Data Reduction Techniques

Diane DeOcampo - Data_607.001

What is Data Reduction?

Ch 12, pg 304 - Data Science for Business textbook

Data Reduction - A large/wide set of data and replace with a smaller set that preserves much of the important information

Also known as Dimensionality Reduction

Pros: easier to process, less resource extensive

Tradeoff: details in insight for manageability gained for the information lost

Data Reduction Technique: Missing Values Ratio

Determining data columns with too many missing values

Columns with missing data greater than a threshold can be removed

The higher the threshold, the more aggressive the reduction

Data Reduction Technique: Low Variance Filtering

The variance is range dependent, favors wide variance

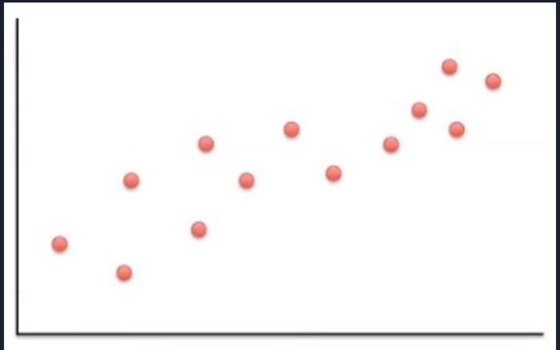
Similar to Missing Values Ratio, the higher the set threshold, the more aggressive the data reduction

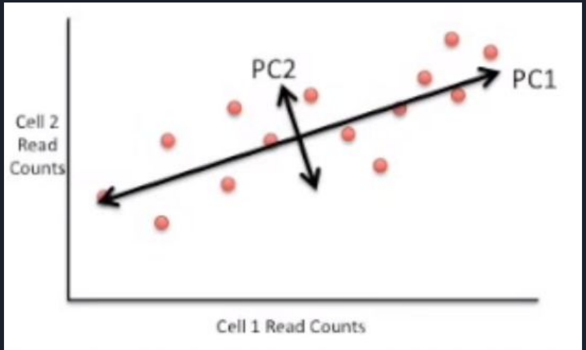
Data Reduction Technique: Principal Component Analysis (PCA)

Linear mapping original variables

where the greatest variance ie important variables (first principal component) and second greatest variance is on the second coord, etc

The data is normalized to be on the same scale





Principal Component Analysis (PCA) Example

Usage of prcomp (from stats package) returns Standard Deviation and Cumulative Proportion

```
pca_res <- prcomp(gapminder_life, scale=TRUE)
```

```
summary(pca_res)

## Importance of components:
##                          PC1     PC2     PC3     PC4     PC5
## Standard deviation     3.360 0.69114 0.40463 0.19246 0.11371
## Proportion of Variance 0.941 0.03981 0.01364 0.00309 0.00108
## Cumulative Proportion  0.941 0.98083 0.99448 0.99756 0.99864
```

```
names(pca_res)

[1] "sdev"     "rotation" "center"   "scale"    "x"
```

Principal Component Analysis (PCA) Example

```
pca_res$x[1:5,1:3]

##                             PC1        PC2        PC3
## Africa_Algeria        0.4518264 -1.3208553  0.3848907
## Africa_Angola        -5.2217443  0.2876153 -0.2574503
## Africa_Benin         -2.2956809 -0.2847236  0.0080484
## Africa_Botswana      -0.6460076  1.1788076  0.8922150
## Africa_Burkina Faso -3.3832822 -0.3287683  0.1598156
```

```
pca_res$center[1:5]

## lifeExp_1952 lifeExp_1957 lifeExp_1962 lifeExp_1967 lifeExp_1972
##     48.38172     50.57246     52.54620     54.26249     55.98415
```

```
head(pca_res$scale^2, n=5)

## lifeExp_1952 lifeExp_1957 lifeExp_1962 lifeExp_1967 lifeExp_1972
##     181.2131     181.8949     177.7479     168.3315     157.4524
```

Other Data Reduction Techniques

Random Forests

Backward Feature Elimination

Forward Feature Construction

High Correlation Filter

Reference/Additional Readings:

Prcomp function:
https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/prcomp

PCA in R: https://blog.learningtree.com/dimensionality-reduction-in-r/

https://www.r-bloggers.com/principal-component-analysis-using-r/

https://cmdlinetips.com/2019/04/introduction-to-pca-with-r-using-prcomp/

Dimensionality Reduction Techniques (with Python codes):
https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/