1906003132015

Doğal Dil İşleme

BAİBÜ Bilgisayar Müh.

Dr. Öğr. Üyesi İsmail Hakkı Parlak

ismail.parlak@ibu.edu.tr

Oda: 335

Markov Modeli

Andrey Andreyevich Markov (1856 - 1922), en çok *stokastik süreçler* üzerine yaptığı çalışmalarla tanınan Rus matematikçidir.

https://tr.wikipedia.org/wiki/Andrey_Markov

Markov Zinciri

- Markov Zinciri, Markov özelliğine sahip bir stokastik süreçtir. Markov özelliğine (Markov property) sahip olmak, mevcut durum verildiğinde, gelecek durumların geçmiş durumlardan bağımsız olması anlamına gelir.
- Mevcut durumun açıklaması, sürecin gelecekteki evrimini etkileyebilecek tüm bilgiyi kapsar. Gelecek durumlara belirli bir şekilde değil, olasılıksal bir süreçle ulaşılacaktır.

https://tr.wikipedia.org/wiki/Markov_zinciri

Koşullu Olasılık

- P(A | B): B'nin gerçekleştiği biliniyorken A'nın gerçekleşme olasılığı.

- A: zarın çift olma olasılığı, B: atılan zar 3'ten büyük gelmiş.

Markov Özelliği

- $p(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1) = p(x_t \mid x_{t-1})$
- x_t, sadece x_{t-1}'e bağlıdır; x_{t-2}, x_{t-3}, vb., bağlı değildir.

Zincir Kuralı

- $p(x_1, x_2, \ldots, x_t) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)\ldots p(x_t|x_{t-1},x_{t-2},..x_1)$

Çıkarım

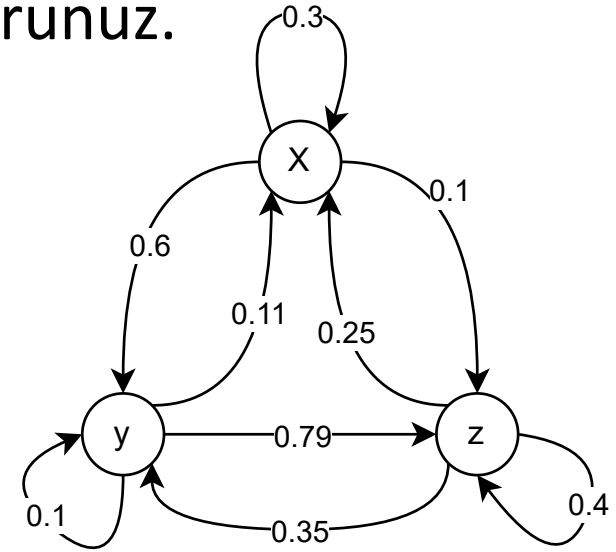
- $p(x_1, x_2, \ldots, x_t) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_t|x_{t-1})$

Notasyon

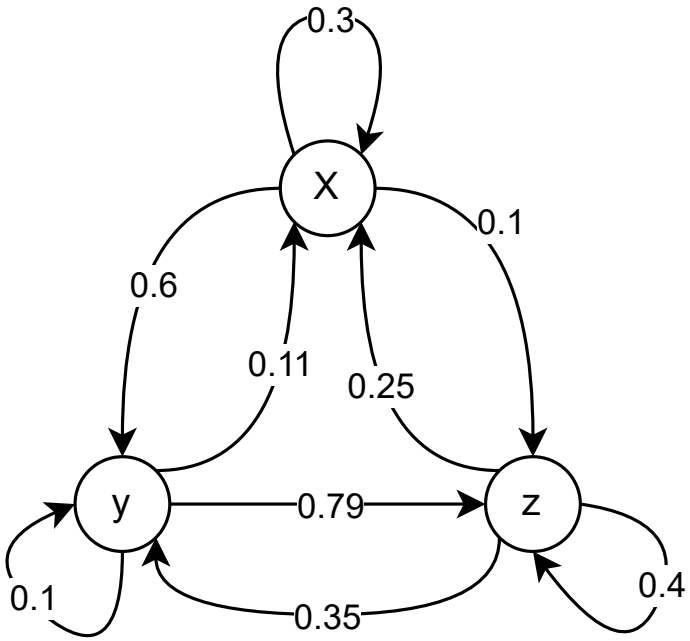
- $s(t) = s_t$ = *t* anındaki durum (state)
- $p(s_t = i)$: *t* anındaki durumun *i* olma olasılığı
- $p(s_{cuma} = güneşli)$: Cuma günü havanın güneşli olma olasılığı nedir?
- $p(s_t = j \mid s_{t-1} = i)$: t-1 anındaki durumun i olduğu bilindiğinde, t anındaki durumun j olma olasılığı.
  - Eğer i ve j 1..M arasındaki değerlere sahip olabiliyorsa bunun gibi kaç tane koşullu olasılık bulunabilir?
  - $p(s_t = 1 \mid s_{t-1} = 1)$, $p(s_t = 2 \mid s_{t-1} = 1)$, …, $p(s_t = M \mid s_{t-1} = 1)$, $p(s_t = 1 \mid s_{t-1} = 2)$, $p(s_t = 2 \mid s_{t-1} = 2)$, …, $p(s_t = M \mid s_{t-1} = 2)$, ..., $p(s_t = 1 \mid s_{t-1} = M)$, $p(s_t = 2 \mid s_{t-1} = M)$, …, $p(s_t = M \mid s_{t-1} = M) \rightarrow M^2$

Durum Geçiş Matrisi

- A_{MxM}: durum geçiş matrisi.
- $A_{ij} = p(s_t = j \mid s_{t-1} = i), \forall i = 1..M, j = 1..M$
- Örn: x, y, z sembollerinden oluşan derlemdeki metinler incelendiğinde aşağıdaki geçiş olasılıkları hesaplandıysa bu sistemi ifade eden durum geçiş matrisini oluşturunuz.



