DOĞAL DİL İŞLEMEYE GİRİŞ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BURSA TEKNİK ÜNİVERSİTESİ

DR. ÖĞR. ÜYESİ HAYRİ VOLKAN AGUN

Özet

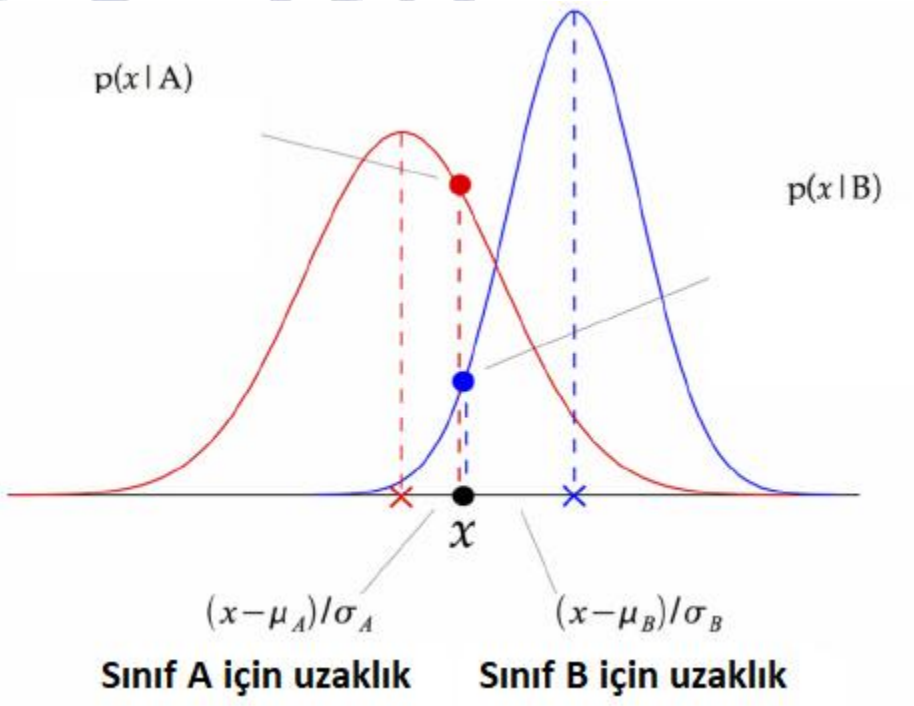
- Üretici sınıflandırma (generative)
- Ayırt edici sınıflandırma (discrimantive)
- Saklı Markov Modeller (hidden Markov models)
- Viterbi Algoritması
- Saklı Markov Model örnekleri

Sınıflandırma Türleri – Üretici (Generative)

- Gözetimli sınıflandırma yöntemleri makine öğrenmesi literatüründe en temel olarak iki ayrı türede ifade edilir.
- Bunlar üretici sınıflandırma ve ayırt edici sınıflandırmadır.
- Üretici sınıflandırma veri üzerinde sonsal (posterior) olasılık değerini hesaplamak için o sınıfa ait olan verinin istatistik analizinden yararlanır.
- İstatiksel analiz için gereken parametreleri girdiye ait veri üzerine yakınsayarak hesaplar.
- Sınıflandırma için bu yakınsanan parametrelerin ayırt edilecek verileri ne ölçüde kapsadığını bulmak için sonsal olasılığı hesaplar.
- Sonsal olasılık hangi sınıf modeli için en yüksek ise o zaman girdi/öğe o sınıftadır.

