Makine Öğrenmesi Güz 2017 1. Ödevi "Naïve Bayes Yöntemi ile Doküman Sınıflandırma"

14505013 İhsan Ozan YILDIRIM

Yıldız Teknik Üniversitesi

Not

Bu çalışmanın kaynak kodlarına aşağıdaki adresten erişilebilir.

https://github.com/ozantoteles/naiveBayesHW2

Özet

Bu ödevde Naïve Bayes yönteminin doküman sınıflandırmadaki başarısı değerlendirilmiştir. Veri seti olarak https://www.kaggle.com/benhamner/clinton-trump-tweets/data adresindeki Twitter'dan toplanılmış "HillaryClinton" ve "realDonaldTrump" kullanıcılarının "tweet"leri kullanılmıştır. Ödev sonucunda test setindeki "tweet"lerin hangi kullanıcıya ait olduğu %92 doğrulukla belirlenmiştir.

Anahtar Kelimeler: Makine Öğrenmesi, Naïve Bayes, Doküman Sınıflandırma, Twitter, Donald Trump, Hillary Clinton

14505013 İhsan Ozan YILDIRIM

Naïve Bayes Yöntemi ile Doküman Sınıflandırma

Giriş

Naïve Bayes basit ama güçlü bir Makine Öğrenmesi yöntemidir. Son yıllarda Makine Öğrenmesi alanındaki önemli gelişmelere bakıldığında basitliğinin yanında hızlı, doğru ve güvenilir bir yöntem olmasıyla öne çıkmaktadır.

Bir Makine Öğrenmesi modeli kurmak için öncelikle kullanılacak verinin modeli ilgilendirecek özelliklerini belirlemek gerekmektedir. Bu ödevde doküman sınıflandırma yapılacağı için özellik olarak kelimelerin sıklıkları kullanılmıştır. Bunun dışında, Naïve Bayes yöntemi değişkenlerin birbirinden bağımsız olduğu varsayımına dayandığı için, kelimelerin sırası ve yapısı özellik olarak değerlendirilmemiş, bu özellikler yok sayılmıştır.

Giriş bölümünün kalanında yöntemi açıklamak için aşağıdaki 5 cümle ve 1 test cümlesi örnek olarak kullanılacaktır. Dilbilgisi ve sözcük yapısı göz önünde bulundurularak ödev için İngilizce dilindeki metinler tercih edilmiştir.

Metin	**Sınıf**
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports
"A very close game"	?

Örnek veri seti (Stecanella, 2017).

İstatistik ve Bayes Teoremi

Bayes Teoremi koşullu olasılıklarla çalışmak için oldukça kullanışlıdır.

$$P(A|B) = \frac{P(B|A)xP(A)}{P(B)}$$

Bir değişkenin koşullu olasılığını tersine çevirmek bilinmeyen olasılıkları hesaplamak için kullanılır.

Doküman sınıflandırma için kelimelerin geçme sıklıkları olasılıklarını hesaplamak için kullanılabilir. Bizim örnek veri setimizde Bayes Teoremini uygulamak istersek $P(sports|a\ very\ close\ game)$ olasılığı tersine çevrilerek aşağıdaki gibi hesaplanabilir.

$$P(sports|a\ very\ close\ game) = \frac{P(a\ very\ close\ game|sports)xP(sports)}{P(a\ very\ close\ game)}$$

Sınıflandırma yapmak için $P(sports|a\ very\ close\ game)$ ve $P(Not\ sports|a\ very\ close\ game)$ ifadeleri karşılaştırılacaktır. Bu nedenle hesaplamalar yapılırken Bayes Teoremi'nin payda kısmındaki ifadeleri kullanılmayabilir. Bu durumda bir cümlenin hangi sınıfa ait olduğunu belirlemek için aşağıdaki iki ifade kullanılır.

$$P(a\ very\ close\ game|sports)xP(sports)$$

Ve

$$P(a\ very\ close\ game|Not\ sports)xP(Not\ sports)$$

Bu şekilde "sports" sınıfında "A very close game" ifadesinin kaç defa geçtiği hesaplayarak bu olasılıkları hesaplayabiliriz. Bunun için ise yöntemin adındaki Naïve (naif) bakış açısından faydalanırız. Dokümanlarımızdaki bütün kelimelerin olasılıklarının birbirinden bağımsız olduğu varsayımında bulunursak $P(a\ very\ close\ game)$ ifadesi aşağıdaki gibi yazılabilir.

$$P(a\ very\ close\ game) = P(a)xP(very)xP(close)xP(game)$$

Aynı varsayımı daha önceki karşılaştırma ifadelerine uygularsak:

$$P(a\ very\ close\ game|sports)$$
$$= P(a|sports)xP(very|sports)xP(close|sports)xP(game|sports)$$

Artık bu kelimelerin veri setimizde kaç defa geçtiğini sayarak olasılıklarını hesaplayabiliriz.

