CS 175: Project in Artificial Intelligence

Slides 1: Introduction
Logistics: Staff

• Teaching staff:
  – Instructor: Arthur Asuncion
    • asuncion@uci
    • Office hours: Friday 3-4pm, BH 4059
  – TA: Yutian Chen
    • yutian.chen@uci
    • Office hours: Tuesday 2-3:30pm, BH 4059
Logistics: Prerequisites

CS 175 (a project course)

CS 171

ICS 23

“You are a good programmer”

Math 2A/B

“You know how to compute a derivative”

Stats/Math 67

“You are familiar with probability”
Logistics: Grading

• Project: 70%

• 3-4 HW assignments: 30%

• No Midterms/Finals
  – But we may use Finals week for presentations

• Late HW policy: 20% off for each day late
Goals

• Create a working and useful AI system
  – Our focus: data mining / machine learning

• Become exposed to a wide variety of ML algorithms

• Learn skills such as Matlab, LaTeX (optional), data processing techniques

• Learn how to effectively collaborate
Lectures

• I will teach on various topics in machine learning
  – See course web page: https://eee.uci.edu/10s/34340

• Some classes will be devoted to assessing project progress

• Please do not hesitate to ask questions
Course Project

1. Start with an interesting task and find real-world data

2. Perform research to find out appropriate data mining / machine learning algorithms

3. Implement several different algorithms

4. Evaluate the performance of the algorithms on data (if unsuccessful, return to step 3, 2, or 1)

5. Write up results
Team project details

• Teams of 3 people are ideal

• The amount of work performed should be proportional to the number of people on the team

• Select team members wisely!

• Multiple teams can work on the same project (e.g. Yahoo’s “Learning to Rank” challenge may be a popular project)
Rough Timeline

• Week 3: Project Proposal

• Week 5: Progress Report 1

• Week 7: Progress Report 2

• End of Quarter:
  – Code deliverables
  – Technical report
  – Final presentation

Warning: The quarter goes by quickly, especially for a project course like CS175.
Required Software: Matlab

• Matlab: “MATrix LABoratory”

• Available in CS364 labs (certain computers) and MSTB labs. Also, can purchase from UCI bookstore

• Required for HW assignments

• Projects can be in Matlab or in Java/C++.

• I will give a brief demo later on if we have time.
Matlab = Blue dots

- FINCH
- SANDHILL-CRANE
- BEACH-MOUSE
- WHITE-STURGEON
- GULF-STURGEON
- STICKLEBACK
- STEELHEAD
- SQUAWFISH
- SILVERSIDE
- SPIKEDACE
- DELTA-SMELT
- BEAUTIFUL SHINER
- BONTO-BLUEGILL
- RED-SPOTTED-BLACKBASS
- WASHINGTON-BLACKBASS
- SOUTHERN-CUTTHROAT
- GIANT-CATFISH
- WESTERN-CATFISH
- BONTAI-FISH
- YELLOW-POND-FISH
- SOUTHERN-COHO-SALMON
- WYOMING-TOAD
- WAQUA-CATFISH
- OZARK-CATFISH
- BONITA-CHUB
- OREGON-CHUB
- SONORAN-CHUB
- RIVER-PELICAN
- SAND-GOOSE
- HAWK-BIRD
- MAUI-PARROT
- SPOTTED-OWL
- GULL-EGRET
- KINGFISHER
How to do well in this course
(and in life)

• Be proactive
• Be productive
• Be resourceful
• Be a leader
• Don’t procrastinate
• Don’t be afraid
• Don’t give up
Topic 1: Data Mining

Slides taken from Prof. Smyth (with slight modifications)
Introduction to Data Mining

• What is data mining?

• Data sets
  – The “data matrix”
  – Other data formats

• Data mining tasks
  – Exploration
  – Description
  – Prediction
  – Pattern finding

• Data mining algorithms
  – Score functions, models, and optimization methods

• The dark side of data mining
What is data mining?
What is data mining?

“The magic phrase used to .......
- put in your resume
- use in a proposal to funding agencies
- use to get venture capital funding
- sell database software
- sell statistical analysis software
- sell parallel computing hardware
- sell consulting services”
What is data mining?

“Data-driven discovery of models and patterns from massive observational data sets”
What is data mining?

“Data-driven discovery of models and patterns from massive observational data sets”

Statistics, Inference
What is data mining?

"Data-driven discovery of models and patterns from massive observational data sets"

Languages and Representations

Statistics, Inference
What is data mining?

“Data-driven discovery of models and patterns from massive observational data sets”

Statistics, Inference

Languages and Representations

Engineering, Data Management
What is data mining?

“Data-driven discovery of models and patterns from massive observational data sets”

Languages and Representations

Engineering, Data Management

Statistics, Inference

Retrospective Analysis
In simple terms….two primary goals
Technological Driving Factors

• Larger, cheaper memory
  – Moore’s law for magnetic disk density
    “capacity doubles every 18 months”
  – storage cost per byte falling rapidly

• Faster, cheaper processors
  – the CRAY of 15 years ago is now on your desk

• Success of relational databases and the Web
  – everybody is a “data owner”

• New ideas in machine learning/statistics
  – Boosting, SVMs, decision trees, non-parametric Bayes, text models, etc
Examples of massive data sets

- MEDLINE text database
  - Records for 19 million published articles

- Web search engines
  - Multiple billion Web pages indexed
  - 100’s of millions of site visitors per day

- CALTRANS loop sensor data
  - Every 30 seconds, thousands of sensors, 2Gbytes per day

- NASA MODIS satellite
  - Coverage at 250m resolution, 37 bands, whole earth, every day

- Retail transaction data
  - Ebay, Amazon, Walmart: >100 million transactions per day
  - Visa, Mastercard: similar or larger numbers
Instant Messenger Data

Jure Leskovec and Eric Horvitz, 2007

240 million IM users over 1 month
1.3 billion edges in the graph
The $1 Million Question

Data set with 480,000 users, 17,000 movies, and 100 million movie ratings
Two Types of Data

- **Experimental Data**
  - Hypothesis H
  - design an experiment to test H
  - collect data, infer how likely it is that H is true
  - e.g., clinical trials in medicine

- **Observational or Retrospective or Secondary Data**
  - massive non-experimental data sets
    - e.g., Web logs, human genome, atmospheric simulations, etc
  - assumptions of experimental design no longer valid
  - how can we use such data to do science?
    - use the data to support model exploration, hypothesis testing
Data-Driven Discovery

• Observational data
  – cheap relative to experimental data
  • Examples:
    – Transaction data archives for retail stores, airlines, etc
    – Web logs for Amazon, Google, etc
    – The human/mouse/rat genome
    – Etc., etc
  ⇒ makes sense to leverage available data
  ⇒ useful (?) information may be hidden in vast archives of data
Data Mining v. Statistics

• Traditional statistics
  – first hypothesize, then collect data, then analyze
  – often model-oriented (strong parametric models)

• Data mining:
  – few if any a priori hypotheses
  – data is usually already collected a priori
  – analysis is typically data-driven not hypothesis-driven
  – Often algorithm-oriented rather than model-oriented

• Different?
  – Yes, in terms of culture, motivation: however…..
  – statistical ideas are very useful in data mining, e.g., in validating whether discovered knowledge is useful
  – Increasing overlap at the boundary of statistics and DM e.g., exploratory data analysis (work of John Tukey in the 1960’s)
Data Mining v. Machine Learning

• To first-order, very little difference....
  – Data mining relies heavily on ideas from machine learning (and from statistics)

• Some differences between DM and ML:
  – More emphasis in DM on scalability, e.g.,
    • algorithms that can work on data that is outside main memory
    • analyzing data in a relational database (reflects database “roots” of DM)
    • analyzing data streams
  – DM is somewhat more applications-oriented
    • Higher visibility in industry and in public
    • ML is somewhat more theoretical, research oriented
Origins of Data Mining

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Pattern Recognition</td>
<td>Hardware (sensors, storage, computation)</td>
<td>Relational Databases</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>“Pencil and Paper”</td>
<td>“Data Dredging”</td>
<td>EDA</td>
<td>“Flexible Models”</td>
<td></td>
</tr>
</tbody>
</table>
DM: Intersection of Many Fields

- Machine Learning (ML)
- Statistics (stats)
- Visualization (viz)
- Computer Science (CS)
- Databases (DB)
- Human Computer Interaction (HCI)
- High-Performance Parallel Computing
Data in Matrix Form

### Measurements

<table>
<thead>
<tr>
<th>ID</th>
<th>Income</th>
<th>Age</th>
<th>....</th>
<th>Monthly Debt</th>
<th>Good Risk?</th>
</tr>
</thead>
<tbody>
<tr>
<td>18276</td>
<td>65,000</td>
<td>55</td>
<td>....</td>
<td>2200</td>
<td>Yes</td>
</tr>
<tr>
<td>72514</td>
<td>28,000</td>
<td>19</td>
<td>....</td>
<td>1500</td>
<td>No</td>
</tr>
<tr>
<td>28163</td>
<td>120,000</td>
<td>62</td>
<td>....</td>
<td>1800</td>
<td>Yes</td>
</tr>
<tr>
<td>17265</td>
<td>90,000</td>
<td>35</td>
<td>....</td>
<td>4500</td>
<td>No</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>.....</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>.....</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>61524</td>
<td>35,000</td>
<td>22</td>
<td>....</td>
<td>900</td>
<td>Yes</td>
</tr>
</tbody>
</table>

“Measurements” may be called “variables”, “features”, “attributes”, “fields”, etc.
Sparse Matrix (Text) Data
Sequence (Web) Data

<table>
<thead>
<tr>
<th>User  1</th>
<th>2</th>
<th>3</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>1</th>
<th>3</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>User 3</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>User 4</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>User 5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Data Mining Lectures

Slides 1: Introduction

Padhraic Smyth, UC Irvine
Time Series Data

TRAJECTORIES OF CENTROIDS OF MOVING HAND IN VIDEO STREAMS

From Scott Gaffney, PhD Thesis, UC Irvine, 2005
Spatio-temporal data

From Scott Gaffney, PhD Thesis, UC Irvine, 2005
Image from work of Tucker Balch and Frank Dellaert, Computer Science Department, Georgia Tech
Relational Data

HP Labs email network
500 people, 20k relationships

How does this network evolve over time?
Different Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modeling
- Predictive Modeling
- Discovering Patterns and Rules
- + others....
Exploratory Data Analysis

• Getting an overall sense of the data set
  – Computing summary statistics:
    • Number of distinct values, max, min, mean, median, variance, skewness,..

• Visualization is widely used
  – 1d histograms
  – 2d scatter plots
  – Higher-dimensional methods

• Useful for data checking
  – E.g., finding that a variable is always integer valued or positive
  – Finding the some variables are highly skewed

• Simple exploratory analysis can be extremely valuable
  – You should always “look” at your data before applying any data mining algorithms
Example of Exploratory Data Analysis
(Pima Indians data, scatter plot matrix)
Different Data Mining Tasks

• Exploratory Data Analysis

• Descriptive Modeling

• Predictive Modeling

• Discovering Patterns and Rules

• + others…. 
Descriptive Modeling

- Goal is to build a “descriptive” model
  - e.g., a model that could simulate the data if needed
  - models the underlying process

- Examples:
  - Density estimation:
    • estimate the joint distribution $P(x_1, \ldots, x_p)$
  - Cluster analysis:
    • Find natural groups in the data
  - Dependency models among the $p$ variables
    • Learning a Bayesian network for the data
Example of Descriptive Modeling
Example of Descriptive Modeling

ANEMIA PATIENTS AND CONTROLS

EM ITERATION 25
Another Example of Descriptive Modeling

- Directed Graphical Models (aka Bayes Nets)
  - goal: learn a probability model with directed relationships among variables
  - representation: directed graphs
  - challenge: distinguishing between correlation and causation

  - example: Do yellow fingers cause lung cancer?

    hidden cause: smoking
Graphical Model for Gene Expression Data

Lin et al, PLOS Genetics, 2009

Data Mining Lectures                                            Slides 1: Introduction                     Padhraic Smyth, UC Irvine
Different Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modeling
- Predictive Modeling
- Discovering Patterns and Rules
- + others…. 
Predictive Modeling

• Predict one variable $Y$ given a set of other variables $X$
  – Here $X$ could be a $p$-dimensional vector
  – Classification: $Y$ is categorical
  – Regression: $Y$ is real-valued

• In effect this is function approximation, learning the relationship between $Y$ and $X$

• Many, many algorithms for predictive modeling in statistics and machine learning

• Often the emphasis is on predictive accuracy, less emphasis on understanding the model
Predictive Modeling: Fraud Detection

• Credit card fraud detection
  – Credit card losses in the US are over 1 billion $ per year
  – Roughly 1 in 50k transactions are fraudulent

• Approach
  – For each transaction estimate $p(\text{fraudulent} \mid \text{transaction})$
  – Model is built on historical data of known fraud/non-fraud
  – High probability transactions investigated by fraud police

• Example:
  – Fair-Isaac/HNC’s fraud detection software based on neural networks, led to reported fraud decreases of 30 to 50%

• Issues
  – Significant feature engineering/preprocessing
  – false alarm rate vs missed detection – what is the tradeoff?
Predictive Modeling: Customer Scoring

• Example: a bank has a database of 1 million past customers, 10% of whom took out mortgages

• Use machine learning to rank new customers as a function of $p(\text{defaults on mortgage} \mid \text{customer data})$

• Customer data
  – History of transactions with the bank
  – Other credit data (obtained from Experian, etc)
  – Demographic data on the customer or where they live

• Techniques
  – Binary classification: logistic regression, decision trees, etc
  – Many, many applications of this nature
Different Data Mining Tasks

• Exploratory Data Analysis

• Descriptive Modeling

• Predictive Modeling

• Discovering Patterns and Rules

• + others….
Pattern Discovery

• Goal is to discover interesting “local” patterns in the data rather than to characterize the data globally.

• Given market basket data we might discover that
  • If customers buy wine and bread then they buy cheese with probability 0.9
  • These are known as “association rules”

• Given multivariate data on astronomical objects
  • We might find a small group of previously undiscovered objects that are very self-similar in our feature space, but are very far away in feature space from all other objects.
Example of Pattern Discovery

ADACABDABAABBBDDBCCADDDDBCCDDBCCBCBCCDADADAADADADBDDABABBCDD
DCDDABDCCBDCBDBCBBBABBBCBBAABCBBACBBDDBBAACCADDADBBDCCBBBBCBB
BDCABDBBADDDBBBCCACDABBABDDCDDBBABBDDBDDDBCACDBBCBCCBBCAC
DCADCBACCADCCACCCACCDDADBCADADBAACCDDEDDBDBCDCCCACACACCDAB
DDBCADADBCBDDADABCCABDAACACABCACBDDBDCCBADCDBADDCCDCCADC
CBBADABAADDAAABCBCBCABDBAADCBCDADDCCBCABABCCBCBACBDABDDDAAA
BADCDCCDBBDCBDDDCBCBDBAADDADBCCAAADBDCAADDBBBBCDCBCCBCCCD
CCADAADACABDABAABBDDCADDDBDBCDDDBCBDCCBCBCBCDADADACCDCDABAABBC
BDBDBADBBBCDADABABBDACDCCDDDBBBCBCCBCCDABCADDADBACBABC
CDBAAADDDBDCCABACBCACDACDCBAAADDAADDADAABBBACCBB
Example of Pattern Discovery

ADACABDABABAABBBDDBCADDDBCDDBCDBCDDBCCBBCCDADADADABDBBDABABBCDD
DCDDABDCCBBDDBDBCCBABBBCBACBCBBDBBAAACCADDADBDBBCBBFCBBCCBB
BDCABDDBBADDBBBBCCACDABBABDDCDDBBBABBDDBDBBDCACDBBCCBBAC
DCADCBACCADCCACCADDACBCADADBAACCACDDCDBDCCCCACACACCDAB
DDBCADADBCBDDADABCCABDAACABCABACBDDDBCBADCBDADDADDCDCDCAD
CBBADABBAAADAAAABCCBCABDBAADCBCDABCABABBCCBCBBABDDADDADAA
BADCDCCDBBCCDBDADDCCBBCCDBAADDADBCCAADDDBDADBDBBCDCCBCCCD
CCADAADADACABDABAAABBBDCADDDBCDDDBCDDBCBCCCBCCCDADADACCDBABAABC
BDBDDBADBBBCBDADABABBDACDCDDDBBCDBBCBCCDBBBCCDADBABACB
CDBAAADDDBDDBABCDABCADCDDBAAADDADCADADAABBAACCBB
Example of Pattern Discovery

• IBM “Advanced Scout” System
  – Bhandari et al. (1997)
  – Every NBA basketball game is annotated,
    • e.g., time = 6 mins, 32 seconds
      event = 3 point basket
      player = Michael Jordan
    • This creates a huge untapped database of information
  
  – IBM algorithms search for rules of the form
    “If player A is in the game, player B’s scoring rate increases from 3.2 points per quarter to 8.7 points per quarter”

  – IBM claimed around 1998 that all NBA teams except 1 were using this software...... the “other team” was Chicago.
General Issues in Data Mining

- Scalability
  - Time and space complexity
  - Parallelization, e.g., MAP-Reduce and Hadoop

- Evaluation
  - Do our results generalize to new data?

- Operational use
  - Will the algorithm require 6 PhDs to “babysit” it?

- Data Privacy
  - Often underestimated by technologists
Data Mining: the Dark Side

• Hype

• Data dredging, snooping and fishing
  – Finding spurious structure in data that is not real

• Historically, ‘data mining’ was a derogatory term in the statistics community
  – making inferences from small samples

• The challenges of being interdisciplinary
  – computer science, statistics, domain discipline
Today's Random Medical News

- Exercise
- Fatty foods
- Stress
- Red wine
- Computer terminals
- Coffee
- Pancake

Can cause:
- Hypothermia
- Heart disease
- Breast cancer
- Spontaneous remission
- Glaucoma
- Depression
- Sexual desire
- A feeling of well-being
- Sexual anxiety

In:
- Tremors
- Two-income families
- Arthritis sufferers
- 7 out of 10 women
- Men 25-40
- Overweight smokers
- Rats

According to a report released today...
Example of “data fishing”

- Example: data set with
  - 50 data vectors
  - 100 variables
  - Even if data are entirely random (no dependence) there is a very high probability some variables will appear dependent just by chance.
Rhine Paradox – (1)

- A parapsychologist in the 1950’s hypothesized that some people had Extra-Sensory Perception.
- He devised an experiment where subjects were asked to guess 10 hidden cards – red or blue.
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right.
Rhine Paradox – (2)

- He told these people they had ESP and called them in for another test of the same type
- Alas, he discovered that almost all of them had lost their ESP
- What did he conclude?

- He concluded that you shouldn’t tell people they have ESP; it causes them to lose it. 😊
Topic 2: Exploratory Data Analysis and Visualization

Slides taken from Prof. Smyth
(with slight modifications)
Exploratory Data Analysis (EDA)

• Get a general sense of the data

• Interactive and visual
  – (cleverly/creatively) exploit human visual power to see patterns
    • 1 to 5 dimensions (e.g. spatial, color, time, sound)
  – e.g. plot raw data/statistics, reduce dimensions as needed

• Data-driven (model-free)

• especially useful in early stages of data mining
  – detect outliers (e.g. assess data quality)
  – test assumptions (e.g. normal distributions or skewed?)
  – identify useful raw data & transforms (e.g. log(x))

• Bottom line: it is always well worth looking at your data!
Summary Statistics

Empirical statistics of data $X = X_1, \ldots, X_n$

- mean: $\mu = \frac{\sum_i X_i}{n}$ \{ $\mu$ minimizes $\sum_i (X_i - \mu)^2$ \}

- mode: most common value in $X$ (e.g., for integer or categorical data)

- median: $X=$sort($X$), median = $X_{n/2}$ (half below, half above)

- quartiles of sorted $X$: $Q1$ value = $X_{0.25n}$, $Q3$ value = $X_{0.75n}$
  - interquartile range: $value(Q3) - value(Q1)$
  - range: $max(X) - min(X) = X_n - X_1$

- variance: $\sigma^2 = \frac{\sum_i (X_i - \mu)^2}{n}$

- skewness: $\sum_i (X_i - \mu)^3 / \left[ (\sum_i (X_i - \mu)^2)^{3/2} \right]$
  - zero if symmetric; right-skewed more common (e.g. you v. Bill Gates)

- number of distinct values for a categorical variable (see unique.m in MATLAB)

- Note: all of these are estimates based on the sample at hand – they may be different from the “true” values (e.g., median age in US).
Exploratory Data Analysis

Tools for Displaying Single Variables
Histogram

- Most common form:
  - split data range into equal-sized bins.
  - for each bin, count the number of points from the data set that fall into the bin
  - y axis: frequency (e.g., counts for each bin)
  - x axis: values of the variable

- The histogram can illustrate features related to the distribution of the data, e.g.,
  - center (i.e., the location)
  - spread (i.e., the scale)
  - skewness
  - presence of outliers
  - presence of multiple modes

However, important to note that the histogram can also obscure these properties!
ZipCode Data: Population

K = 50

K = 500

K = 50
MATLAB code for ZipCode Data

• MATLAB code to generate previous slide:

\[
X = \text{zipcode\_data}(:,2) \quad \% \text{second column from zipcode array}
\]
\[
\text{histogram}(X, 50) \quad \% \text{histogram of } X \text{ with 50 bins}
\]
\[
\text{histogram}(X, 500) \quad \% 500 \text{ bins}
\]
\[
\text{index} = X < 5000; \quad \% \text{identify } X \text{ values lower than 5000}
\]
\[
\text{histogram}(X(\text{index}), 50) \quad \% \text{now plot just these } X \text{ values}
\]
Histogram Detecting Outlier (Missing Data)

blood pressure = 0 ?
Issues with Histograms

• For small data sets, histograms can be misleading.
  – Small changes in the data or bucket boundaries can result in very different histograms.
  – Modes may be missed or falsely introduced
  – Produces non-smooth estimate of the distribution

• Interactive bin-width example (online applet)
  – http://www.stat.sc.edu/~west/javahtml/Histogram.html

• For large data sets, histograms can be quite effective at illustrating general properties of the distribution.

• