Gaussian Kontrolu

Diyelim ki Gaussian dagilimina sahip oldugunu dusundugumuz $\{x_i\}$ verilerimiz var. Bu verilerin Gaussian dagilimina uyup uymadigini nasil kontrol edecegiz? Normal bir dagilimin her veri noktasi icin soyle temsil edebiliriz,

$$y_i = \Phi\bigg(\frac{x_i - \mu}{\sigma}\bigg)$$

Burada Φ standart Gaussian'i temsil ediyor (detaylar icin *Istatistik Ders 1*) ve CDF fonksiyonuna tekabul ediyor. CDF fonksiyonunun ayni zamanda ceyregi (quantile) hesapladigi soylenir, aslinda CDF son derece detayli bir olasilik degeri verir fakat evet, dolayli yoldan noktanin hangi ceyrek icine dustugu de gorulecektir.

Simdi bir numara yapalim, iki tarafa ters Gaussian formulunu uygulayalim, yani Φ^{-1}.

$$\Phi^{-1}(y_i) = \Phi^{-1}\bigg(\Phi\bigg(\frac{x_i - \mu}{\sigma}\bigg)\bigg)$$

$$\Phi^{-1}(y_i) = \frac{x_i - \mu}{\sigma}$$

$$x_i = \Phi^{-1}(y_i)\sigma + \mu$$

Bu demektir ki elimizdeki verileri $\Phi^{-1}(y_i)$ bazinda grafiklersek, bu noktalar egimi σ, baslangici (intercept) μ olan bir duz cizgi olmalidir. Eger kabaca noktalar duz cizgi olusturmuyorsa, verimizin Gaussian dagilima sahip olmadigina karar verebiliriz.

Ustte tarif edilen grafik, olasilik grafigi (probability plot) olarak bilinir.

Ters Gaussian teorik fonksiyonunu burada vermeyecegiz, Scipy `scipy.stats.invgauss` hesaplar icin kullanilabilir. Fakat y_i'nin kendisi nereden geliyor? Eger y_i, CDF'in bir sonucu ise, pur veriye bakarak bir CDF degeri de hesaplayabilmemiz gerekir. Bunu yapmak icin bir baska numara lazim.

1. Eldeki sayilari artan sekilde siralayin

2. Her veri noktasina bir derece (rank) atayin (siralama sonrasi hangi seviyede oldugu yeterli, 1'den baslayarak).

3. Ceyrek degeri y_i bu sira / $n+1$, n eldeki verinin buyuklugu.

Bu teknik niye isliyor? x'in CDF'i $x_i < x$ sartina uyan x_i'lerin orani degil midir? Yani bir siralama soz konusu ve ustteki teknik te bu siralamayi biz elle yapmis olduk, ve bu siralamadan gereken bilgiyi aldik.

Ozet Istatistikleri

Genellikle istatistik kitaplari hemen ortalama (mean), medyan (median) ve baglantili ozet istatistiklerinden (summary statistics) bahsederek ise girerler. Bu istatistikleri dikkatle kullanmak gerekir, cunku her turlu veri, her yerde gecerli degildirler. Mesela ortalama sadece tek merkezi bir tepesi olan (unimodal) dagilimlar icin gecerlidir. Eger bu temel varsayim gecerli degilse, ortalama kullanarak yapilan hesaplar bizi yanlis yollara goturur. Bu uyaridan sonra ortalama ve standart sapmayi (standart deviation) gorelim.

$$m = \frac{1}{n}\sum_i x_i$$

Standart sapma veri noktalarin "ortalamadan farkinin ortalamasini" verir. Tabii bazen noktalar ortalamanin altinda, bazen ustunde olacaktir, bizi bu negatiflik, pozitiflik ilgilendirmez, biz sadece farkla alakaliyiz. O yuzden her sapmanin karesini aliriz, bunlari toplayip nokta sayisina boleriz .

$$s^2 = \frac{1}{n}\sum_i (x_i - m)^2$$

Eger m tanimini ustte yerine koyarsak,

$$= \frac{1}{n}\sum_i x_i^2 + \frac{1}{n}\sum_i m^2 - \frac{2}{n}\sum_i x_i m$$

$$= \frac{1}{n}\sum_i x_i^2 + \frac{m^2 n}{n} - \frac{2mn}{n} m$$

$$= \frac{1}{n}\sum_i x_i^2 + m^2 - 2m^2$$

$$= \frac{1}{n}\sum_i x_i^2 - m^2$$

Bu olcuye varyans (variance) denir ve teorik olarak ortalamadan daha onemli oldugu soylenebilir. Fakat dagilimin yayilma olcusu olarak biz bu olcuyu oldugu gibi degil, onun karesini kullanacagiz (ki standart sapma buna deniyor aslinda). Niye? Cunku o zaman veri noktalarinin ve yayilma olcusunun birimleri birbiri ile ayni olacak. Eger veri setimiz bir alisveris sepetindeki malzemelerin lira cinsinden degerleri olsaydi, varyans bize sonucu "karekok lira" olarak verecekti ve bunun pek anlami olmayacakti.