Durum Geçiş Matrisi



x:0, y:1, z:2

A =

	x	y	z
x	0.3	0.6	0.1
y	0.11	0.1	0.79
z	0.25	0.35	0.4

A_{12} = 0.79 = 1'den sonra 2'nin gelme olasılığı = y'den sonra z gelme olasılığı.

İlk Durum Olasılık Dağılımı

- Bir gün yolda yürürken…
  - p(gün | yol), p(yürürken | yolda), …
  - p(bir | ???)

- Sembollerin dizinin en başta bulunma olasılıklarının dağılımını elde etmek için ilk durum olasılık dağılımları vektörünü (π) hesaplamak gerekir.
  - $\pi_i = p(s_1 = i)$
- $A_{ij} = p(s_t = j \mid s_{t-1} = i)$
- $\pi_i = p(s_1 = i)$

Olasılık Dağılımı Hesaplama

- $\pi_i = p(s_1 = i) = i$ sembolünün en başta bulunma olasılığı.
- $\pi_i = \frac{count(s_1=i)}{N}$, N = veri setindeki dizi sayısı.
- $A_{ij} = \frac{count(i \to j)}{count(i)}$
- $s_1\ s_2 \ldots s_T$ dizisinin bir veri setinde bulunma olasılığı $= p(s_1, s_2, \ldots, s_T) = \pi_{s1} \prod_{t=2}^{T} A_{s_{t-1}, s_t}$ 🫣
- Herhangi bir π_i veya A_{ij} değeri 0'sa?

Olasılık Düzenlemesi (Add-One Smoothing)

- $A_{ij} = \frac{count(i \to j)+1}{count(i)+M}$

- $\pi_i = \frac{count(s_1 = i)+1}{N+M}$, N: cümle sayısı; M: len(vocab)

- $p(s_1, s_2, \ldots, s_T) = \pi_{s1} \prod_{t=2}^{T} A_{s_{t-1}, s_t}$
- $\log p(s_1, s_2, \ldots, s_T) = \log \pi_{s1} + \sum_{t=2}^{T} \log A_{s_{t-1}, s_t}$
  log(AB) = log(A) + log(B)

Örnek

- "x y x x y z", "y x y z z y", "x z x z z x", "x y z".
  π = ?, A = ?

1. sözlük = {x:0, y:1, z:2}
2. π = [(3+1)/(4+3), (1+1)/(4+3), (0+1)/(4+3)] = [4/7, 2/7, 1/7]
3. $A_{00} = \frac{count("x\ x")+1}{count(x)+3} = 2/10$, $A_{01} = \frac{count("x\ y")+1}{count(x)+3} = 5/10$, $A_{02} = \frac{count("x\ z")+1}{count(x)+3} = 3/10$,
   $A_{10} = \frac{count("y\ x")+1}{count(y)+3} = 3/8$, $A_{11} = \frac{count("y\ y")+1}{count(y)+3} = 1/8$, $A_{12} = \frac{count("y\ z")+1}{count(y)+3} = 4/8$,
   $A_{20} = \frac{count("z\ x")+1}{count(z)+3} = 3/8$, $A_{21} = \frac{count("z\ y")+1}{count(z)+3} = 2/8$, $A_{22} = \frac{count("z\ x")+1}{count(z)+3} = 3/8$
4. A =

0.2	0.5	0.3
0.375	0.125	0.5
0.375	0.25	0.375

Dil Modeli

- 1. Derece MM: p(st | st-1). Her kelimenin bulunma olasılığı sadece kendinden 1 önceki kelimeye bağlıdır.
  $A^{(1)}_{ij} = p(s_t = j \mid s_{t-1} = i)$
- 2. Derece MM: $p(s_t \mid s_{t-1}, s_{t-2})$. Her kelimenin bulunma olasılığı kendinden önceki 2 kelimeye bağlıdır.
  $A^{(2)}_{ijk} = p(s_t = k \mid s_{t-1} = j, s_{t-2} = i)$
- Dil modeli = $\{\pi, A^{(1)}, A^{(2)}\}$
- Cümle başındaki olasılık dağılımları için π, cümle başındaki 2. kelimelerin olasılık dağımları için $A^{(1)}$, geriye kalan kelimelerin olasılık dağılımları için ise $A^{(2)}$ kullanılır.

Otomatik Cümle Oluşturma

1. π'den olasılık dağılımlarına göre 1. kelime seçilerek cümleye başlanır.
   `np.random.choice(["Y", "T"], [0.3, 0.7])`
2. $A^{(1)}$'den olasılık dağılımlarına göre 2. kelime seçilir.
3. $A^{(2)}$'den olasılık dağılımlarına göre 3. kelime seçilir.
4. $A^{(2)}$'den olasılık dağılımlarına göre 4. kelime seçilir.
5. …
6. Cümleleri bitirmek için . işareti yerine "END" yazabiliriz.