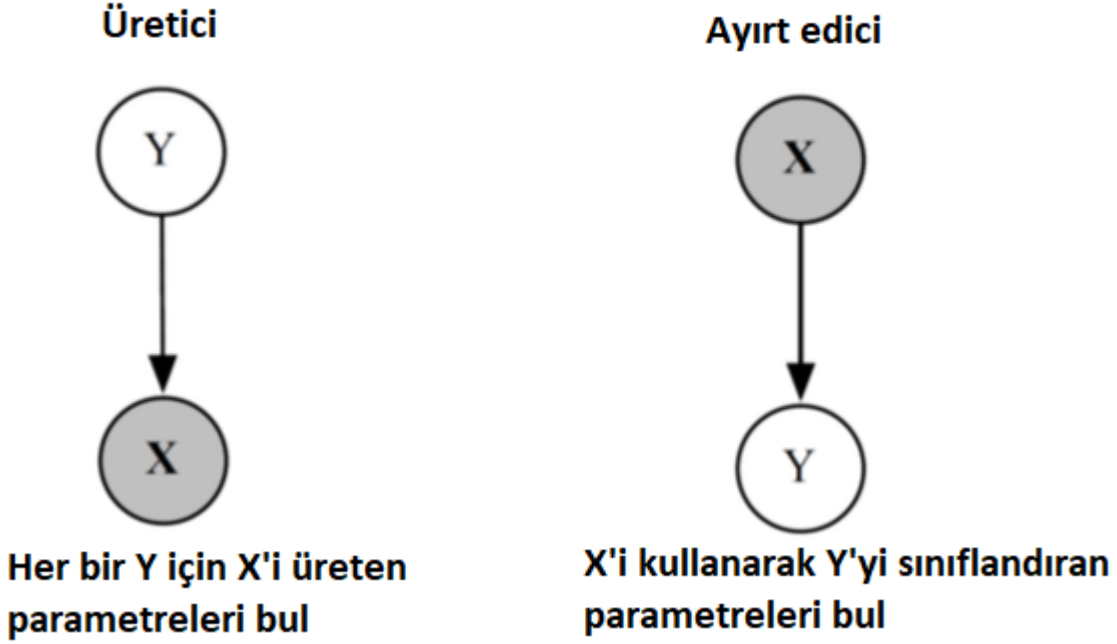
Sınıflandırma Türleri – Üretici (Generative)

- Sınıflandırma yöntemleri makine öğrenmesi literatüründe en temel olarak iki ayrı türede ifade edilir.
- Bunlar üretici sınıflandırma ve ayırt edici sınıflandırmadır.
- Genel yaklaşımda naive Bayes kullanılarak P(x | A) olasılığından P(A | x) hesaplanır.
- Yanda bir X girdi değeri A ve B sınıfları için
  - Elde edilen iki farklı ortalama ve standart sapma parametreleri ve normal dağılım kullanılarak yapılan sınıflandırma yandaki şekilde gösterilmektedir.
  - Sınıflandırma için tek bir sınır yerine iki farklı sınıf için tek bir uzaklık foksiyonu yerine iki farklı uzaklık kullanılmaktadır.
  - Bunun sebebi sizce nedir?



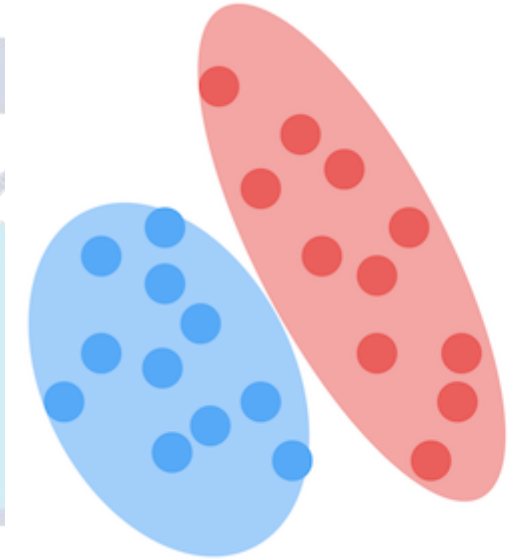
Sınıflandırma Türleri – Ayırt Edici (Discriminative)

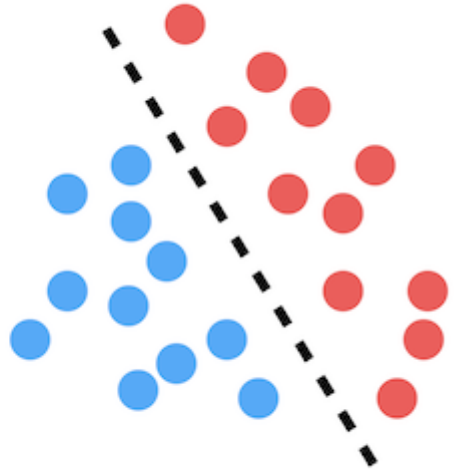
- Ayırt edici sınıflandırmada veri dağılımına bakılmaksızın bir sınır fonksiyonu elde edilmektedir. Kullanılan sınır fonksiyonu girilen bir girdi için doğru sınıfın bulunmasında kullanılır.



