YILDIZ TEKNİK ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ

Makine Öğrenmesi 2. Ödevi

Ana Bileşen Analizi Yöntemi İle El Yazısı Karakter Tanıma

Melike Nur Mermer - 15501010
27.11.2016

Giriş:

Bu çalışma 0-9 arası sayıların grayscale görüntülerinin piksel değerlerini tutan MNIST veriseti kullanılarak Ana Bileşen Analizi yöntemi ile en belirleyici özellikleri içeren eğitim datasının belirlenmesi ve test verilerinin eğitim örneklerine olan öklit mesafesi ile sınıflandırılmasını konu almaktadır. Sınıflandırma sonucunda Ana Bileşen Analizi ile elde edilen en belirleyici özelliklerden %40, %60, %80'i kullanılarak her rakamın doğru tahmin edilme sayıları hesaplanmış ve yanlış tahmin edildiği durumlarda en çok hangi rakam ile karıştırıldığını gösteren karışıklık matrisi gösterilmiştir.

Yöntem:

Verisetinden eğitim ve test örneklerinin çıkarılması

Ana bileşen analizi uygulanmadan önce kullanılacak eğitim seti belirlenmiştir. Verilen datasetteki her sınıf için ilk 5 örnek eğitim setine alınmıştır. Yani toplamda 50 örnekten oluşan bir eğitim setimiz olmuştur. Eğitim örnekleri belirlendikten sonra bu örneklerin dışında kalan örnekler test örnekleri olarak belirlenmektedir. Bu işlem için datasetteki tüm elemanların eğitim örneği içinde olup olmadığı teker teker kontrol edilir. Üzerinde çalışılan eleman eğer eğitim setinde varsa "var" değişkeni 1 yapılır. "var" değişkeninin döngü sonunda 0 olarak kalması o örneğin eğitim setinde olmadığını gösterir ve o örnek test setine eklenir.

Ana bileşen analizi

En belirleyici özelliklerin çıkarılması işlemi için öncelikle elde edilen eğitim datasının transpozu alınır ve tüm sütunların ortalaması her bir sütundan çıkartılarak "mean centered sample"lar elde edilir. Elde edilen matris ile transpozunun çarpılması sonucu kovaryans matrisi hesaplanır ve bu matristen özdeğerler ve bu özdeğerlere karşılık gelen özvektörler belirlenir. Bu özvektörler matrisinin transpozu ile "mean centered sample"ların çarpılması sonucu özellikler başka bir uzaya taşınır ve "eigenfaces" elde edilir. Belirlenen özdeğerlerden ilk aşamada %40 'ı ve bunların karşılık geldiği özvektörler kullanılarak "emat4" %40lık eğitim matrisi oluşturulur. Aynı şekilde "emat6" %60lık eğitim matrisi ve "emat8" %80lik eğitim matrisini ifade etmektedir.

Test örneklerinin sınıflandırılması

Test için belirlenen örneklerin öncelikle eğitim örneklerinin bulunduğu uzaya taşınması gereklidir. Bu işlem için öncelikle test örnekleri "mean centered sample" olarak kaydedilir ve özvektörlerin transpoz matrisi ile çarpılarak eğitim örnekleri ile aynı uzaya taşınır. Sırasıyla %40lık, %60lık ve %80lik eğitim matrisleri için test örneklerinin her bir eğitim örneğine olan öklit mesafesi bulunur ve en yakın olduğu eğitim örneğinin sınıfında olarak karar verilir. Farklı boyuttaki eğitim matrisleri için her bir sınıfın doğru tahmin edilme sayıları uygulama bölümünde verilecektir.

Uygulama:

a. %40, %60 ve %80lik eğitim matrisleri için elde edilen doğru tahmin sayıları

	%40	**%60**	**%80**
0	367	366	365
1	252	265	268
2	345	345	343
3	357	357	356
4	162	141	141
5	271	284	279
6	365	367	367

7	374	375	373
8	204	231	216
9	73	66	65
toplam	2770	2797	2773

Bu sonuçlara göre toplamda en çok doğru tahmin %60lık eğitim matrisi ile elde edilmiştir. Toplamda 3773 test örneği olduğu düşünülürse bu testin başarısı 2797/3773=%74 olarak belirlenmektedir. Buradan anlaşılmaktadır ki en belirleyici özelliklerin belli bir orana kadar eğitim datasına alınmasının test başarısına olumlu etkisi olurken bu değerden sonra yeni özellik eklemenin olumlu bir etkisi olmamaktadır.

b. 0-9 arası rakamların tanıma sırasında en çok hangi rakamlarla karıştırıldığını gösteren "confusion matrix" %40, %60 ve %80lik eğitim matrisleri için ayrı ayrı hesaplanmıştır.

