Data Mining - What is it and why would anyone want to do it?

Dr. Cheryl Howard
Kilmer Middle School Career Day
November 18, 2008

Cheryl@datamininglab.com
What is Data Mining?

- Data mining is the discovery of new and unexpected patterns from large data sets that can be used to solve problems
  - It takes advantage of large databases, fast computers, and advanced statistics and math
  - Data mining often “learns by example”
  - It works well on complex problems that don’t always have obvious solutions
  - We are usually trying to classify things (e.g., fraud/not fraud) or predict (e.g., “What else would Joe like to buy from Amazon?”)
What Do Data Miners Do?

- Make online shopping more convenient for the user (and more profitable for the site!)
- Help drug companies discover new drugs more quickly
- Help businesses identify and keep their most valuable customers (and sell them more stuff!)
- Catch bad guys
  - People who cheat on their taxes
  - Identity thieves
  - Terrorists
Who Needs Data Miners?
Why Do Data Mining?

• Most organizations have more data than they can handle
• They can use this data to make their business more efficient and profitable
• Data Mining systems can automatically do the tedious work, giving experts more time to deal with the tough, interesting problems
• It can do things that a few experts are good at, but can’t quite explain how they do it
What Kinds of Data are Used?

- Structured data
  - Numbers (age, income, amount of taxes paid)
  - Ordered sets {low, medium, high}
  - Unordered sets {blond, brunette, redhead}
- Text
  - Strings of characters
  - Complete documents
  - Web pages, blogs, e-mails
  - Phone numbers, addresses, URLs
- Images
  - Photographs
  - Radar and infrared
  - Medical images (e.g., x-rays or MRI’s)
- Sound
What Makes Someone Good at Data Mining?

- Strong math and analytical skills
- An understanding of experimental design and the scientific method
- Interest in different areas, like computer science, linguistics, and business
- Attention to detail
How Long Do I Have to Go to School*?

*Informal survey of my 11 colleagues at Elder Research, Inc.
What do I have to Study*?

Major Field of Study

*Informal survey my 11 colleagues at Elder Research, Inc.
So, where does the hard math come in?

Support Vector Machines

Given training vectors \( x_i \in R^d, i = 1, \ldots, l \) in two classes, and a vector \( y \in R^d \) such that \( y_i \in \{-1, 1\} \). SVM solves the following dual problem:

\[
\min \quad f(\alpha) = \frac{1}{2} \alpha^T Q \alpha - \epsilon^T \alpha
\]

such that \( 0 \leq \alpha_i \leq C, i = 1, \ldots, l \) and \( y^T \alpha = 0 \), where

\[
\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l)^T
\]

and \( Q \) is an \( l \times l \) matrix, \( Q(\alpha_i, \alpha_j) = y_i y_j K(x_i, x_j) \)

The decision function is

\[
\text{sign} \left( \sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b \right)
\]

where \( b \) is a constant term.

Decision Tree

Kernel

Nearest Neighbor

Neural Network (or Polynomial Network)
Good Things About Being a Data Miner

• Solving lots of different, interesting, important problems
• There’s always more to learn
• You don’t have to wear a suit (very often)
• Good pay and job security
  (there’s no shortage of bad guys out there!)