Sınıflandırma Türleri – Ayırt Edici (Discriminative)

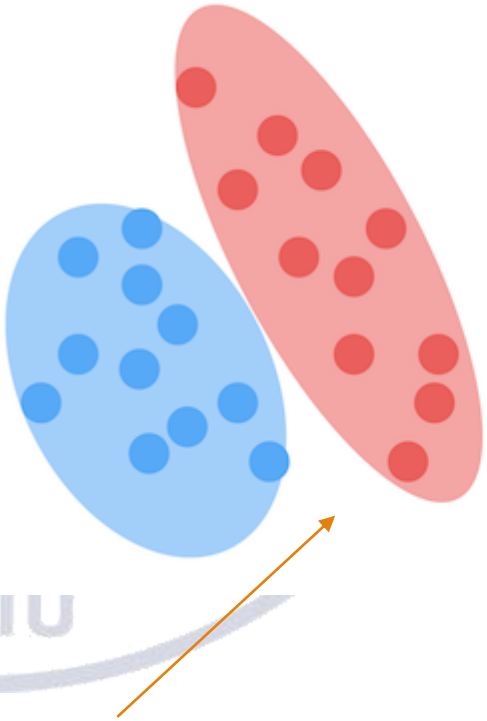
- Ayırt edici sınıflandırmada veri dağılımına bakılmaksızın bir sınır fonksiyonu elde edilmektedir. Kullanılan sınır fonksiyonu girilen bir girdi için doğru sınıfın bulunmasında kullanılır.



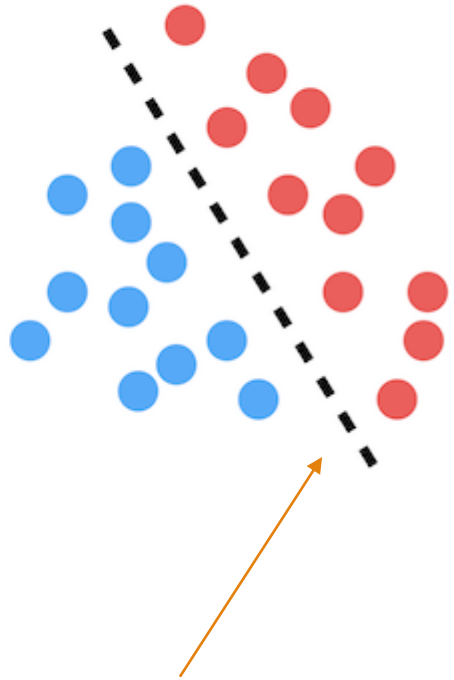


Sınıflandırma Türleri – Ayırt Edici (Discriminative)

- Ayırt edici sınıflandırmada veri dağılımına bakılmaksızın bir sınır fonksiyonu elde edilmektedir. Kullanılan sınır fonksiyonu girilen bir girdi için doğru sınıfın bulunmasında kullanılır.
- Ayırt edici sınıflandırma ve üretici sınıflandırma modellerinde kullanılan matematik birbirine çok benzer olabilir ancak temel bu iki sınıflandırma yöntemi ya girdiyi yada ayırt edici modeli oluşturmada kullanılır.



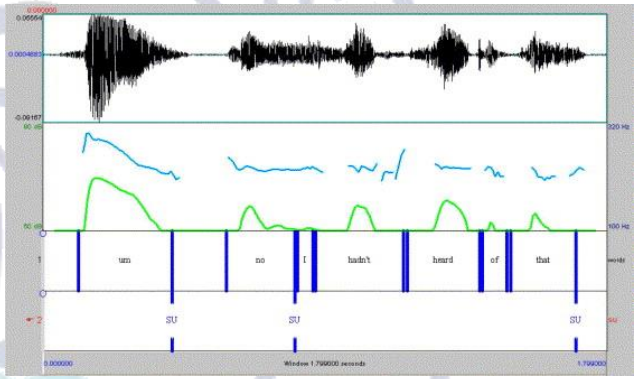
Girdi modellemesi



Sınır modellemesi

Sınıflandırma – Sınır Bulma

- Bir paragraftaki her bir cümlenin başlangıç ve bitişinin bulunması.
- Bir kelime yada bir deyimin başlangıç ve bitişinin bulunması
- Bir metin içerisinde geçen özel isimlerin başlangıç ve bitişlerini bulunması
- Bir metin içerisinde geçen ardışık kelimelerin yada eklerin belirli bir sınıfa ait olma durumunun bulunması.
- Sınır bulma doğal dil işleme dışında en sık ses ve görüntü işlemede kullanılmaktadır.



Good Supply S.A.
Taxpayer Registry: K7B00961L31
Via 2724 Regina Throughway
IT-00000 Roma
Italy

Good Trader Ltd.
6888-6890 Wallermouth Avenue
Estate Suite 00999
Praque,
Czech Republic
Phone 282-5175-0798

Costumer contact: Lisa Williams
E-mail: lisa.williams@email.com

Invoice Number: 173A2-0019

Date: October-04-2019

Sınır Bulma Problemi

Tim Cook eski Apple CEO'su yeniden Apple'ın başına geçti.

- Yukarıdaki örnek için sınıflandırma her bir kelimenin yada her bir karakterin bir özel isim başlangıcı yada bitişi olduğu şeklinde yapılabilir.

Sınır Bulma Problemi

- Sınır bulma probleminde kullanılan ardışık öğeler (kelimeler, heceler, ekler yada karakterler) birbirinden bağımsız değildir. Bu durumda;
- $P(k2 \mid k1) = P(k1 \cap k2) \neq P(k1) * P(k2)$
- Sadece karakterler bağımsız değildir. Sınıf bilgisi verildiğinde bir kelimenin başlangıç karakteri kişi sınıfına ait başlangıcı temsil ediyorsa o zaman bitiş karakterinin de başka bir sınıfa ait olma olasılığı 0 dır. Bu durumda sınıf olasılıkları da birbirine bağlıdır veya biririnden bağımsız olamaz.
- $P(k2, k1|sinif) = P(k2 \cap k1|sinif) \neq P(k1|sinif) * P(k2|sinif)$