%40lık eğitimde elde edilen karışıklık matrisleri

Örnek sayısına göre:

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	**367**	0	0	1	1	0	1	0	1	0
1	0	**252**	34	7	5	11	3	10	57	5
2	0	1	**345**	13	0	0	8	6	2	0
3	0	7	7	**357**	0	5	0	2	1	5
4	27	57	0	9	**162**	39	36	20	1	31
5	21	1	1	51	0	**271**	12	3	3	8
6	1	6	0	0	0	0	**365**	0	0	0
7	0	3	1	0	1	0	0	**374**	0	3
8	3	9	21	92	5	3	11	22	**204**	5
9	3	0	6	**186**	26	54	1	9	19	73

Toplam yüzdeye göre:

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	**98,92%**	0,00%	0,00%	0,27%	0,27%	0,00%	0,27%	0,00%	0,27%	0,00%
1	0,00%	**65,63%**	8,85%	1,82%	1,30%	2,86%	0,78%	2,60%	14,84%	1,30%
2	0,00%	0,27%	**92,00%**	3,47%	0,00%	0,00%	2,13%	1,60%	0,53%	0,00%
3	0,00%	1,82%	1,82%	**92,97%**	0,00%	1,30%	0,00%	0,52%	0,26%	1,30%
4	7,07%	14,92%	0,00%	2,36%	**42,41%**	10,21%	9,42%	5,24%	0,26%	8,12%
5	5,66%	0,27%	0,27%	13,75%	0,00%	**73,05%**	3,23%	0,81%	0,81%	2,16%
6	0,27%	1,61%	0,00%	0,00%	0,00%	0,00%	**98,12%**	0,00%	0,00%	0,00%
7	0,00%	0,79%	0,26%	0,00%	0,26%	0,00%	0,00%	**97,91%**	0,00%	0,79%
8	0,80%	2,40%	5,60%	24,53%	1,33%	0,80%	2,93%	5,87%	**54,40%**	1,33%
9	0,80%	0,00%	1,59%	**49,34%**	6,90%	14,32%	0,27%	2,39%	5,04%	19,36%

%60lık eğitimde elde edilen karışıklık matrisleri

Örnek sayısına göre:

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	**366**	0	0	2	1	0	1	0	1	0
1	0	**265**	28	7	4	12	2	12	52	2
2	0	1	**345**	14	0	0	8	5	2	0
3	0	8	7	**357**	0	5	0	2	0	5
4	34	67	0	11	**141**	35	32	21	1	40
5	23	0	0	42	0	**284**	13	1	1	7
6	1	4	0	0	0	0	**367**	0	0	0
7	0	2	1	0	2	0	0	**375**	0	2
8	3	12	19	66	5	3	11	20	**231**	5
9	6	0	8	**188**	27	51	2	9	20	66

Toplam yüzdeye göre:

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	**98,65%**	0,00%	0,00%	0,54%	0,27%	0,00%	0,27%	0,00%	0,27%	0,00%
1	0,00%	**69,01%**	7,29%	1,82%	1,04%	3,13%	0,52%	3,13%	13,54%	0,52%
2	0,00%	0,27%	**92,00%**	3,73%	0,00%	0,00%	2,13%	1,33%	0,53%	0,00%
3	0,00%	2,08%	1,82%	**92,97%**	0,00%	1,30%	0,00%	0,52%	0,00%	1,30%
4	8,90%	17,54%	0,00%	2,88%	**36,91%**	9,16%	8,38%	5,50%	0,26%	10,47%
5	6,20%	0,00%	0,00%	11,32%	0,00%	**76,55%**	3,50%	0,27%	0,27%	1,89%
6	0,27%	1,08%	0,00%	0,00%	0,00%	0,00%	**98,66%**	0,00%	0,00%	0,00%
7	0,00%	0,52%	0,26%	0,00%	0,52%	0,00%	0,00%	**98,17%**	0,00%	0,52%
8	0,80%	3,20%	5,07%	17,60%	1,33%	0,80%	2,93%	5,33%	**61,60%**	1,33%
9	1,59%	0,00%	2,12%	**49,87%**	7,16%	13,53%	0,53%	2,39%	5,31%	17,51%