$P(sports)$ ve $P(Not\ sports)$ ifadelerinin değerleri veri setinden kolayca görülebileceği gibi $3/5$ ve $2/5$'tir.

$P(game|sports)$ ifadesi ise yine veri setinden $2/11$ olarak hesaplanabilir. Ancak $P(close|sports)$ hesaplanmak istediğinden değeri 0 olacağından koşullu olasılıklar hesaplanırken Laplace Smooting yöntemi (ya da benzer başka bir yöntem) uygulanmalıdır. Bunun için pay ifadesine 1 eklenir ve paydaya ise toplam kelime sayısı eklenir. Bu durumda $P(game|sports) = \frac{2+1}{11+14}$ ve diğer kelimelerin olasılıkları aşağıdaki gibi olur.

Kelime	$P(kelime\|sports)$	$P(kelime\|Not\ Sports)$
a	$\frac{2+1}{11+14}$	$\frac{1+1}{9+14}$
very	$\frac{1+1}{11+14}$	$\frac{0+1}{9+14}$
close	$\frac{0+1}{11+14}$	$\frac{1+1}{9+14}$
game	$\frac{2+1}{11+14}$	$\frac{0+1}{9+14}$

Son olarak bu olasılıkları birbiriyle çarparak hangisinin daha büyük olduğunu hesaplayabiliriz.

$$P(a|sports)xP(very|sports)xP(close|sports)xP(game|sports)x$$

$P(sports) = 5{,}97197E\text{-}06$

$$P(a|Not\ sports)xP(very|Not\ sports)xP(close|Not\ sports)x$$

$$P(game|Not\ sports)xP(Not\ sports) = 2{,}6214E\text{-}07$$

Sonuç olarak $5{,}97197E-06 > 2{,}6214E-07$ olduğu için *"a very close game"* cümlesi "sports" sınıfına dahildir diyebiliriz.

Uygulama

Bu ödevde Naïve Bayes yöntemi Python (3.6.2) dili, JetBrains PyCharm Professional (2017.2) geliştirme ortamı, SQLite3 veritabanı ve *json*, *csv*, *os*, *sqlite3*, *time*, *numpy*, *tweepy* Python kütüphaneleri kullanılarak geliştirilmiştir.

Sınıflandırma yapılacak ve seti https://www.kaggle.com/benhamner/clinton-trump-tweets/data adresinden edinilmiştir. Geliştirme aşamasında önce Twitter API *tweepy* kütüphanesi üzerinden kullanılarak veri seti elde edilmiş ancak sonrasında etiketleme işlemindeki zorluklar nedeniyle hazır kütüphane kullanımına yönelinmiştir.

Kaggle.com internet sitesinden edinilen veri seti 6953 "tweet"ten oluşmaktadır. "tweet"ler yazan kullanıcının kim olduğuna göre iki sınıftan (HillaryClinton, realDonaldTrump) etikenlemiştir. Veri setindeki dokümanlar iki sınıf arasında yaklaşık olarak yarı yarıya dağılmıştır (3226 HillaryClinton, 3218 realDonaldTrump). Benzer şekilde 10543 kelimeden oluşan sözlük de yaklaşık olarak yarı yarıya dağılmış kelimelerden (31011 HillaryClinton, 32184 realDonaldTrump) oluşmaktadır.

İlk aşamada Twitter'dan indirilen daha sonra ise Kaggle'dan edinilen "tweet"ler her biri *json* formatından ayrı birer doküman olarak diske kaydedilmiştir. Daha sonra dokümanlar ön işlemden geçirilmiş ve veritabanına yazılmıştır.

Ön İşlem

"tweet" dokümanları noktalama işaretleri ve "stop word"lerden arındırıldıktan sonra sınıflandırma işlemine tabi tutulmuştur. Dokümanlardan çıkarılan noktalama işaretleri ve "stop word"ler aşağıdaki gibidir.

