

## SVD, Toplu Tavsiye (Collaborative Filtering)

Diyelim ki Star Trek (ST) dizisini ne kadar beğendigini 4 tane kullanıcı sezonlara göre isaretlemiş. Bu örnek veriyi alttaki gibi gösterelim.

```
%pylab inline
from pandas import *

d = np.array(
    [[5, 5, 0, 5],
     [5, 0, 3, 4],
     [3, 4, 0, 3],
     [0, 0, 5, 3],
     [5, 4, 4, 5],
     [5, 4, 5, 5]])

data = DataFrame (d.T,
    columns=['S1', 'S2', 'S3', 'S4', 'S5', 'S6'],
    index=['Ben', 'Tom', 'John', 'Fred'])
print data
```

	S1	S2	S3	S4	S5	S6
Ben	5	5	3	0	5	5
Tom	5	0	4	0	4	4
John	0	3	0	5	4	5
Fred	5	4	3	3	5	5

Veriye göre Tom, ST dizisinin 3. sezonunu 4 seviyesinde sevmiş. 0 değeri o sezonun seyredilmediğini gösteriyor.

Toplu Tavsiye algoritmaları verideki diğer kişilerin bir ürünü, diziyi, vs. ne kadar beğendiklerinin verisinin diğer "benzer" kişilere tavsiye olarak sunulabilir, ya da ondan önce, bir kişinin daha almadığı ürünü, seyretmediği sezonu, dinlemediği müziği ne kadar "beğeneceğini" tahmin eder. Kaggle sitesi üzerinden yapılan unlu Netflix yarışmasının amacı buydu - ayrıca tahmin edilen ve gerçek beğeni notunun hata payının hesabı için RMSE hesabı kullanılmıştı.

Peki benzerliğin kriteri nedir, ve benzerlik nelerin arasında ölçülür?

Benzerlik, ürün seviyesinde, ya da kişi seviyesinde yapılabilir. Eğer ürün seviyesinde ise, tek bir ürün için tüm kullanıcıların verdiği nota bakılır. Eğer kullanıcı seviyesinde ise, tek kullanıcının tüm ürünlere verdiği beğeni notları vektörü kullanılır.

Mesela 1. sezondan hareketle, o sezonu beğenen kişilere o sezona benzer diğer sezonlar tavsiye edilebilir. Kisiden hareketle, mesela John'a benzeyen diğer kişiler bulunarak onların beğendiği ürünler John'a tavsiye edilebilir.

Ürün ya da kişi bazında olsun, benzerliği hesaplamanın da farklı yolları var. Genel olarak benzerlik ölçütünün 0 ile 1 arasında değişen bir sayı olmasını tercih ediyoruz

ve tum ayarlari ona gore yapiyoruz. Diyelim ki ki elimizde begeni notlarini tasiyan  $A, B$  vektorleri var (baska veri turu de tasiyor olabilir tabii), ve bu vektorlerin icinde begeni notlari var. Benzerlik cesitleri soyle:

#### Oklit Benzerligi (Euclidian Similarity)

Bu benzerlik  $1/(1+mesafe)$  olarak hesaplanir. Mesafe karelerin toplamının karekoku (yani Oklitsel mesafe, ki isim buradan geliyor). Bu yuzden mesafe 0 ise (yani iki "sey" arasinda hic mesafe yok, birbirlerine cok yakinlar), o zaman hesap 1 dondurur (mukemmel benzerlik). Mesafe arttikca bolen buyudugu icin benzerlik sifira yaklasir.

#### Pearson Benzerligi

Bu benzerligin Oklit'ten farkliligi, sayi buyuklugune hassas olmamasidir. Diyelim ki birisi her sezonu 1 ile begenmis, digeri 5 ile begenmis, bu iki vektorun Pearson benzerligine gore birbirine esit cikar. Pearson -1 ile +1 arasinda bir deger dondurur, alttaki hesap onu normalize ederek 0 ile 1 arasina ceker.