%80lık eğitimde elde edilen karışıklık matrisleri

Örnek sayısına göre:

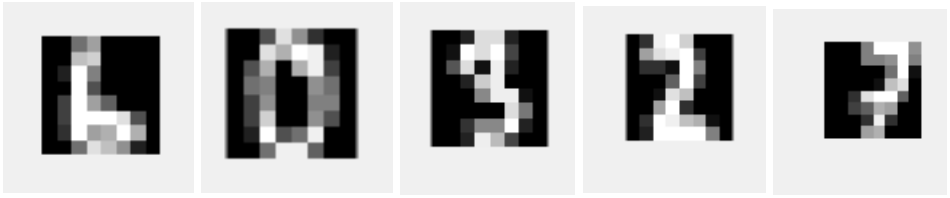
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	**365**	0	0	2	1	0	1	0	2	0
1	0	**268**	26	7	4	12	3	13	49	2
2	0	1	**343**	15	0	0	8	6	2	0
3	0	6	9	**356**	0	6	0	2	0	5
4	29	71	0	10	**141**	37	38	21	1	34
5	24	0	1	47	0	**279**	11	1	1	7
6	1	4	0	0	0	0	**367**	0	0	0
7	0	2	1	0	2	0	0	**373**	0	4
8	4	12	20	79	5	2	9	22	**216**	6
9	4	0	8	**189**	27	51	3	9	21	65

Toplam yüzdeye göre:

	0	1	2	3	4	5	6	7	8	9
0	**98,38%**	0,00%	0,00%	0,54%	0,27%	0,00%	0,27%	0,00%	0,54%	0,00%
1	0,00%	**69,79%**	6,77%	1,82%	1,04%	3,13%	0,78%	3,39%	12,76%	0,52%
2	0,00%	0,27%	**91,47%**	4,00%	0,00%	0,00%	2,13%	1,60%	0,53%	0,00%
3	0,00%	1,56%	2,34%	**92,71%**	0,00%	1,56%	0,00%	0,52%	0,00%	1,30%
4	7,59%	18,59%	0,00%	2,62%	**36,91%**	9,69%	9,95%	5,50%	0,26%	8,90%
5	6,47%	0,00%	0,27%	12,67%	0,00%	**75,20%**	2,96%	0,27%	0,27%	1,89%
6	0,27%	1,08%	0,00%	0,00%	0,00%	0,00%	**98,66%**	0,00%	0,00%	0,00%
7	0,00%	0,52%	0,26%	0,00%	0,52%	0,00%	0,00%	**97,64%**	0,00%	1,05%
8	1,07%	3,20%	5,33%	21,07%	1,33%	0,53%	2,40%	5,87%	**57,60%**	1,60%
9	1,06%	0,00%	2,12%	**50,13%**	7,16%	13,53%	0,80%	2,39%	5,57%	17,24%

Bu matrislerden anlaşıldığı üzere test örneği olarak verilen rakamlardan en çok karıştırılan rakam 9 olmuştur. 9 rakamı her 3 eğitimde de yaklaşık %50 oranında 3 olarak tahmin edilmiştir. Bunun dışında en çok yanlış tahmin edilen rakamlar 8 ve 4 olmuştur.

c. Test verilerine indislerinden ulaşabileceğimiz doğru tahmin edilenlere 1, yanlış tahmin edilenlere 0 atanan bir dizi "crt" oluşturulmuştur. Bu diziden alınan indislere göre bazı doğru tahmin edilen resimler şunlardır:



Gerçek=6 Gerçek=0 Gerçek=3 Gerçek=2 Gerçek=7

Yanlış tahmin edilen bazı resimler ve tahmin edilen sınıflar da aşağıda verilmiştir:

Tahmin=6 Gerçek=4 | Tahmin=3 Gerçek=9 | Tahmin=0 Gerçek=5 | Tahmin=3 Gerçek=8 | Tahmin=9 Gerçek=7

Sonuç:

Hesaplanan karışıklık matrislerine bakıldığında en çok karıştırılan rakamın 9 olduğu görülmektedir. Bunun sebebi 9 rakamının yazılışının 3 rakamına benzemesidir. Diğer bazı rakamların karıştırılmasında ise el yazılarının tanınamaz halde olması etkili olmuştur. Uygulama kısmında verilen resimlerde başarısız örneklerin bir kısmının bakıldığında bile anlaşılamayacak kadar kötü yazıldığı görülmektedir. Eğitim örneklerinde de bu gibi verilerin olması yani eğitimde kullanılan verilerde kötü örneklerin olması test örneklerinin yanlış sınıflandırılmasına neden olabilmektedir.

