Data Mining

Unit # 12

Principal Component Analysis

- Principal comonent analysis (PCA) or Factor Analysis (FA) statistical techniques applied to a single set of variables where the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another.
- Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors.
- Factors are thought to reflect underlying processes that have created the correlations among variables.

Fundamental Steps

- Steps in PCA or FA include
  - Selecting and measuring a set of variables
  - Preparing the covariance matrix
  - Extracting a set of factors from the covariance matrix
  - Determining the number of factors
  - interpreting the results
- Although there are relevant statistical considerations to most of these steps, an important test of the analysis is its interpretability.

Application

- PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension.

Limitation

- One of the problems with PCA and FA is that there is no criterion variable against which to test the solution.
- In regression analysis, for instance, the dependent variable (DV) is a criterion and the correlation between observed and predicted DV scores serves as a test of the solution
- In classification, the solution is judged by how well it predicts group membership.
- But in PCA or FA there is no external criterion such as group membership against which to test the solution.

Practical Issues

- Because FA and PCA are exquisitely sensitive to the sizes of correlations, it is critical that honest, reliable correlations be employed.
- Sensitivity to outlying cases, problems created by missing data, and degradation of correlations between poorly distributed variables all plague FA and PCA.

Variance and Covariance

- Standard deviation and variance only operate on one dimension.
- However, it is useful to have a similar measure to find out how much the dimensions vary from the mean *with respect to each other.*
- Covariance is such a measure. Covariance is always measured *between* 2 dimensions.
- If you calculate the covariance between one dimension and *itself, you get the* variance.
- So, if you had a 3-dimensional data set (x, y, z), then you could measure the covariance between the x an y dimensions, the x and z dimensions, and the y and z dimensions.

Variance and Covariance (Cont'd)

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Year	Gold	Dollar
2006	1233	60.8
2007	1581	61.5
2008	2075	79.1
2009	3070	83.5
2010	3835	86.5
2011	4720	90.2
2012	5500	97.8
2013	4250	110.1

Eigen Values and Eigen Vectors

- Let *A* be a square matrix. A non-zero vector *x* is called an **eigenvector** of *A* if and only if there exists a number (real or complex) λ such that

$$Ax = \lambda x$$

- If such a number λ exists, it is called an **eigenvalue** of *A*. The vector *x* is called eigenvector associated to the eigenvalue .

Eigen Vectors

- Eigenvectors can only be found for square matrices.
- And, not every square matrix has eigenvectors.
- And, given an nxn matrix that does have eigenvectors, there are n of them.
- For example, given a 3x3 matrix, there are 3 eigenvectors.

Example (Source: Wikipedia)

For the matrix *A*

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

the vector

$$\mathbf{x} = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

is an eigenvector with eigenvalue 1. Indeed,

$$A\mathbf{x} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 3 + 1 \cdot (-3) \\ 1 \cdot 3 + 2 \cdot (-3) \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1 \cdot \begin{bmatrix} 3 \\ -3 \end{bmatrix}.$$

On the other hand the vector

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

is *not* an eigenvector, since

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \cdot 0 + 1 \cdot 1 \\ 1 \cdot 0 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

and this vector is not a multiple of the original vector **x**.

Eigen Value Computation

- When a transformation is represented by a square matrix A, the eigen value equation can be expressed as $Ax - \lambda x = 0$
- Where I is the identify matrix. This can be rearranged to $(A - \lambda I)x = 0$
- If there exists an inverse $(A - \lambda I)^{-1}$ then both sides can be left multiplied by the inverse to obtain the trivial solutions: $x = 0$. Thus we require there to be no inverse by assuming from linear algebra that the determinants equals zero:
- $\det(A - \lambda I) = 0$
- To compute eigen vectors, solve for $Ax = \lambda x$ for all values of λ.

Example

Exercises

For the following square matrix:

$$\begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & 2 \\ -6 & 0 & -2 \end{pmatrix}$$

Decide which, if any, of the following vectors are eigenvectors of that matrix and give the corresponding eigenvalue.

$$\begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

Working of PCA

- The original data are projected onto a much smaller space, resulting in dimensionality reduction.
- Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA “combines” the essence of attributes by creating an alternative, smaller set of variables.
- The initial data can then be projected onto this smaller set.

Steps 1 and 2

- Step 1:
  - The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.
- Step 2:
  - PCA computes *k orthonormal vectors that provide a basis for the normalized input* data.
  - These vectors are referred to as the *principal components.*
  - *The input data are a linear* combination of the principal components.