#### Kosinus Benzerligi (Cosine Similarity)

iki vektoru geometrik vektor olarak gorur ve bu vektorlerin arasinda olusan aciye (daha dogrusu onun kosinusunu) farklilik olcutu olarak kullanir.

$$\cos \theta = \frac{A \cdot B}{||A|| ||B||}$$

```
%pylab inline
from numpy import linalg as la
def euclid(inA,inB):
    return 1.0/(1.0 + la.norm(inA - inB))

def pearson(inA,inB):
    if len(inA) < 3 : return 1.0
    return 0.5+0.5*np.corrcoef(inA, inB, rowvar = 0)[0][1]

def cos_sim(inA,inB):
    num = float(np.dot(inA.T,inB))
    denom = la.norm(inA)*la.norm(inB)
    return 0.5+0.5*(num/denom)

print np.array(data.ix['Fred'])
print np.array(data.ix['John'])
print np.array(data.ix['Ben'])
print pearson(data.ix['Fred'],data.ix['John'])
print pearson(data.ix['Fred'],data.ix['Ben'])

[5 4 3 3 5 5]
[0 3 0 5 4 5]
```

```
[5 5 3 0 5 5]
0.551221949943
0.906922851283
```

```
print cos_sim(data.ix['Fred'],data.ix['John'])
print cos_sim(data.ix['Fred'],data.ix['Ben'])

0.898160909799
0.977064220183
```

Simdi tavsiye mekanigine gelelim. En basit tavsiye yontemi, mesela kisi bazli olarak, bir kisiye en yakin diger kisileri bulmak (matrisin tamamina bakarak) ve onların begendikleri urunu istenilen kisiye tavsiye etmek. Benzerlik icin ustteki olcutlerden birini kullanmak.

Fakat belki de elimizde cok fazla urun, ya da kullanıcı var. Bir boyut azaltma islemi yapamaz miyiz?

Evet. SVD yontemi burada da isimize yarar.

$$A = USV$$

elde edecegimiz icin, ve  $S$  icindeki en buyuk degerlere tekabul eden  $U, V$  degerleri siralanmis olarak geldigi icin  $U, V$ 'nin en bastaki degerlerini almak bize "en onemli" bloklari verir. Bu en onemli kolon ya da satirlari alarak azaltilmis bir boyut icinde benzerlik hesabi yapmak islemlerimizi hizlandirir. Bu azaltilmis boyutta kumeleme algoritmalarini devreye sokabiliriz;  $U$ 'nun mesela en onemli iki kolonu bize iki boyut-taki sezon kumelerini verebilir,  $V$ 'nin en onemli iki (en ust) satiri bize iki boyutta bir kisi kumesi verebilir.

O zaman begeni matrisi uzerinde SVD uygulayalim,

```
from numpy.linalg import linalg as la
U,Sigma,V=la.svd(data, full_matrices=False)
print data.shape
print U.shape, Sigma.shape, V.shape
u = U[:, :2]
vt=V[:, :2].T
print 'u', u
print 'vt', vt
print u.shape, vt.shape

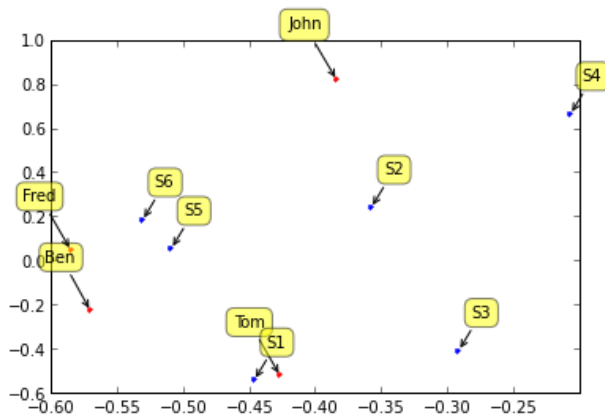
(4, 6)
(4, 4) (4,) (4, 6)
u [[-0.57098887 -0.22279713]
   [-0.4274751  -0.51723555]
   [-0.38459931  0.82462029]
   [-0.58593526  0.05319973]]
```

```
vt [[-0.44721867 -0.53728743]
    [-0.35861531  0.24605053]
    [-0.29246336 -0.40329582]
    [-0.20779151  0.67004393]
    [-0.50993331  0.05969518]
    [-0.53164501  0.18870999]]
(4, 2) (6, 2)
```

degerleri elimize gecer. U ve VT matrisleri

```
def label_points(d,xx,yy,style):
    for label, x, y in zip(d, xx, yy):
        plt.annotate(
            label,
            xy = (x, y), xytext = style,
            textcoords = 'offset points', ha = 'right', va = 'bottom',
            bbox = dict(boxstyle = 'round,pad=0.5', fc = 'yellow', alpha = 0.5),
            arrowprops = dict(arrowstyle = '->', connectionstyle = 'arc3,rad=0'))

plt.plot(u[:,0],u[:,1], 'r.')
label_points(data.index, u[:, 0], u[:, 1],style=(-10, 30))
plt.plot(vt[:,0],vt[:,1], 'b.')
label_points(data.columns, vt[:, 0], vt[:, 1],style=(20, 20))
plt.savefig('svdrecom_1.png')
```



Cok guzel! SVD bize urun bazinda sezon 5 ve 6'nin bir kume olusturdugunu, Ben ve Fred'in de kisi bazinda ayri bir kume oldugunu gosterdi.

