MCMC, Degisim Nokta Hesabi, Gibbs Orneklemesi, Bayes Teorisi

Ingiltere'de 1851 ve 1962 yillari arasinda komur madenlerinde olan kazalarin sayisi yillik olarak kayitli (ekteki zip dosyasinda `coal.txt` dosyasinda). Acaba bu kazalarin "oraninin" degisimine bakarak, degisimin oldugu seneyi bulabilir miyiz? Boyle bir degisim ani neyi gosterir? Belki madenlerle alakali regulasyonlarda, denetimlerde bir degisiklik olmustur, ve kaza orani azalmistir.

Bu hesabi yapabilmek icin "degisim noktasi" hesabi (change-point analysis), ve Bayes kurali ile Bayes formullerini hesaplamamizi saglayan Markov Chain Monte Carlo (MCMC) teknigine bakacagiz. Kazalarin sayisinin tumunu iki Poisson dagiliminin birlesimi (joint distribution) uzerinden modelleyecegiz, ve bu dagilimlarin birinci Poisson'dan ikincisine gectigi ani hesaplamaya ugrasacagiz.

Once Bayes, dagilimlar konusuna bir bakalim:

Poisson dagilimi

$$p(y|\theta) = \frac{e^{-\theta}\theta^y}{y!}$$

Eldeki n tane veri noktasi $y = y_0, y_1, ..., y_n$'nin hep birlikte θ ile tanimli bir Poisson dagilimindan gelip gelmediginin ne kadar mumkun oldugu (likelihood) hesabi soyledir:

$$p(y|\theta) = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i!}$$

Formulun bolunen kismindaki tum y noktalari toplaniyor, bolen kisminde ise tum y degerleri teker teker faktoryel hesabi sonrasi birbiri ile carpiliyor.

Simdi yukaridaki θ degiskeni de noktasal bir deger yerine bir "dagilima", mesela θ Gamma dagilimina sahip olabilirdi: $Gamma(\alpha, \beta)$. Formulde α, β sabit degerlerdir (fonksiyon degiskeni degil). Gamma olasilik formulu soyledir:

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

O zaman $p(y|\theta)$ formulunu bulmak icin Bayes teorisini kullanmamiz gerekecekti. Bayes teorisi bilindigi gibi

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Ikinci formule dikkat, esitlik yerine orantili olma (proportional to) isaretini kullaniyor. Sebep: bolen kismindaki p(y)'yi kaldirdik, sonuc olarak soldaki $p(\theta|y)$ degeri artik bir dagilim degil – bu bir bakimdan onemli ama ornekleme amaci icin bir fark yaratmiyor, basitlestirme amaciyla bunu yaptik, boylece $p(y)$'yi hesaplamamiz gerekmeyecek, ama ornekleme uzerinden diger tum hesaplari hala yapabiliriz. Tamam.

Simdi Bayes Teorisini Gamma oncul (apriori) ve Poisson mumkunlugu (likelihood) uzerinden kullanirsak,

$$p(\theta|y) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \times \frac{e^{-n\theta}\theta^{\sum y}}{\prod y!}$$

Benzer terimleri yanyana getirelim:

$$p(\theta|y) = \frac{\beta^\alpha}{\Gamma(\alpha)\prod y!}\theta^{\alpha-1}\theta^{\sum y}e^{-\beta\theta}e^{-n\theta}$$

Simdi sol taraftaki bolumu atalim; yine usttekine benzer numara, bu kisim gidince geri galan dagilim olamayacak, ama ona "oranli" baska bir formul olacak.

$$p(\theta|y) \propto \theta^{\alpha-1}\theta^{\sum y}e^{-\beta\theta}e^{-n\theta}$$

$$\propto \theta^{\alpha-1+\sum y}e^{-(\beta+n)\theta}$$

Bu dagilim nedir? Formulun sag tarafi Gamma dagiliminin formulune benzemiyor mu? Evet, formulun sag tarafi $Gamma(\alpha+\sum y, \beta+n)$ dagilimi, yani ona orantili olan bir formul. Yani Bayes teorisi uzerinden sunu anlamis olduk; eger oncul dagilim Gamma ise, Poisson mumkunluk bizi tekrar Gamma sonuc dagilimina goturuyor. Gamma'dan baslayinca tekrar Gamma'ya ulasiyoruz. Bu bir rahatlik, bir kolaylik, bir matematiksel numara olarak kullanilabilir. Sonuc (posterior) dagilimlarin sekli, hesaplanma, cebirsel islemler acisindan onemli, eger temiz, kisa, oz olurlarsa hesap islerimiz kolaylasir.