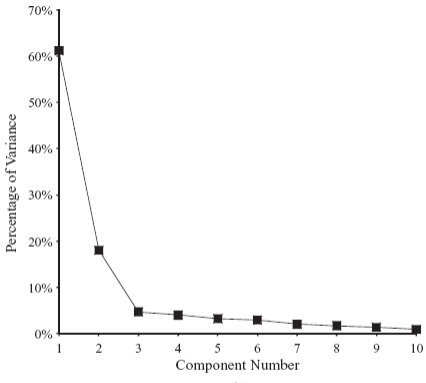
Steps 3 and 4

- Step 3:
  - The principal components are sorted in order of decreasing "significance" or strength.
  - The principal components essentially serve as a new set of axes for the data, providing important information about variance.
  - That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on.
- Step 4:
  - Because the components are sorted according to decreasing order of "significance," the size of the data can be reduced by eliminating the weaker components, that is, those with low variance.
  - Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

Example
(Source: Witten et al.)

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%

(a)



(b)

FIGURE 7.5

Principal components transform of a dataset: (a) variance of each component and (b) variance plot.

KNIME: PCA DEMO

Abstract

- The Big Data challenge is becoming one of the most exciting opportunities for the next years.
- Big Data is a new term used to identify the datasets that due to their large size and complexity, we can not manage them with our current methodologies or data mining soft-ware tools.
- Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it.

What Is Big Data?

- There is not a consensus as to how to define big data

"Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population." - *Teradata Magazine article, 2011*

"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." - *The McKinsey Global Institute, 2011*

Big Data Mining

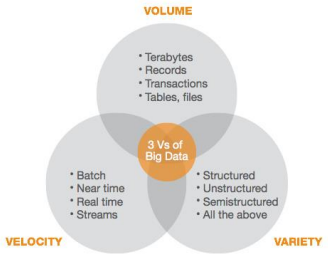
- The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress" [9].
- Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [34] .
- However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [8].
- The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day.

Big Data Mining (Cont'd)

- Usama Fayyad [11] in his invited talk at the KDD BigMine'12 Work-shop presented amazing data numbers about internet usage, among them the following:
  - each day Google has more than 1 billion queries per day,
  - Twitter has more than 250 milion tweets per day,
  - Facebook has more than 800 million updates per day, and
  - YouTube has more than 4 billion views per day.
- The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year.

3 V's

- Doug Laney[19] was the first one in talking about 3 V's in Big Data management:
  - Volume: there is more data than ever before, its size continues increasing, but not the percent of data that our tools can process
  - Variety: there are many different types of data, as text, sensor data, audio, video, graph, and more
  - Velocity: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time



How Is Big Data Different?

1) Automatically generated by a machine
   (e.g. Sensor embedded in an engine)

2) Typically an entirely new source of data
   (e.g. Use of the internet)

3) Not designed to be friendly
   (e.g. Text streams)

4) May not have much values
   - Need to focus on the important part

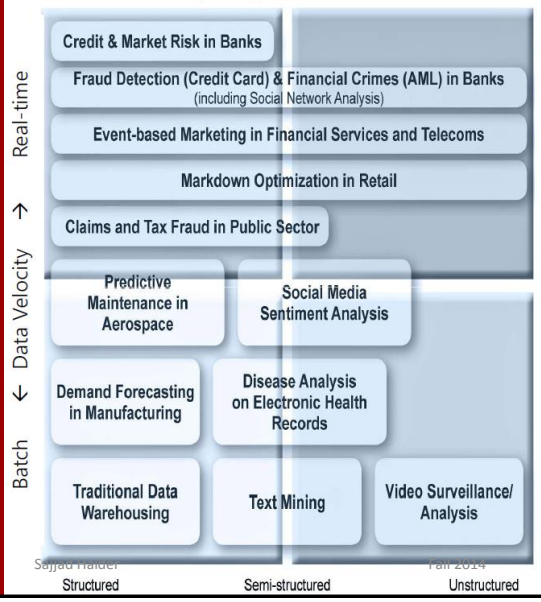
Risks of Big Data

- Will be so overwhelmed
  - Need the right people and solve the right problems
- Costs escalate too fast
  - Isn't necessary to capture 100%
- Many sources of big data is privacy
  - self-regulation
  - Legal regulation

Why You Need to Tame Big Data

- Analyzing big data is already standard
  (e.g. ecommerce)
- Be left behind in a few years
  - So far, only missed the chance on the bleeding edge
- Capturing data, using analysis to make decisions
  - Just an extension of what you are already doing today

The Structure of Big Data

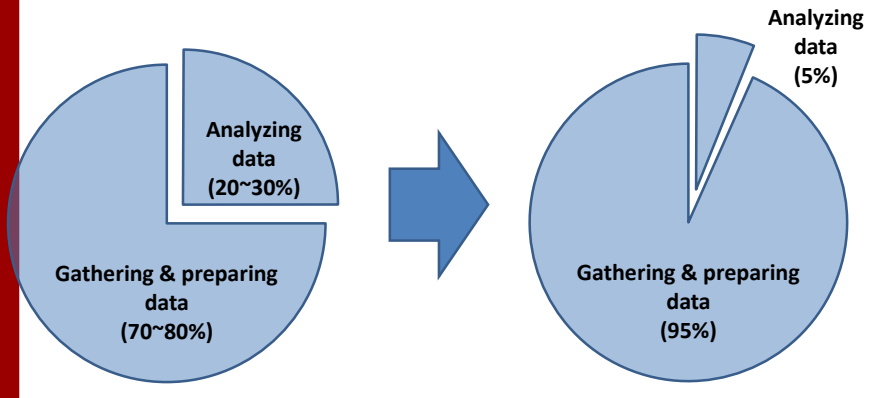


- Structured
  - Most traditional data sources
- Semi-structured
  - Many sources of big data
- Unstructured
  - Video data, audio data