Azaltilmis boyutlari nasil kullaniriz? Yeni bir kisiyi (mesela Bob) ele alinca, bu kisinin verisini oncelikle aynen diger verilerin indirgendigi gibi azaltilmis boyuta "indirgememiz" gerekiyor. Cunku artik islem yaptigimiz boyut orasi. Peki bu indirgemeyi nasil yapariz? SVD genel formulu hatirlarsak,

$$A = USV$$

Azaltılmış ortamda

$$A = U_k S_k V_k$$

Diyelim ki gitmek istedigimiz nokta azaltılmış  $U$ , o zaman  $U_k$ 'yi tek basına birakalım,

$$AV_k^{-1} = U_k S V_k V_k^{-1}$$

$U_k, V_k$  matrisleri ortonormal, o zaman  $V_k^{-1} V_k = I$  olacak, yani yokolacak

$$AV_k^{-1} = U_k S$$

Benzer şekilde

$$AV_k^{-1} S^{-1} = U_k$$

Çok fazla ters alma işlemi var, her iki tarafın devriginin alalım

$$(S^{-1})^T (V_k^{-1})^T A^T = U_k^T$$

$V_k^{-1} = V_k^T$  o zaman üstteki formül devrigin devriginin almak demektir, yani tekrar basa dönmüş oluyoruz, demek ki  $V_k$  değişmeden kalıyor

$$(S^{-1})^T V_k A^T = U_k^T$$

$S$  ise köşegen matris, onun tersi yine köşegen, köşegen matrisin devriği yine kendisi

$$S^{-1} V_k A^T = U_k^T$$

Bazı kod ispatları,  $u$ 'nın ortonormal olması:

```
print np.dot(u.T,u)
```

```
[[ 1.00000000e+00  4.83147593e-18]
 [ 4.83147593e-18  1.00000000e+00]]
```

Doğal olarak  $1e-17$  gibi bir sayı sifıra çok yakın, yani sıfır kabul edilebilir. Devrik ve tersin aynı olduğunu gösterelim: İki matrisi birbirinden çıkartıp, çok küçük bir sayıdan büyüklüğe göre filtreleme yapalım, ve sonuç içinde bir tane bile True olup olmadığını kontrol edelim,

```
print not any(U.T-la.inv(U) > 1e-15)
```

True

Yeni Bob verisi

```
bob = np.array([5,5,0,0,0,5])
```

O zaman

```
print bob.T.shape
print u.shape
S_k = np.eye(2)*Sigma[:2]
bob_2d = np.dot(np.dot(la.inv(S_k),vt.T),bob.T)
print bob_2d

(6,)
(4, 2)
[-0.37752201 -0.08020351]
```

Ustte `eye` ve `Sigma` ile ufak bir takla attik, bunun sebebi `svd` cagrisindan gelen `Sigma` sonucunun bir vektor olması ama ustteki islem icin kosegen bir "matrise" ihtiyacimiz olması. Eger birim (identity) matrisini alip onu `Sigma` ile carparsak, bu kosegen matrisi elde ederiz.

Simdi mesela kosinus benzerligi kullanarak bu izdusumlennmis yeni vektorun hangi diger vektorlere benzedigini bulalim.

```
for i,user in enumerate(u):
    print data.index[i],cos_sim(user,bob_2d)
```

```
Ben 0.993397525045
Tom 0.891664622942
John 0.612561691287
Fred 0.977685793579
```

Sonuca gore yeni kullanıcı Bob, en çok Ben ve Fred'e benziyor. Sonuca eristik! Artık bu iki kullanıcının yüksek not verdiği ama Bob'un hiç not vermediği sezonları alıp Bob'a tavsiye olarak sunabiliriz.

Movielens 1M Verisi

Bu veri seti 6000 kullanıcı tarafından yaklaşık 4000 tane filme verilen not / derece (rating) verisini içeriyor, 1 milyon tane not verilmiş, yani  $4000 * 6000 = 24$  milyon olasılık içinde sadece 1 milyon veri noktası dolu. Bu oldukça seyrek bir matris demektir.