Kaynak kod:

```
clear all;
dataset=importdata('C:\Users\Melike Nur
Mermer\Desktop\PCA\PCA\sayi.dat');
[sample,feature]=size(dataset);
for i=1:sample
    for j=1:feature-1
    x(i,j)=dataset(i,j);
    end
    y(i,1)=dataset(i,feature);
end
nofclass=max(y)+1;
a=1;
for i=1:nofclass
nofclasssample=0;
        for k=1:sample
        if y(k,1)==(i-1)
            traindatax(a,:)=x(k,:);
            traindatay(a,1)=y(k,1);
            a=a+1;
            nofclasssample=nofclasssample+1;
        end
        if nofclasssample==5
            break;
        end
        end
end

a=1;
for i=1:sample
var=0;
    for j=1:50
        if x(i,:)==traindatax(j,:)
           var=1;
           break;
        end
    end
    if var==0
            testdatax(a,:)=x(i,:);
            testdatay(a,1)=y(i,:);
            a=a+1;
    end
end
%train örneklerini sütunlar haline getir
trainx=transpose(traindatax);
averages(:,1)=mean(trainx.');
%hesaplanan ortalamadan mean centered samplelar elde edilir
for i=1:50
    mcs(:,i)=trainx(:,i)-averages(:,1);%mean centered sample
end
mcst=transpose(mcs);
%covariance matrix hesaplanır
cov=mcs*mcst;
%eigenvalue-eigenvectorler hesaplanır
```

```
eval=eig(cov);%eigenvaluelar ve eigenvectorler sıralı geliyor
[evect,v]=eig(cov);
evect=transpose(evect);
efaces=evect*mcs;
%64 tanenin 11 tanesi "0"
%ilk %40 evale karşılık gelen
for i=1:53*0.4
    emat4(i,:)=efaces(i,:);
end
%ilk %60 evale karşılık gelen
for i=1:53*0.6
    emat6(i,:)=efaces(i,:);
end
%ilk %80 evale karşılık gelen
for i=1:floor(53*0.8)
    emat8(i,:)=efaces(i,:);
end

%eğitim bitti bu örneklere olan benzerliğine bakılarak test yapılır
testx=transpose(testdatax);
for i=1:length(testx)
mctestx(:,i)=testx(:,i)-averages(:,1);
end
efacetestx=evect*mctestx;

%test datalarının train datalarına olan uzaklıklarını tutan matris
distances=zeros(length(testdatax),50);
for i=1:length(testdatax)
    for k=1:50
    dist=0;
    for j=1:21
        dist=dist+(efacetestx(j,i)-emat4(j,k))^2;
    end
    distances(i,k)=dist^(1/2);
    end
end
correct=0;
for i=1:length(testdatax)
[m,nearest(i,1)]=min(distances(i,:));
predictions(i,1)=traindatay(nearest(i,1),1);
end
nofcp=zeros(10,1);%number of correct predictions
confmat=zeros(10,10);%karışıklık matrisi
crt=[];%doğru tahmin edilen örnekler 1, yanlışlar 0
for j=1:10
    for i=1:length(testdatax)
        if predictions(i,1)==j-1 && predictions(i,1)==testdatay(i,1)
        nofcp(j,1)=nofcp(j,1)+1;
        crt(i)=1;
        end
    end
end
correct=sum(nofcp);

for i=1:length(testdatax)
    realvalue=testdatay(i,1)+1;
```

```
    for j=1:10
        if predictions(i,1)==j-1
            confmat(realvalue,j)=confmat(realvalue,j)+1;
        end
    end
end
for i=1:10
    for j=1:10
        confmat2(i,j)=confmat(i,j)/sum(confmat(i,:));
    end
end
%örnek tahmin resimleri
for i=1:8
    for j=1:8
        mat1(i,j)=testdatax(3679,8*(i-1)+j);
    end
end
im1=mat2gray(mat1);
imshow(im1);
```