SI2020_dataprep_exercise_svd_v1

PFR

August,2020

PFR data prep exercises for dimension reduction

This is an R Markdown document for data prep exercises.

This exercise is to run SVD and possibly reduce dimensions of the data

##load data

```
W_df_orig = read.table('weather_orig.csv',
                       header=TRUE,sep=",",
                       stringsAsFactors = TRUE)  #try TRUE

#Keep rows that are NOT missing data
keep_ind  = complete.cases(W_df_orig)
W_df      = W_df_orig[keep_ind,]

Y=as.numeric(W_df[,'RainTomorrow'])  #save this for later

# remove this redundant column
W_df = subset(W_df, select=-c(RISK_MM))

dim(W_df)
```

```
## [1] 328  23
```

##select numeric columns First, SVD and PCA only work on numeric columns, so we have to only keep the numeric columns

```
# Get numeric or integer columns only
col_classes    = sapply(W_df,class)   #get column classes as a list
num_inds       = c(which(col_classes=='numeric'),
                   which(col_classes=='integer'))

W_dfnum       = W_df[,num_inds]
dim(W_dfnum)
```

```
## [1] 328  16
```

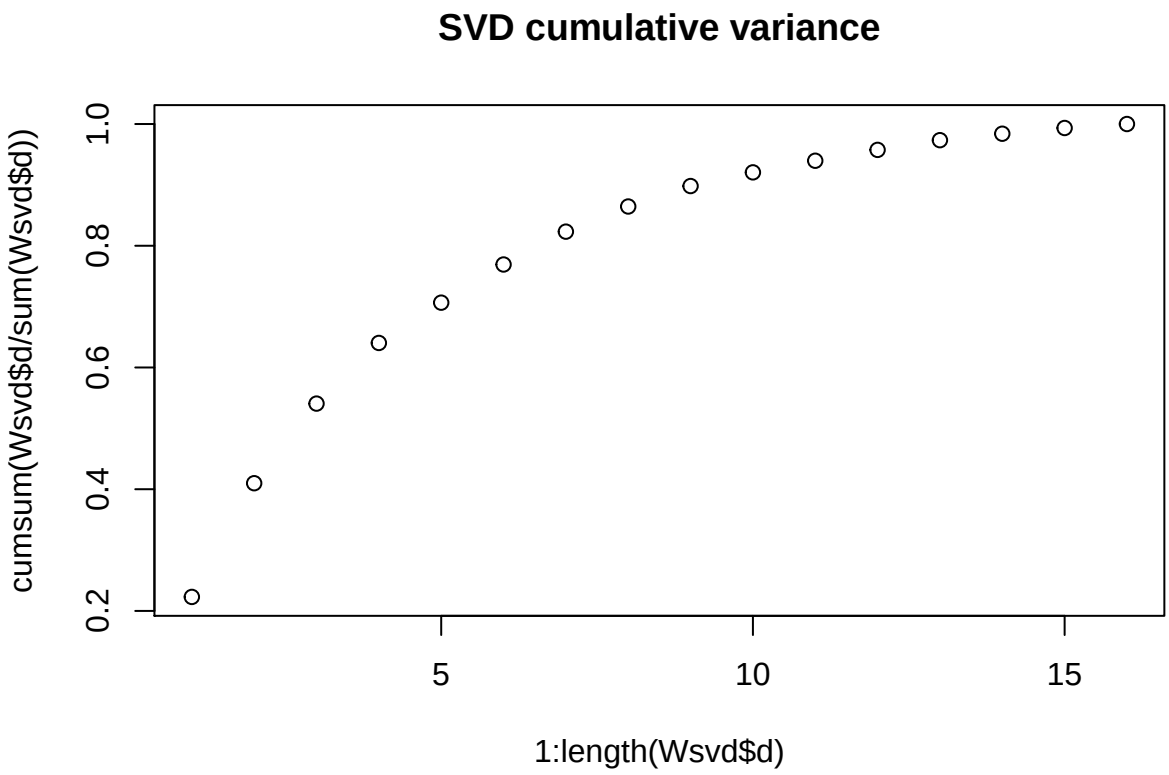
##Now mean center data

```
#use 'scale' function
W_mncntr=scale(W_dfnum,center=TRUE,scale=FALSE)
```

##Now run SVD

```
# the singular values are in the 'Wsvd$d' variable
# the factors are in the 'Wsvd$u' and 'Wsvd$v' variables
Wsvd=svd(W_mncntr)
```

```r
#plot the cumulative variance that each factor accounts for
plot(1:length(Wsvd$d),cumsum(Wsvd$d/sum(Wsvd$d)),main='SVD cumulative variance')
```



##Now lets reduce the dimensions, what's a reasonable amount of total variance that we have captured; conversely how much can we ignore

```r
#One could take first 3 components as an approximation to original data, for example

numcomp = 3

#NOTE  the %*% is matrix multiplication
W_dfred = Wsvd$u[,1:numcomp] %*% diag(Wsvd$d[1:numcomp]) %*% Wsvd$v[1:numcomp,1:3]

dim(W_dfred)
```

```
## [1] 328   3
```