Verinin ham hali diğer ders notlarımızı içeren üst dizinlerde var, veriyi SVD ile kullanılır hale getirmek için bu dizindeki `movielens_prep.py` adlı script kullanılır. İşlem bitince `movielens.csv` adlı bir dosya script'te görülen yere yazılacak. Bu

dosyada olmayan derecelendirmeler, verilmemiş notlar boş olacaktır. Bu boşlukları sıfırlarsak, seyrek matrisi o noktaları atlar. Ardından bu seyrek matris üzerinde seyrek SVD işletilebilir. Bu normal SVD'den daha hızlı işleyecektir.

Tavsiye kodlamamız için yazının başında anlatılan tekniği kullanacağız, film verisi üzerinde boyut azaltılması yapılacak, benzer kullanıcı bulunacak, ve herhangi bir yeni kullanıcı / film kombinasyonu için bu diğer benzer kullanıcının o filme verdiği not baz alınacak.

Veriyi eğitim ve test olarak iki parçaya böleceğiz. SVD eğitim bölümü üzerinde işletilecek.

Bu bağlamda, önemli bir diğer konu eksik veri noktalarının SVD sonuçlarını nasıl etkileyeceği. Sonuçta eksik yerler `nan`, oradan sıfır yapıp ardından seyrek matris kodlaması üzerinden "atlanıyor" olabilir, fakat bu değerler atlanıyor (yani hızlı işleniyor, depolanıyor) olsa bile, onların sıfır olmasının bir anlamı yok mudur? Evet vardır. Not bakımından sıfır da bir not'tur, ve bu sebeple sonuçları istenmeyen biçimde etkileyebilir.

O zaman mevcut veriyi öyle bir değiştirelim ki verilmemiş notlar, yani sıfır değerleri sonucu fazla değiştirmesin.

Bunu yapmanın yollarından biri her film için bir ortalama not değeri hesaplamak, ve bu ortalama değeri o filme verilen tüm not değerlerinden çıkartmaktır. Bu işleme "sıfır çevresinde merkezlemek" ismi de verilir, hakikaten mesela film *j* için ortalama 3 ise, 5 değeri 2, 3 değeri sıfır, 2 değeri -1 haline gelecektir. Bu bir ilerlemedir çünkü ortalama 3 değeri zaten bizim için "önemsiz" bir değerdir, tavsiye problemi bağlamında bizim en çok ilgilendığımız sevilen filmler, ve sevilmeyen filmler. Bu değerler sırasıyla artı ve eksi değerlere dönüşecekler, ve SVD bu farklılığı matematiksel olarak kullanabilme yeteneğine sahip.

Altta Pandas `mean` çağırışı ile bu işlemin yapıldığını görüyoruz, dikkat, Pandas dataframe içinde `nan` değerleri olacaktır, ve Pandas bu değerleri atlaması gerektiğini bilir, yani bu değerler ortalamaya etki etmez. Ardından merkezleme işlemi eğitim verisi üzerinde uygulanıyor.

```
%pylab inline
import pandas as pd, os
import scipy.sparse as sps
df = pd.read_csv("%s/Downloads/movielens.csv" % os.environ['HOME'], sep=';')
print df.shape
df = df.ix[:,1:] # id kolonunu atla
df = df.ix[:, :3700] # sadece filmleri al
df_train = df.copy().ix[:5000,:]
df_test = df.copy().ix[5001:,:]
df_train[np.isnan(df_train)] = 0.0
movie_avg_rating = np.array(df_train.mean(axis=0))
df_train = df_train - movie_avg_rating
dfs_train = sps.coo_matrix(df_train)
```

```

df_train = np.array(df_train)
df_test = np.array(df_test)

print df_train.shape
print df_test.shape

__top_k__ = 10
import scipy.sparse.linalg as slin
import scipy.linalg as la
U,Sigma,V=slin.svds(dfs_train,k=__top_k__)
print U.shape, Sigma.shape, V.shape
Sigma = np.diag(Sigma)

(6040, 3731)
(5001, 3700)
(1039, 3700)
(5001, 10) (10,) (10, 3700)

```

Altta test verisi üzerinde satır satır ilerliyoruz, ve her satır (test kullanıcısı) içinde film film ilerliyoruz. "Verilmiş bir not" arıyoruz (cogunlukla not verilmemiş oluyor cunku), ve buldugumuz zaman artık elimizde test edebilecegimiz bir sey var, o notu "sifirlayip" vektorun geri kalanini azaltılmış boyuta yansitıyoruz, ve sonra o boyut-taki tum diger  $U$  vektorleri içinde arama yapıyoruz, en yakin diger kullanıcıyı buluyoruz ve onun bu filme verdigi notu tahminimiz olarak kullanıyoruz.

