

Istatistik - Ders 1

Bu notlar makine öğrenimi, veri madenciliği gibi konularda gerekli olasılık ve istatistik bilgisini paylaşmak için hazırlanıyor. Notlarda olasılık ve istatistik aynı anda anlatılacak, ve uygulamalara ağırlık verilecek.

Orneklem Uzayı (Sample Space)

Orneklem uzayı Ω bir deneyin mümkün tüm olasılıksal sonuçların (outcome) kümesidir. Eğer deneyimiz ardi ardına iki kere yazi (T) tura (H) atıp sonucu kaydetmek ise, bu deneyin mümkün tüm sonuçları şöyledir

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

Sonuçlar ve Olaylar (Outcomes and Events)

Ω içindeki her nokta bir sonuctur (outcome). Olaylar Ω 'nin herhangi bir alt kümesidir ve sonuçlardan oluşurlar. Mesela üstteki yazi-tura deneyinde “iki atisin icinden ilk atisin her zaman H gelmesi olayi” böyle bir alt kümedir, bu olaya A diyelim, $A = \{HH, HT\}$.

Ya da bir deneyin sonucu ω fiziksel bir ölçüm , diyelin ki sıcaklık ölçümü. Sıcaklık \pm , reel bir sayı olduğuna göre, $\Omega = (-\infty, +\infty)$, ve sıcaklık ölçümünün 10'dan büyük ama 23'ten küçük ya da esit olma “olayı” $A = (10, 23]$. 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 için en kolayi, hic tura gelmeme, yani tamamen yazi gelme olasiligini, A^c 'yi hesaplamak, ve onu 1'den cikartmaktır. c sembolu “tamamlayici (complement)” kelimesinden geliyor.

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - P(\text{hepsi yazi}) \\ &= 1 - P(T_1)P(T_2)\dots P(T_{10}) \\ &= 1 - \left(\frac{1}{2}\right)^{10} \approx .999 \end{aligned}$$

Rasgele Degiskenler (Random Variables)

Bir rasgele degisken X bir eslemedir, ki bu esleme $X : \Omega \rightarrow \Re$ 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 yapıyoruz. 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 $F_X : \Re \rightarrow [0, 1]$ tanimi

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

Eger X ayriskal ise, yani sayilabilir bir kume $\{x_1, x_2, \dots\}$ 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 integral hesabi yapmamiz gereklidir. Ek olarak PDF 1'den buyuk olabilir, ama PMF olamaz. PDF'in 1'den buyuk olabilmesi integrali bozmaz mi? Unutmayalim, integral hesabi yapiyoruz, noktasal degerlerin 1 olması 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)

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

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 eqind Y eger $F_X(x) = F_Y(x)$, $\forall x$. Bu X, Y birbirine esit, birbirinin aynisi demek

degildir. Bu degiskenler hakkındaki 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 yaklasik-sallik 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 $X \sim \text{Bernoulli}(p)$ 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

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

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.

Geleneye göre standart Normal dağılım rasgele değişkeni Z ile gösterilmelidir, PDF ve CDF $\phi(z)$ ve $\Phi(z)$ olarak gösterilir.

$\Phi(z)$ 'nin kapalı form (closed-form) tanımı yoktur. Bu, matematikte “analitik bir forma sahip değil” demektir, formülü bulunamamaktadır, bunun sebebi ise Normal PDF’in integralinin analitik olarak alınamıyor olmasıdır.

Bazı faydalı özellik noktaları

1. Eğer $X \sim N(\mu, \sigma^2)$ ise, o zaman $Z = (X - \mu)/\sigma \sim N(0, 1)$.
2. Eğer $Z \sim N(0, 1)$ ise, o zaman $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$
3. Eğer $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots$ ve her X_i diğerlerinden bağımsız ise, o zaman

$$\sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

Tekrar $X \sim N(\mu, \sigma^2)$ alırsak ve 1. kuraldan devam edersek / temel alırsak şu da doğru olacaktır.

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

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

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

İlk geçisi nasıl elde ettik? Bir olasılık ifadesi $P(\cdot)$ içinde eşitliğin iki tarafına aynı anda aynı toplama, çıkarma operasyonlarını yapabiliriz.

Son ifadenin anlamı şudur. Eğer standart Normal’in CDF’ini hesaplayabiliyorsak, istediğimiz Normal olasılık hesabını yapabiliriz demektir, çünkü artık X içeren bir hesabın Z ’ye nasıl tercüme edildiğini görüyoruz.

Tüm istatistik yazılımları $\Phi(z)$ ve $\Phi(z)^{-1}$ hesabi için gerekli rutinlere sahiptir. Tüm istatistik kitaplarında $\Phi(z)$ ’nin belli değerlerini taşıyan bir tablo vardır.