Data Mining: How Companies use Linear Algebra

Ralph Abbey, Carl Meyer

NCSU

MAA Southeastern Section: 26th, March 2010
1. Introduction
   - Data Mining

2. Linear Regression

3. Eigenvalues & Eigenvectors

4. Latent Semantic Indexing
Why should you care about linear algebra?

Ralph Abbey, Carl Meyer

Data Mining: How Companies use Linear Algebra
Why should you care about linear algebra?
The process of extracting meaningful information from data.

Who does this, why?

Search Engines, Stock Services, Banks, Retail Chains, etc.

Data mining offers a huge potential for increased profits.

Why doesn't everyone use data mining?

Not enough resources, not enough potential for gain for the cost, more pressing short term concerns.

Ralph Abbey, Carl Meyer

Data Mining: How Companies use Linear Algebra
Data Mining

- The process of extracting meaningful information from data.
The process of extracting meaningful information from data.

Who does this, why?

Search Engines, Stock Services, Banks, Retail Chains, etc. Data mining offers a huge potential for increased profits. Why doesn’t everyone use data mining?
The process of extracting meaningful information from data.

Who does this, why?

Search Engines, Stock Services, Banks, Retail Chains, etc. Data mining offers a huge potential for increased profits. Why doesn’t everyone use data mining?

Not enough resources, not enough potential for gain for the cost, more pressing short term concerns.
One of the most common procedures in data mining. A very simple and cheap way of mining data. Often seen more in statistics books than math books. We would be more used to seeing the linear system $Ax = b$. 

Ralph Abbey, Carl Meyer

Data Mining: How Companies use Linear Algebra
Linear Regression

- One of the most common procedures in data mining.
One of the most common procedures in data mining.

A very simple and cheap way of mining data.
Linear Regression

- One of the most common procedures in data mining.
- A very simple and cheap way of mining data.
- Often seen more in statistics books than math books.
Linear Regression

- One of the most common procedures in data mining.
- A very simple and cheap way of mining data.
- Often seen more in statistics books than math books.
- We would be more used to seeing the linear system $Ax = b$. 
There are many methods for solving including: Gaussian Elimination, Multiplying by Inverse, Conjugate Gradient Method, GMRES, etc.

For Conjugate Gradient, for example, we need $A$ from $Ax = b$ to be symmetric, positive-definite (spd). $A = A^T$ $x^T Ax > 0$ for all $x > 0$ (each entry in $x$ is positive).
Ax = b

There are many methods for solving including:
\[ Ax = b \]

- There are many methods for solving including:
  - Gaussian Elimination, Multiplying by Inverse, Conjugate Gradient Method, GMRES, etc.
There are many methods for solving including:

- Gaussian Elimination, Multiplying by Inverse, Conjugate Gradient Method, GMRES, etc.

For Conjugate Gradient, for example, we need $A$ from $Ax = b$ to be symmetric, positive-definite (spd).
There are many methods for solving including:

- Gaussian Elimination, Multiplying by Inverse, Conjugate Gradient Method, GMRES, etc.

For Conjugate Gradient, for example, we need $A$ from $Ax = b$ to be symmetric, positive-definite (spd).

- $A = A^T$
There are many methods for solving including:

- Gaussian Elimination, Multiplying by Inverse, Conjugate Gradient Method, GMRES, etc.

For Conjugate Gradient, for example, we need $A$ from $Ax = b$ to be symmetric, positive-definite (spd).

- $A = A^T$
- $x^tAx > 0$ for all $x > 0$ (each entry in $x$ is positive).
Why do Linear Algebraists love Eigenvalues and Eigenvectors more than their wives?

Ax = λx: λ is the eigenvalue corresponding to the eigenvector x

Used in Principal Component Analysis, studying the behavior of Markov Chains, (differential equations), other clustering methods.
Why do Linear Algebraists love Eigenvalues and Eigenvectors more than their wives?

- Lots of beautiful theory - and it’s everywhere!
Why do Linear Algebraists love Eigenvalues and Eigenvectors more than their wives?

- Lots of beautiful theory - and it’s everywhere!
- $Ax = \lambda x$: $\lambda$ is the eigenvalue corresponding to the eigenvector $x$
Why do Linear Algebraists love Eigenvalues and Eigenvectors more than their wives?

- Lots of beautiful theory - and it’s everywhere!

- $Ax = \lambda x$: $\lambda$ is the eigenvalue corresponding to the eigenvector $x$

- Used in Principal Component Analysis, studying the behavior of Markov Chains, (differential equations), other clustering methods.
Principal Component Analysis

Google finds over 4 million for a normal search, and over 3 million for a scholar search

Used in clustering, categorizing, finding direction of maximal variance

Ralph Abbey, Carl Meyer

Data Mining: How Companies use Linear Algebra
Principal Component Analysis

- $X$ is the data matrix, and the mean of the each row is stored in the vector $u$
Principal Component Analysis

- $X$ is the data matrix, and the mean of the each row is stored in the vector $u$

- $B = X - u \ast e^T$ ($e$ is the vector of all ones)
**Principal Component Analysis**

- $X$ is the data matrix, and the mean of the each row is stored in the vector $u$

- $B = X - u \cdot e^T$ ($e$ is the vector of all ones)

- Find the eigenvalues and eigenvector of the covariance matrix $C = B^T B$
Principal Component Analysis

- $X$ is the data matrix, and the mean of the each row is stored in the vector $u$

- $B = X - u \cdot e^T$ ($e$ is the vector of all ones)

- Find the eigenvalues and eigenvector of the covariance matrix $C = B^T B$

- Google finds over 4 million for a normal search, and over 3 million for a scholar search
Principal Component Analysis

- $X$ is the data matrix, and the mean of the each row is stored in the vector $u$

- $B = X - u \cdot e^T$ ($e$ is the vector of all ones)

- Find the eigenvalues and eigenvector of the covariance matrix $C = B^T B$

- Google finds over 4 million for a normal search, and over 3 million for a scholar search

- Used in clustering, categorizing, finding direction of maximal variance
Latent Semantic Indexing

Precursor to modern search engines
Finds 'latent' semantic meaning
Makes use of the Singular Value Decomposition (SVD)
Latent Semantic Indexing

- Precursor to modern search engines
Latent Semantic Indexing

- Precursor to modern search engines

- Finds ‘latent’ semantic meaning
Latent Semantic Indexing

- Precursor to modern search engines
- Finds ‘latent’ semantic meaning
- Makes use of the Singular Value Decomposition (SVD)