Altta eger bulunan diger kullanıcı o filme not vermemisse, basitlestirme amaclı olarak, o filmi atlادık. Gercek dunya sartlarında filme not vermiş ve yakin olan (en yakin olmasa da) ikinci, ucuncu kullanıcılar bulunup onun notu kullanılabilir.

```

def euclid(inA,inB):
    return 1.0/(1.0 + la.norm(inA - inB))

rmse = 0; n = 0
for i,test_row in enumerate(df_test):
    for j, test_val in enumerate(test_row):
        if np.isnan(test_val): continue
        curr = test_row.copy()
        curr[j] = np.nan
        curr[np.isnan(curr)] = 0.

        proj_row = np.dot(np.dot(la.inv(Sigma),V),curr)

        sims = np.array(map(lambda x: euclid(x, proj_row), U[:,:__top_k__]))
        isim = np.argmax(sims)

        # eger bulunan kullanıcı o filme not vermemisse atla
        if np.isnan(df.ix[isim, j]): continue

```



```

# egitim verisinde notlar sifir etrafinda ortalannmis, tekrar
# normal haline dondur
est = df_train[isim, j]+movie_avg_rating[j]

# gercek not
real = df_test[i, j]

print i, 'icin en yakin', isim, 'urun',j, 'icin oy', est, 'gercek', real
rmse += (real-est)**2
n += 1
break # her kullanicici icin tek film test et
if i == 20: break # 20 kullanicici test et

print "rmse", np.sqrt(rmse / n)

0 icin en yakin 1903 urun 144 icin oy 5.0 gercek 5.0
1 icin en yakin 239 urun 144 icin oy 5.0 gercek 5.0
2 icin en yakin 2045 urun 844 icin oy 4.0 gercek 4.0
3 icin en yakin 4636 urun 0 icin oy 3.0 gercek 4.0
4 icin en yakin 139 urun 845 icin oy 4.0 gercek 5.0
5 icin en yakin 427 urun 1107 icin oy 4.0 gercek 5.0
6 icin en yakin 3620 urun 31 icin oy 4.0 gercek 4.0
7 icin en yakin 1870 urun 0 icin oy 4.0 gercek 3.0
8 icin en yakin 4816 urun 106 icin oy 5.0 gercek 5.0
9 icin en yakin 3511 urun 0 icin oy 3.0 gercek 4.0
10 icin en yakin 3973 urun 1212 icin oy 5.0 gercek 4.0
11 icin en yakin 2554 urun 287 icin oy 4.0 gercek 5.0
12 icin en yakin 4733 urun 31 icin oy 4.0 gercek 3.0
13 icin en yakin 2339 urun 9 icin oy 4.0 gercek 3.0
14 icin en yakin 3036 urun 10 icin oy 4.0 gercek 3.0
15 icin en yakin 2748 urun 253 icin oy 5.0 gercek 5.0
16 icin en yakin 450 urun 16 icin oy 4.0 gercek 4.0
17 icin en yakin 1133 urun 9 icin oy 5.0 gercek 2.0
18 icin en yakin 3037 urun 253 icin oy 5.0 gercek 4.0
19 icin en yakin 1266 urun 107 icin oy 3.0 gercek 3.0
20 icin en yakin 537 urun 253 icin oy 5.0 gercek 5.0
rmse 0.975900072949

```

Sonuc fena degil. Tavsiye programlarında RMSE 0.9 civari iyi olarak bilinir, Netflix yarismasinda [4] mesela kazanan algoritma RMSE 0.85'e erismistir.

Kaynaklar

[1] <http://www.igvita.com/2007/01/15/svd-recommendation-system-in-ruby/>

[2] Harrington, P., Machine Learning in Action

[3] <http://stats.stackexchange.com/questions/31096/how-do-i-use-the-svd-in-collaborat>

[4] [http://en.wikipedia.org/wiki/Netflix\\_Prize](http://en.wikipedia.org/wiki/Netflix_Prize)