Exploring Big Data

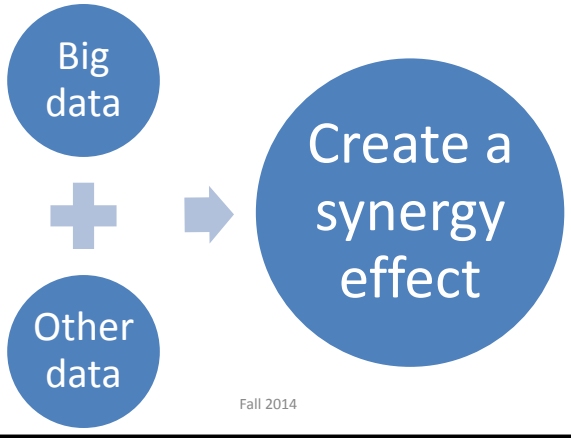
- The time for developing an analysis

- The time for developing an analysis (Initially working with big data)



Mixing Big Data with Traditional Data

- The biggest value in big data can be driven by combing big data with other corporate data



Mixing Big Data with Traditional Data

- Browsing history
  - Knowing how valuable a customer is
  - What they have bought in the past
- Text (Online chat and e-mails)
  - Knowing the detailed product specification being discussed
  - The sales data related those products

Who Benefits from Big Data

- Companies with a tradition of fact-based decision making
- Engineering and research functions
- The best web-native companies

Big Data Controversies

- There is no need to distinguish Big Data analytics from data analytics, as data will continue growing, and it will never be small again.
- Big Data may be a hype to sell Hadoop based computing systems. Hadoop is not always the best tool [23]. It seems that data management system sellers try to sell systems based in Hadoop, and MapReduce may be not always the best programming platform, for example for medium-size companies.
- In real time analytics, data may be changing. In that case, what it is important is not the size of the data, it is its recency.

Big Data Controversies (Cont'd)

- Claims to accuracy are misleading. As Taleb explains in his new book [32], when the number of variables grow, the number of fake correlations also grow. For example, Leinweber [21] showed that the S&P 500 stock index was correlated with butter production in Bangladesh, and other funny correlations.
- Bigger data are not always better data. It depends if the data is noisy or not, and if it is representative of what we are looking for. For example, some times Twitter users are assumed to be a representative of the global population, when this is not always the case.

Big Data Controversies (Cont'd)

- Ethical concerns about accessibility. The main issue is if it is ethical that people can be analyzed without knowing it.
- Limited access to Big Data creates new digital divides. There may be a digital divide between people or organizations being able to analyze Big Data or not. Also organizations with access to Big Data will be able to extract knowledge that without this Big Data is not possible to get. We may create a division between Big Data rich and poor organizations.

Tools: Open Source Revolution

- Apache Hadoop [3]: software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes.
- A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.

Big Data Mining Tools

- Apache Mahout [4]: Scalable machine learning and datamining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.
- R [29]: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets.

Concluding Remarks on Big Data

- Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year.
- This data is going to be more diverse, larger, and faster.
- Big Data is becoming the new Final Frontier for scientific data research and for business applications.
- We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before.

Analytics and Decision Making

- Google, Amazon, and others have prospered not by giving customers information but by giving them shortcuts to decisions and actions. (Analytics 3.0, Harvard Business Review December 2013)

Course Recap

Course Objectives

- Know the knowledge discovery process
- Understand the different categories of algorithms
- Be able to judge which algorithms fit different problems
- Have practical experience choosing algorithms for a specific problem
- Have practical experience working in technical teams
- Have practical experience executing data mining projects
- Have practical experience using open source data mining software

Course Outlines

- Classification Techniques
  - Classification/Decision Trees
  - Naïve Bayes
  - Neural Networks
- Clustering
  - Partitioning Methods
  - Hierarchical Methods
- Feature Selection
- Model Evaluation
- Patterns and Association Mining
- Text Mining
- Papers/Case Studies/Applications Reading

What Next?

- An excellent online book on Data Mining with R by Yanchang Zhao
  - http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf
- Another great online book on Mining of Massive Data by several professors of Stanford
  - http://infolab.stanford.edu/~ullman/mmds/book.pdf

What Next? (Cont'd)

- Cloudera has provided a Virtual Machine of Hadoop. You can simply download the VM file and open it in Virtual Box to learn and to experiment with Hadoop and its echo system.
  - http://www.cloudera.com/content/support/en/downloads.html
- Keep visiting Kaggle.com and other similar websites to keep your knowledge and skill up to date. Wish you a strong career in analytics!