Not: Hatta uzerinde calistigimiz problem sebebiyle eger Poisson mumkunluk olacagini biliyorsak, sadece bu sebeple bile oncul dagilimi, ustteki kolaylik bilindigi icin, ozellikle Gamma secebiliriz, cunku biliriz ki Gamma ile baslarsak elimize tekrar Gamma gececektir.

Simdi komur madeni verisine gelelim. Bu madendeki kazalarin sayisinin Poisson dagilimindan geldigini one suruyoruz, ve kazalarin "iki turlu" oldugunu bildigimizden hareketle, birinci tur kazalarin ikinci tur kazalardan degisik Poisson parametresi kullandigini one surecegiz.

O zaman degisim anini, degisim senesini nasil hesaplariz?

Kazalarin ilk k senede ortalama θ ile, ve k ve n arasindaki senelerde ortalama λ Poisson ile dagildigini soyleyelim: Yani

$Y_i = Poisson(\theta) \quad i = 1, .., k$

$Y_i = Poisson(\lambda) \quad i = k+1, .., n$

Burada Y_i sene i sirasinda olan kazalarin sayisini belirtiyor. Bayes kuralini hatirlarsak θ ve λ parametrelerine oncul dagilim atayacagiz. Bu dagilim Gamma olacak. Yani $\theta \sim Gamma(a_1, b_1)$ ve $\lambda \sim Gamma(a_2, b_2)$.

Ayrica k degerini de bilmiyoruz, k degeri yani "degisim noktasi" Poisson dagilimlarin birinden otekine gectigi andir. Bu seneyi bulmaya calisiyoruz. Simdi tum verinin, tum seneleri kapsayacak sekilde modelini kurmaya baslayalim. k parametresinin aynen oteki parametreler gibi bir oncul dagilimi olacak (ki sonradan elimize k icin de bir sonuc dagilimi gececek), ama bu parametre elimizdeki 112 senenin herhangi birinde "esit olasilikta" olabilecegi icin onun oncul dagilimi Gamma degil $k \sim Unif(1, 112)$ olacak. Yani ilk basta her senenin olasiligi birbiriyle esit, her sene $\frac{1}{112}$ olasilik degeri tasiyor.

Bu modelin tamaminin mumkunlugu nedir?

$$L(\theta, \lambda, k|y) = \frac{1}{112} \times \prod_{i=1}^{k} \frac{e^{-\theta}\theta^{y_i}}{y_i!} \times \prod_{i=k+1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

Eger sonuc (posterior) gecisini yapinca yukarida oldugu gibi Gamma dagilimlarini elde ederiz:

$$L(\theta, \lambda, k|y) \propto \theta^{a_1 - 1 + \sum_{i=1}^{k} y_i} e^{-(b_1+k)\theta} \lambda^{a_2 - 1 + \sum_{i=k+1}^{n} y_i} e^{-(b_2+n-k)\lambda}$$

$\frac{1}{112}$'yi bir sabit oldugu icin formulden attik, bu durum orantili hali etkilemiyor. Ustteki formul icindeki Gamma dagilimlarini gorebiliyoruz, hemen yerlerine koyalim:

$$L(\theta, \lambda, k|y) \propto Gamma(a_1 + \sum_{i=1}^{k} y_i, b_1 + k) \ Gamma(a_2 + \sum_{i=k+1}^{n} y_i, b_2 + n - k)$$

Gibbs orneklemeye gelelim. Bu orneklemeye gore sartsal dagilim (conditional distribution) formulu bulunmaya ugrasilir, hangi degiskenlerin verili olduguna gore, o degiskenler sabit kabul edilebilir, ve orantisal formulden atilabilir. Bu her degisken icin teker teker yapilir.

Sorna hesap sirasinda her sartsal dagilima teker teker zar attirilir, ve elde edilen deger, bu sefer diger sartsal dagilimlara deger olarak gecilir. Bu islem sonuca erisilinceye kadar ozyineli (iterative) olarak tekrar edilir (mesela 1000 kere). O zaman,

$$\theta|Y_1, .., Y_n, k \sim Gamma(a_1 + \sum_{i=1}^{k} y_i, b_1 + k)$$

$$\lambda|Y_1, .., Y_n, k \sim Gamma(a_2 + \sum_{i=k+1}^{n} y_i, b_2 + n - k)$$

$$p(k|Y_1, .., Y_n) \propto \theta^{\sum_{i=1}^{k} y_i} e^{-k\theta} \lambda^{\sum_{i=k+1}^{n} y_i} e^{k\lambda}$$

