele degiskenler ise $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$ olabilir. Goruldugu gibi bir sonuc ile reel sayi arasinda esleme var. X rasgele degiskeni bir sonucu x'e eslemis, yani (x, y) icinden sadece x'i cekip cikartmis. Benzer sekilde Y, Z degiskenleri var.

Toplamsal Dagilim Fonksiyonu (Cumulative Distribution Function -CDF-)

Tanim

X rasgele degiskeninin CDF'i $F_X : \Re \to [0, 1]$ tanimi

$F_X(x) = P(X \geq x)$

Eger X ayriksal ise, yani sayilabilir bir kume $\{x_1, x_2, ...\}$ icinden degerler aliyorsa olasilik fonksiyonu (probability function), ya da olasilik kutle fonksiyonu (probability mass function -PMF-)

$f_X(x) = P(X = x)$

Bazen f_X, ve F_X yerine sadece f ve F yazariz.

Tanim

Eger X surekli (continuous) ise, yani tum x'ler icin $f_X(x) > 0$, $\int_{-\infty}^{+\infty} f(x)dx =$

1 olacak sekilde bir f_X mevcut ise, o zaman her $a \leq b$ icin

$$P(a < X < b) = \int_a^b f_X(x)dx$$

Bu durumda f_X olasilik yogunluk fonksiyonudur (probability density function -PDF-).

$$F_X = \int_\infty^x f_X(t)dt$$

Ayrica $F_X(x)$'in turevi alinabildigi her x noktasinda $f_X(x) = F'_X(x)$ demektir.

Dikkat! Eger X surekli ise o zaman $P(X = x) = 0$ degerindedir. $f(x)$ fonksiyonunu $P(X = x)$ olarak gormek hatalidir. Bu sadece ayriksal rasgele degiskeninler icin isler. Surekli durumda olasilik hesabi icin belli iki nokta arasinda entegral hesabi yapmamiz gereklidir. Ek olarak PDF 1'den buyuk olabilir, ama PMF olamaz. PDF'in 1'den buyuk olabilmesi entegrali bozmaz mi? Unutmayalim, entegral hesabi yapiyoruz, noktasal degerlerin 1 olmasi tum 1'lerin toplandigi anlamina gelmez. Bakiniz *Entegralleri Nasil Dusunelim* yazimiz.

Tanim

X rasgele degiskeninin CDF'i F olsun. Ters CDF (inverse cdf), ya da ceyrek fonksiyonu (quantile function)

$$F^{-1}(q) = inf\Big\{x : F(x) \leq q\Big\}$$

ki $q \in [0, 1]$. Eger F kesinlikle artan ve surekli bir fonksiyon ise $F^{-1}(q)$ tekil bir x sayisi ortaya cikarir, ki $F(x) = q$.

Eger inf kavramini bilmiyorsak simdilik onu minimum olarak dusunebiliriz.

$F^{-1}(1/4)$ birinci ceyrek

$F^{-1}(1/2)$ medyan (median, ya da ikinci ceyrek),

$F^{-1}(3/4)$ ucuncu ceyrek

olarak bilinir.

Iki rasgele degisken X ve Y dagilimsal olarak birbirine esitligi, yani $X \stackrel{d}{=} Y$ eger $F_X(x) = F_Y(x)$, $\forall x$. Bu X, Y birbirine esit, birbirinin aynisi demek

degildir. Bu degiskenler hakkindaki tum olasiliksal islemler, sonuclar ayni olacak demektir.

Uyari! "X'in dagilimi F'tir" beyanini $X \sim F$ seklinde yazmak bir gelenek. Bu biraz kotu bir gelenek aslinda cunku $\sim$ sembolu ayni zamanda yaklasiksallik kavramini belirtmek icin de kullaniliyor.

