ELM472 – Makine Öğrenmesinin Temelleri

Ödev 2

Naive Bayes ile Spam Mail Detection

Son teslim tarihi: 31.10.2022 – 17:00

Alican Bayındır
a.bayindir2020@gtu.edu.tr
Elektronik Mühendisliği Bölümü, GTÜ, Kocaeli, Türkiye

I. GİRİŞ

Naive Bayes, sınıflandırıcılar oluşturmak için basit bir yöntemdir. Bu modeller, özellik değerlerinin vektörleri olarak temsil edilen problem durumlarına sınıf etiketleri sağlar ve sınıf etiketleri sınırlı bir kümeden seçilir. Bu tür sınıflandırıcıları eğitmek için sadece bir teknik değil, sınıf değişkeni verildiğinde bir özelliğin değerinin diğer tüm özelliklerin değerinden bağımsız olduğu öncülüne dayanan bir algoritma ailesi vardır. Örneğin bir meyve kırmızı, küresel ve kabaca 10 cm çapında ise elma olarak kabul edilebilir. Renk, yuvarlaklık ve çap değişkenleri arasındaki olası bağlantılara rağmen, saf bir Bayes sınıflandırıcısı, bu özelliklerin her birinin, bu meyvenin bir elma olma olasılığına bağımsız olarak katkıda bulunduğunu düşünür.

Naive Bayes modelleri için parametre tahmini sıklıkla maksimum olabilirlik yöntemini kullandığından, Bayes olasılığını benimsemeden veya herhangi bir Bayes tekniğini uygulamadan saf Bayes modeli ile çalışmak mümkündür.

II. UYGULAMA

Bayes teoremi, e-posta filtreleri oluşturmak için uygulanır. Bayes teoremi, spam mail tanımlama konusu ile ilgili olarak kullanılan basit seviye ayrıştırıcıdır:

1- İletide belirli bir ifadenin bulunduğunu bilerek, iletinin ilk kez spam olma olasılığını belirlemek;
2- 2- tüm terimlerini (veya ilgili bir alt kümesini) göz önünde bulundurarak mesajın ikinci kez spam olma olasılığını belirleyin;
3- Alışılmadık kelimelerle, belki üçüncü kez uğraşmak.
4-

Şüpheli mesajda "replika" teriminin göründüğünü varsayalım. E-posta almaya alışmış çoğu kişi, bu iletişimin muhtemelen istenmeyen posta veya daha spesifik olarak popüler saat markalarının kopyalarını satma önerisi olduğunun farkındadır. Ancak, spam algılama programı yalnızca olasılıkları hesaplayabilir; bu gerçekleri -hangi kelime nerde ve ne anlamda kullanıldığını- bilmediğinden dolayı da "Naive" olarak adlandırılır.

Yazılımın bu olasılık hesabını hesaplamak için kullandığı algoritma Bayes teoremine dayanmaktadır.

$$P(S|W) = \frac{P(W|S) \cdot P(S)}{P(W|S).P(S) + P(W|H).P(H)}$$

Denk 1.

P(S|W), "replika" ifadesini içeren bir iletinin istenmeyen posta olma olasılığıdır;
P(S), belirli bir iletinin istenmeyen posta olma genel olasılığıdır;
İstenmeyen postaların "replika" ifadesini içerme olasılığı P(W|S) olarak bilinir;
Belirli bir iletinin spam (veya "ham") olmama olasılığı Pr(H) olarak ifade edilir;
"Replika" teriminin amatör iletişimlerde ortaya çıkma olasılığı Pr(W|H) ile verilir.

A. Python kodunun yazılması

Gerekli araştırma ve konu incelenmesi yapıldıktan sonra Python kodu yazılmış ve ekte iletilmiştir. Formülün koda işlenmesi sırasında ilk olarak izin verilen 2 kütüphane ödevde bize teslim edilen veri dosyası ile içeri aktarılmıştır. (Numpy ve pandas kütüphaneleri).

	labels	text	label_nums
0	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	spam	Subject: photoshop , windows , office . cheap ...	1
4	ham	Subject: re : indian springs\r\nthis deal is t...	0