En son formulde icinde k olan terimleri tuttuk, gerisini attik. Formul e terimleri birlestirilerek biraz daha basitlestirilebilir:

$$p(k|Y_1, .., Y_n) \propto \theta^{\sum_{i=1}^{k} y_i} \lambda^{\sum_{i=k+1}^{n} y_i} e^{(\lambda - \theta)k}$$

Bir basitlestirme daha soyle olabilir

$$K = \sum_{i=1}^{k} y_i$$

$$\lambda^{\sum_{i=k+1}^{n} y_i} = \lambda^{\sum_{i=1}^{n} y_i - \sum_{i=1}^{k} y_i}$$

Ustel islemlerde eksi isareti, ustel degisken ayrilinca bolum islemine donusur:

$$= \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\lambda^{\sum_{i=1}^{k} y_i}}$$

$$= \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\lambda^{K}}$$

$p(k|Y_1,..,Y_n) \propto \theta^K \frac{\lambda^{\sum_{i=1}^n y_i}}{\lambda^K} e^{(\lambda-\theta)k}$

$= \left(\frac{\theta}{\lambda}\right)^K \lambda^{\sum_{i=1}^n y_i} e^{(\lambda-\theta)k}$

$\lambda^{\sum_{i=1}^n y_i}$ terimi k'ye degil n'ye bagli oldugu icin o da final formulden atilabilir

$p(k|Y_1,..,Y_n) \propto \left(\frac{\theta}{\lambda}\right)^K e^{(\lambda-\theta)k}$

$p(k)$ icin ortaya cikan bu formule bakarsak, elimizde verilen her k degeri icin bir olasilik dondurecek bir formul var. Daha onceki Gamma orneginde formule bakarak elimizde hemen bir Gamma dagilimi oldugunu soyleyebilmistik. Bu kodlama sirasinda isimize yarayacak bir seydi, hesaplama icin bir dagilima "zar attirmamiz" gerekiyor, ve Gamma orneginde hemen Python Numpy kutuphanesindeki random.gamma cagrisina Gamma'dan gelen rasgele sayilar urettirebiliriz. Ustteki formule bakarsak, hangi dagilima zar attiracagiz?

Cevap soyle: $p(k|..)$ pdf fonsiyonundaki k degiskeni $1,..,119$ arasindaki tam sayi degerleri alabilir, o zaman ortada bir ayriksal (discrete) dagilim var demektir. Ve her k noktasi icin olabilecek olasilik degerini ustteki $p(k|..)$ formulune hesaplattirabiliyorsak, ayriksal bir dagilimi her nokta icin ustteki cagri, ve bu sonuclari normalize ederek (vektorun her elemanini vektorun toplamina bolerek) bir dagilim sekline donusturebiliriz. Daha sonra bu "vektorsel dagilim" uzerinden zar attiririz. Python kodundaki `w_choice` ya da R dilindeki `sample` cagrisi bu isi yapar.

Kodlari isletince elimize k = 41 degeri gececek, yani degisim ani 1851+41 = 1892 senesidir.

```
import numpy as np
import math
import random

# samples indexes from a sequence of probability table
# based on those probabilities
def w_choice(lst):
    n = random.uniform(0, 1)
    for item, weight in enumerate(lst):
        if n < weight:
            break
```

```
        n = n - weight
    return item

#
# hyperparameters: a1, a2, b1, b2
#
def coal(n,x,init,a1,a2,b1,b2):
    nn=len(x)
    theta=init[0]
    lam=init[1]
    k = init[2]
    z=np.zeros((nn,))
    for i in range(n):
        ca = a1 + sum(x[0:k])
        theta = np.random.gamma(ca, 1/float(k + b1), 1)
        ca = a2 + sum(x[(k+1):nn])
        lam = np.random.gamma(ca, 1/float(nn-k + b2), 1)
        for j in range(nn):
            z[j]=math.exp((lam-theta)*(j+1)) * (theta/lam)**sum(x[0:j])
        # sample
        zz = z / sum(z)
        k = w_choice(zz)
        print float(theta), float(lam), float(k)

if __name__ == "__main__":

    data = np.loadtxt("coal.txt")
    coal(1100, data, init=[1,1,30], a1=1,a2=1,b1=1,b2=1)
```

Kaynaklar:

Ioana A. Cosma and Ludger Evers, Markov Chain Monte Carlo Methods (Lecture)

Charles H. Franklin, Bayesian Models for Social Science Analysis (Lecture)