Bernoulli Dagilimi

X'in bir yazi-tura atisini temsil ettigini dusunelim. O zaman $P(X = 1) = p$, ve $P(X = 0) = 1 - p$ olacaktir, ki $p \in [0, 1]$ olmak uzere. O zaman X'in dagilimi Bernoulli deriz, ve $X \sim Bernoulli(p)$ diye gosteririz. Olasilik fonksiyonu $f(x) = p^x(1-p)^{(1-x)}$, $x \in \{0, 1\}$.

Yani x ya 0, ya da 1. Parametre p, 0 ile 1 arasindaki herhangi bir reel sayi.

Uyari!

X bir rasgele degisken; x bu degiskenin alabilecegi spesifik bir deger; p degeri ise bir **parametre**, yani sabit, onceden belirlenmis reel sayi. Tabii istatistiki problemlerde (olasilik problemlerinin tersi olarak dusunursek) cogunlukla o sabit parametre bilinmez, onun veriden hesaplanmasi, kestirilmesi gerekir. Her halukarda, cogu istatistiki modelde rasgele degiskenler vardir, ve onlardan ayri olarak parametreler vardir. Bu iki kavrami birbiriyle karistirmayalim.

Normal (Gaussian) Dagilim

$X \sim N(\mu, \sigma^2)$ ve PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\},\ x \in \Re$$

ki $\mu \in \Re$ ve $\sigma > 0$ olacak sekilde.

Ileride gorecegiz ki μ bu dagilimin "ortasi", ve σ onun etrafa ne kadar "yayildigi" (spread). Normal dagilim olasilik ve istatistikte cok onemli bir rol oynar. Dogadaki pek cok olay yaklasiksal olarak Normal dagilima sahiptir. Sonra gorecegimiz uzere, mesela bir rasgele degiskenin degerlerinin toplami her zaman Normal dagilima yaklasir (Merkezi Limit Teorisi -Central Limit Theorem-).

Eger $\mu = 0$ ve $\sigma = 1$ ise X'in standart Normal dagilim oldugunu soyleriz.

Gelenege gore standart Normal dagilim rasgele degiskeni Z ile gosterilmelidir, PDF ve CDF $\phi(z)$ ve $\Phi(z)$ olarak gosterilir.

$\Phi(z)$'nin kapali form (closed-form) tanimi yoktur. Bu, matematikte "analitik bir forma sahip degil" demektir, formulu bulunamamaktadir, bunun sebebi ise Normal PDF'in entegralinin analitik olarak alinamiyor olusudur.

Bazi faydali puf noktalari

1. Eger $X \sim N(\mu, \sigma^2)$ ise, o zaman $Z = (X - \mu)/\sigma \sim N(0, 1)$.

2. Eger $Z \sim N(0, 1)$ ise, o zaman $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$

3. Eger $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, ...$ ve her X_i digerlerinden bagimsiz ise, o zaman

$$\sum_{i=1}^{n} X_i = N\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma^2\Big)$$

Tekrar $X \sim N(\mu, \sigma^2)$ alirsak ve 1. kuraldan devam edersek / temel alirsak su da dogru olacaktir.

$P(a < X < b) =?$

$$= P\Big(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\Big)$$

$$= P\Big(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\Big) = \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big)$$

Ilk gecisi nasil elde ettik? Bir olasilik ifadesi $P(\cdot)$ icinde esitligin iki tarafina ayni anda ayni toplama, cikarma operasyonlarini yapabiliriz.

Son ifadenin anlami sudur. Eger standart Normal'in CDF'ini hesaplayabiliyorsak, istedigimiz Normal olasilik hesabini yapabiliriz demektir, cunku artik X iceren bir hesabin Z'ye nasil tercume edildigini goruyoruz.

Tum istatistik yazilimlari $\Phi(z)$ ve $\Phi(z)^{-1}$ hesabi icin gerekli rutinlere sahiptir. Tum istatistik kitaplarinda $\Phi(z)$'nin belli degerlerini tasiyan bir tablo vardir.