Şekil 1Veri seti

Daha sonra içe aktarılan veri dosyası içeriğinde gereksiz bilgileri barındırdığından dolayı örneğin "Unnamed: 0 satırı" temizlenmiştir. Sonrasında elimizdeki mail verilerini test ve train olmak üzere 2 ye ayırıyoruz. Burda training için 0.7 test için 0.3 katsayıları ile veriler çarpıldı ve %70'i train %30'u test olmak üzere ayrıldı. Sonrasında veri içerisinde hala daha gereksiz bulunan karakterler ('.' , ',' , '()' , 'a', 'to') olduğundan dolayı veri tekrar temizlendi ve maillerin hepsi küçük harf olarak düzenlendi. Her mailde bulunan kelimelerin listesi (csr matrix) çıkartıldı. 3 adet kelime listesi bütün kelimelerin listesi, spam maillerde geçen kelimeler ve ham maillerde geçen kelimelerin listesi oluşturuldu. Yukarıda belirtilen Denklem 1 formülü işlendi ve bir olasılık hesabı yapıldı. Bu olasılık hesabının kontrolü için bir fonksiyon yazıldı ve fonksiyon içerisinde gelen mesajın spam mi ham mi ayrımı yapılması için bir kontrol mekanizması oluşturuldu. Formüle göre çıkan olasılık hesabından gelen mailin hangi mail olduğuna karar verildi. Bu işlemler neticesinde modelin doğruluk oranını sadece test etmek için sklearn kütüphanesi kullanıldı ve sınıflandırmanın başarısı raporlandı.

Aşağıdaki resimde fonksiyonun içerisine gönderilen bir mailin başarılı bir şekilde ayrıştırıldığı gösterilmiştir. Fonksiyonun içerisine VSCO uygulamasının gönderdiği bir mail koyulmuştur ve aşağıdaki sonuç elde edilmiştir.

" *Our way of saying thanks*
Thanks for being a part of our community! Building the tools to support and inspire active creators like you is what keeps us going at VSCO.

We love seeing how you use VSCO to express yourself and we want to make sure you have all the tools you need to feel supported on your creative journey."

```
In [69]: classify_test_set('Our way of saying

         It is a ham mail.

Out[69]: 0
```

Bu örnekten sonra aynı fonksiyonun içerisine aşağıdaki cümle yazılmıştır ve spam olduğu başarıyla saptanmıştır.

"*Spank monster daddy is on discount only and only just for you!"*

```
In [70]: classify_test_set("Spank monster daddy

         It is a spam mail.

Out[70]: 1
```

III. SONUÇ

Ödevde bize teslim edilen veriler başarılı bir şekilde analiz edilmiş olup, mail hakkında spam-ham ayrımı yapılmıştır. Bu yöntem kelimelerin önceki örneklerinde geçip geçmemelerine ve geçtiyse hangi sıklıkla geçtiğine bakarak bir olasılık hesabı yapar ve mailin spam mi ham mi olduğuna karar verir. Burada kelimenin bağlamı, anlamı, içeriği ve ne için kullanıldığı gibi bilgilere bakılmaksızın karar verildiğinden dolayı bu yöntem "Naive" olarak adlandırılır. Çünkü her kelimeye başlangıçta eşit şekilde davranır. "The", "a", "some" ve "is" (İngilizce) gibi kelimeler veya bunların yabancı dildeki karşılıkları karar kısmında bir etki sağlamayacağından dolayı göz ardı edilebilir. Dikkate alınan terimler, spam içeriği 0.0'a (geçerli iletişimin ayırt edici göstergeleri) veya 1.0'a yakın olan terimlerdir. (belirli spam işaretleri). Bir strateji, incelenmekte olan mesajdan sadece 10 kelimeyi |0.5 pI| mutlak değeri ile tutmak olabilir. Daha genel olarak, bazı Bayesian filtreleme algoritmaları, karar verme sürecine fazla bir şey katmadıkları için spam değeri 0,5'in altında olan herhangi bir kelimeyi basitçe atar.

IV. KAYNAKÇA

[1] Ahmet Güneş, Dr. Öğr. Üyesi, ELM 472 - Makine Öğrenmesi Temelleri dersi, Gebze Teknik Üniversitesi.

[2] E. Alpaydin, Introduction to Machine Learning, 3. bs. Cambridge, MA, USA: MIT Press, 2014.

[3] Anonym. (2022, May 16). Naive Bayes Spam Filtering. Wikipedia. https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering

[4] Andrew NG, Dr. , CS 229 - Makine öğrenmesi, Harvard Üniversitesi