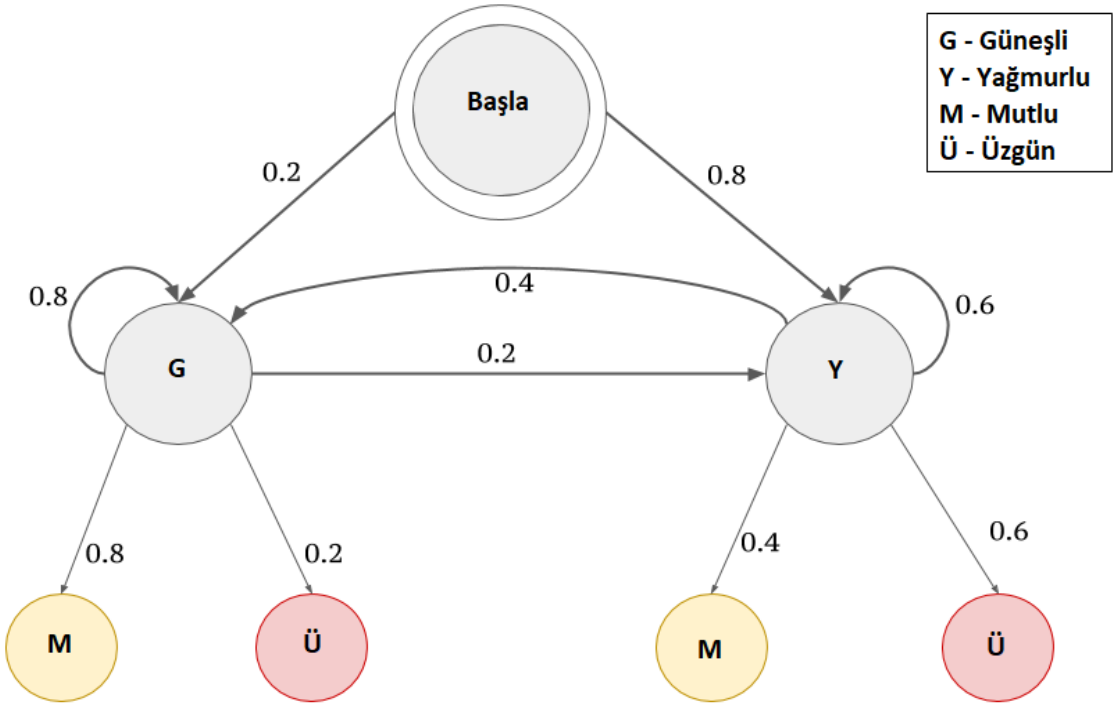
Üretici Sınıflandırma

- Genellikle Normal (Gausian) dağılım kullanılarak sınıflandırılacak sınır öğelerinin geçme frekansları olasılık dağılıma dönüştürülür.
- Olasılık dağılım için normal dağılım parametreleri olan ortalama ve standart sapma her bir sınıf için hesaplanır.
- Hesaplanan değerler sınıflandırılacak öğelerin ardışık olarak geçme dağılımları için 1 ile diğer durumlar için 0 ile çarpılarak tüm ardışık geçen öğelerin o sınıf için olasılığı hesaplanır. Olasılık hangi sınıf için yüksek ise o zaman o sınıf seçilir.
- Sınıfların ardışık geçme olasıkları, karakter yada ardışık öğelerin geçme olasılıkları ve ardışık öğelerin belirli bir sınıfa ait olma olasılıkları olmak üzere toplamda 3 farklı olasılık dağılımı mevcuttur.

Saklı Markov Modelleri

- Saklı Markov modelleri (Hidden Markov Model) ardışık sınıflandırma probleminde geçen 3 olasılık dağılımını birleşik (joint) dağılıma çevirerek modellemektedir.
- Birleşik dağılımda gözlemlenebilir olan ardışık kelime dağılımını (conditional output), gözlemlenemeyen ardışık sınıf dağılımını (conditional hidden) ve ilksel/ön (prior) olasılık dağılımını birleştirilir.
- Gözlemlenebilir dağılım daha önce gördüğümüz dil modelidir. Ardışık olarak geçen kelime, ek gibi durumların şartlı olasılık modelidir.
- Gözlemleneyen dağılım arka plandaki durumlar ve durumlar arası geçişlerin koşullu olasılık modelidir. Bunlar sınıf veya etiket dağılımlarıdır.

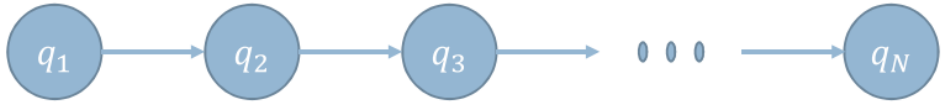
Saklı Markov Modelleri



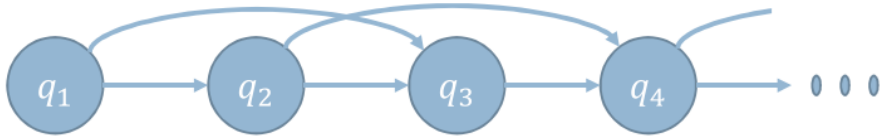
- Yandaki şekilde bir saklı Markov modeli ifade edilmektedir.
- Markov modellerinin tümünde bir durum ve bu durumun gerçekleşme olasılığı sadece ilişkili olduğu durumun koşullu olasılığı olarak belirlenir.
- Yandaki şekilde koşullu olasılıklar ok bağlantıları ile ifade edilmektedir.
- Yandaki şekilde bir kişinin mutlu yada üzgün olma durumu ifade edilmektedir. Buna göre havanın güneşli olması ve kişinin mutlu olması P(mutlu|güneşli)

Saklı Markov Modelleri

- P(mutlu|güneşli) = P(mutlu) * (güneşli|mutlu) / p(güneşli)
- Aşağıdaki ilk model 1. derece Markov modelidir ve ikinci model ise 2. derece Markov modelidir



$$P(q_i|q_1 \dots q_{i-1}) = P(q_i|q_{i-1})$$



$$P(q_i|q_1 \dots q_{i-1}) = P(q_i|q_{i-1}q_{i-2})$$

Saklı Markov Modeli

- Saklı Markov modelinde olasılığın hesaplanması (likelihood), çözümleme (decoding/forward pass) ve öğrenme olmak üzere 2 adım vardır.
- Likelihood: gözlemlenen durum kullanılarak oluşabilecek tüm olası ardışık durum dizilerinin her biri için olasılığın hesaplanması.

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t)$$

Saklı Markov Modeli

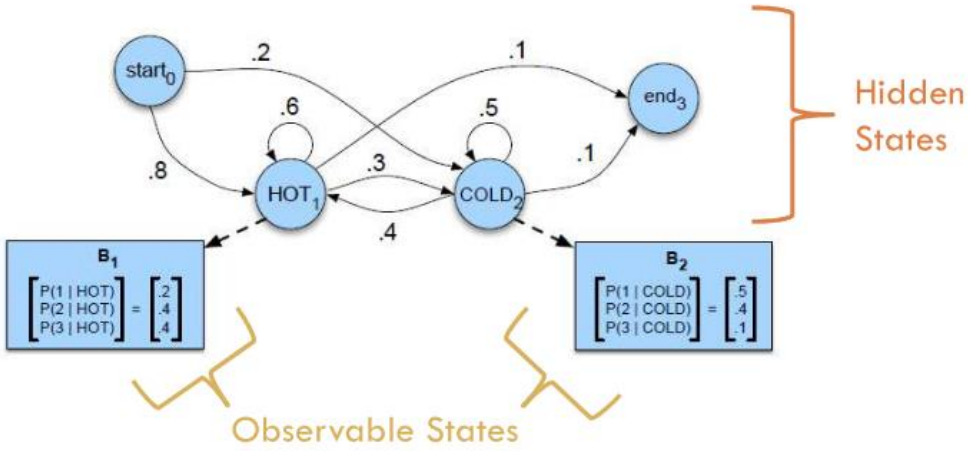
- Çözümleme (decoding/forward) : Sadece maksimum oluşacak olasılığın bulunması.

$$v_t(j) = max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t)$$

- Öğrenme (learning): Durumlar arasında geçişlerin olasılıklarının hesaplanması. Veri üzerinde sayma işlemi ile üretici model ile öğrenilir.

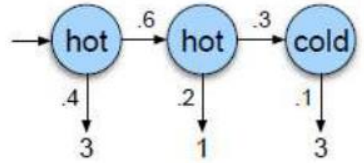
$$a_{ij} = \frac{Count(i \rightarrow j)}{\sum_{q\epsilon} Count(i \rightarrow q)}$$

Saklı Markov Modeli



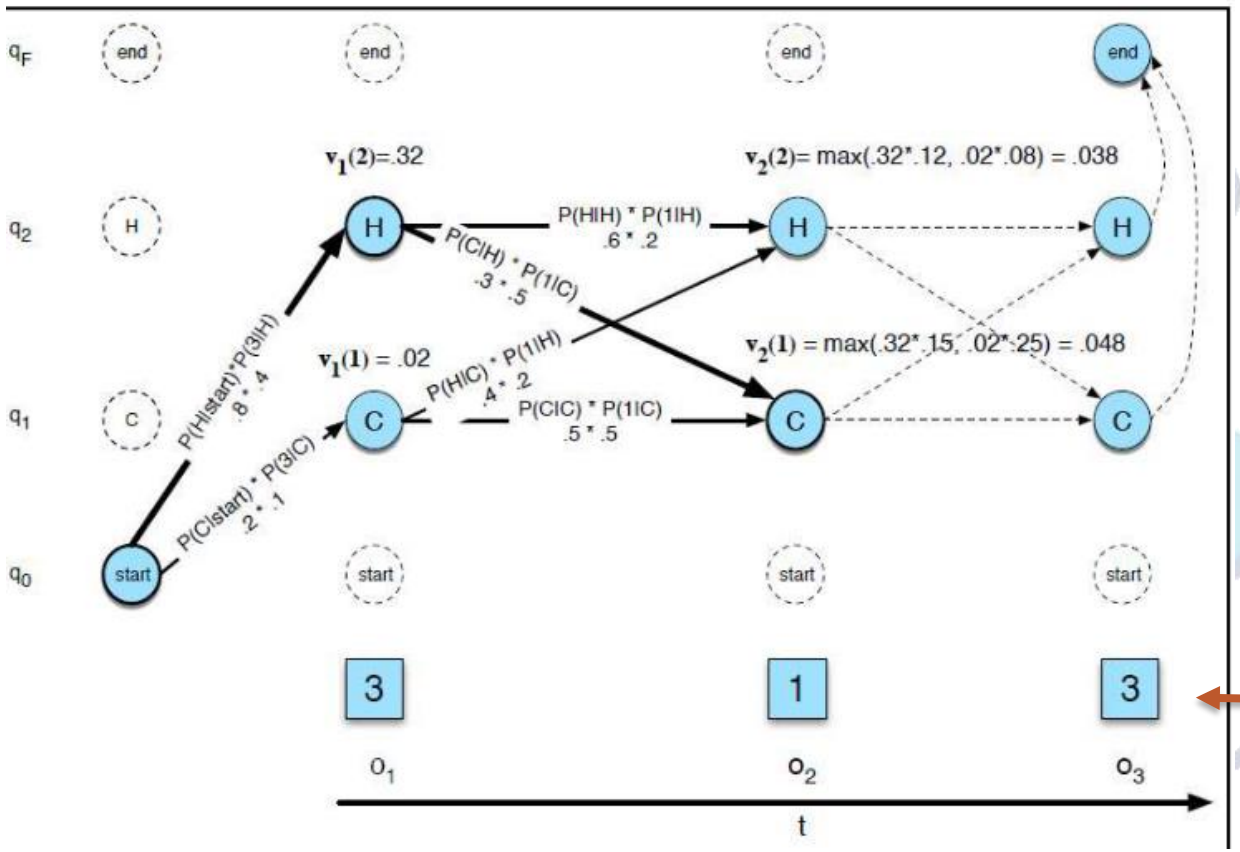
- Yandaki şekilde sıcak ve soğuk gizli/saklı durumları için sayılar gösterilmiştir. Bu sayılar her bir durum için farklı olasılık ile gözlemlenmektedir.
- Bu durumda 3 1 3 için gözlemlenebilecek her bir durum dizisi nedir ve olasılıkları nedir?

HOT, HOT, HOT
HOT, HOT, COLD
HOT, COLD, HOT
HOT, COLD, COLD
COLD, HOT, HOT
COLD, HOT, COLD
COLD, COLD, HOT
COLD, COLD, COLD



$$P(O,Q) = P(O|Q)P(Q)$$
$$= \prod_{i=1}^{T} P(o_i|q_i) \prod_{i=1}^{T} P(q_i|q_{i-1})$$

Saklı Markov Modeli



Saklı durumlar öğeler ve soldan sağa kombinasyonları

Gözlemlenen öğeler

Saklı Markov Modelleri

- Saklı markov modelleri üretici model sınıfındadır.
- Doğal dil işlemede –
- Kelime türü belirleme (Part of Speech Taging) de kullanılabilir.

Alp	dün	akşam	yemeğinde	soslu	makarna	yedi	.
NN/	ADV/	ADV/	NN/	ADJ/	NN/	VBD/	PUNC/

- Kelime isim öbeklerinin bulumasında kullanılabilir.