Noktalama işaretleri: '!'#$%&\'()*+,-./:;<=>?@[\\]^_` {|}~'…—'

"Stop Word"ler:
"rt","a","about","above","after","again","against","all","am","an","and","any","are","arent","as","at","be","because","been","before","being","below","between","both","but","by","cant","cannot","could","couldnt","did","didnt","do","does","doesnt","doing","dont","down","during","each","few","for","from","further","had","hadnt","has","hasnt","have","havent","having","he","hed","hell","hes","her","here","heres","hers","herself","him","himself","his","how","hows","i","id","ill","im","ive","if","in","into","is","isnt","it","its","its","itself","lets","me","more","most","mustnt","my","myself","no","nor","not","of","off","on","once","only","or","other","ought","our","ours","ourselves","out","over","own","same","shant","she","shed","shell","shes","should","shouldnt","so","some","such","than","that","thats","the","their","theirs","them","themselves","then","there","theres","these","they","theyd","theyll","theyre","theyve","this","those","through","to","too","under","until","up","very","was","wasnt","we","wed","well","were","weve","were","werent","what","whats","when","whens","where","wheres","which","while","who","whos","whom","why","whys","with","wont","would","wouldnt","you","youd","youll","youre","youve","your","yours","yourself","yourselves"

Kullanılan veri setindeki "tweet"ler kullanıcıların kendi görüşlerini yansıttığı için "Re-tweet" edilen "tweet"lerin elenmesine gerek duyulmamıştır. Sadece "rt" ifadesi "stop word" olarak değerlendirilmiştir.

Başka kullanıcıların "mention" yapılması durumunda "tweet"lerde beliren "@" işareti noktalama işareti olarak değerlendirilmiş, "mention" yapılan isimlerin modele etkisi göz ardı edilmemiştir.

Benzer şekilde "hashtag"ler için kullanılan "#" işareti de noktalama işareti olarak değerlendirilmiştir.

Sınıflandırma

Sınıflandırma yapılırken kelimelerin kaç defa geçtiği bilgisi de ayrı veritabanı tablolarında tutulmuştur. Bütünlük sağlamak amacıyla bu raporda kaynak kodun son halinde çalıştırılmış iki adet örnek ele alınacaktır.

İlk Örnek

Veri seti rasgele bir şekilde %50 öğrenme ve %50 test etme verisi olacak şekilde ayrılmıştır. Veri setindeki dağılımlar aşağıdaki tablodaki gibidir.

	Öğrenme		Test	
tweet	3362		3083	
HillaryClinton "tweet"leri	2123	63%	1103	36%
realDonaldTrump "tweet"leri	1238	37%	1980	64%

Bu örnekte kodun çalıştırılması öğrenme 210 ve test 12 saniye olmak üzere yaklaşık 222 saniye sürmüştür. [1] Bu örnek için karışıklık (confusion) matrisi ve doğruluk (accuracy) değeri aşağıdaki gibidir.

		Veri Setinin etiketleri	
		HillaryClinton	realDonaldTrump
Modelin Sağladığı Etiketler	HillaryClinton	1092	895
	realDonalTrump	11	1085

$$Doğruluk = \frac{1092 + 895}{1092 + 895 + 11 + 1085} = 0{,}706 = \%70$$

[1] Windows 7 Enterprise 64-bit İşletim Sistemi, Intel® Core™ i5-6200U 2.30GHz İşlemci, 8 GB RAM

İkinci Örnek

Veri seti rasgele bir şekilde yaklaşık %75 öğrenme ve %25 test etme verisi olacak şekilde ayrılmıştır. Veri setindeki dağılımlar aşağıdaki tablodaki gibidir.

	Öğrenme		Test	
tweet	5430		1523	
HillaryClinton "tweet"leri	2611	48%	615	40%
realDonaldTrump "tweet"leri	2818	52%	908	60%

Bu örnekte kodun çalıştırılması öğrenme 300 ve test 6 saniye olmak üzere yaklaşık 306 saniye sürmüştür. Bu örnek için karışıklık (confusion) matrisi ve doğruluk (accuracy) değeri aşağıdaki gibidir.

		Veri Setinin etiketleri	
		HillaryClinton	realDonaldTrump
Modelin Sağladığı Etiketler	HillaryClinton	524	32
	realDonalTrump	91	876

$$Doğruluk = \frac{876 + 524}{876 + 524 + 91 + 32} = 0{,}919 = \%92$$

Sonuçların Yorumlanması

Bütün Makine Öğrenmesi yöntemlerinde olduğu gibi Naïve Bayes yöntemi için de veri setinin doğru seçilmesi ve doğru hazırlanması oldukça önemlidir. Raporda sunulan iki örnek öğrenme ve test seti dağılımı ve sonuçlar dışında geliştirme aşamasında birçok farklı dağılım ve bazı başka veri setleri denenmiştir.

Naïve Bayes yöntemi için öğrenme setinde sınıfların doküman ve kelime sayılarının eşit dağılması olasılık hesabını doğrudan etkilediği için önemlidir. Dağılımın eşit olmadığı durumlarda yöntemin sayısı daha fazla olan sınıfa yönlenebileceği gözlemlenmiştir.

Rapordaki iki örnek arasında da görülebileceği gibi öğrenme setinin büyük olması doğruluk üzerinde olumlu etki yaratmaktadır.

Örnek başarısız "tweet"ler (İkinci Örnek'ten)

Gerçek "tweet" 1:
stonewall: the birthplace of a movement and soon a national monument for equality. https://t.co/yb1u1tucr4 #thanksobama
Ön işlem yapılmış "tweet":
thanksobama equality stonewall monument birthplace movement soon national
Veri Seti: HillaryClinton
Model: realDonaldTrump

Gerçek "tweet" 2:
let's act on coal miner safety so people like don blankenship are held accountable for blatantly disregarding it. https://t.co/eyxdxpvtuh
Ön işlem yapılmış "tweet":
disregarding coal miner don held blankenship accountable safety blatantly people act like
Veri Seti: HillaryClinton
Model: realDonald Trump

Gerçek "tweet" 3:
just a few hours left to vote in the ny primary. confirm your polling place then head over: https://t.co/iwio5b9eal https://t.co/rch0h3ecrn
Ön işlem yapılmış "tweet":
vote primary left place just hours polling head confirm ny
Veri Seti: HillaryClinton
Model: realDonald Trump

Gerçek "tweet" 4:
everybody is talking about the protesters burning the american flags and proudly waving mexican flags. i want america first - so do voters!
Ön işlem yapılmış "tweet":
first want america talking everybody proudly voters mexican protesters burning american flags waving
Veri Seti: realDonaldTrump
Model: HillaryClinton

Gerçek "tweet" 5:
the @uschamber must fight harder for the american worker. China and many others
Ön işlem yapılmış "tweet":
china many harder pacts fight advantage trade terrible worker american must uschamber taking us others
Veri Seti: realDonaldTrump
Model: HillaryClinton

İlk üç örnek "tweet"teki "movement, national ve people" sözcükleri daha çok realDonaldTrump kullanıcısının kullandığı sözcükler olduğu ve "tweet"lerdeki diğer sözcükler olasılıkları düşük sözcükler olduğu için model HillaryClinton yerine realDonaldTrump sonucu vermiştir.

Diğer iki başarısız örnekte ise "america, mexican" gibi her iki kullanıcının da çok kullandığı sözcüklerin olması modelin doğru sonuç vermesine engel olmuştur.

Örnek başarılı "tweet"ler (İkinci Örnek'ten)

Gerçek "tweet":
we've got a candidate in hillary clinton who is a fighter. we got one tough cookie. — @senwarren
https://t.co/vu6pbtlsy4
Ön işlem yapılmış "tweet":
clinton tough hillary senwarren fighter cookie one got candidate

Veri Seti: HillaryClinton
Model: HillaryClinton

Gerçek "tweet":
today's third stop- londonderry new hampshire! thank you! #fitn #votetrumpnh https://t.co/prpcxaz7ov
Ön işlem yapılmış "tweet":
third todays hampshire stop fitn londonderry votetrumpnh thank new

Veri Seti: realDonaldTrump
Model: realDonaldTrump

Gerçek "tweet":
we can't stop fighting until all lgbt americans can live their lives free of prejudice violence
Ön işlem yapılmış "tweet":
prejudice violence free americans fighting live hate lives stop lgbt can

Veri Seti: HillaryClinton
Model: HillaryClinton

Gerçek "tweet":
donald trump is closer than ever to clinching the gop nomination. there's only one candidate who has more votes. https://t.co/luvanujuks
Ön işlem yapılmış "tweet":
ever nomination donald candidate one votes clinching trump gop closer

Veri Seti: HillaryClinton
Model: HillaryClinton

Gerçek "tweet":
@gregusp61: you really rocked them hard in s.c. rubio and cruz were pummled. so glad jeb is gone! next no liar!
Ön işlem yapılmış "tweet":
gregusp61 rocked really hard sc pummled liar next gone glad cruz rubio jeb

Veri Seti: realDonaldTrump
Model: realDonaldTrump

Sonuç

Naïve Bayes yöntemi basit ve kolay uygulanabilir olmasının yanında hızlı ve güvenilir bir yöntemdir. Güvenilirliği garanti altına almak için veri setini doğru seçmek ve hazırlamak oldukça önemlidir. Bu ödevde veri seti olarak "tweet"ler kullanıldığı için her bir dokümanın uzunluğu oldukça kısa kalmasına rağmen model %92 gibi bir doğruluk oranına ulaşabilmiştir. Daha uzun dokümanlarda benzer doğruluk seviyelerine ulaşmak daha kolay olacaktır.
Makine Öğrenmesi yöntemlerinden Naïve Bayes'in önemli ve etkili bir doküman sınıflandırma yöntemi olduğu görülmüştür.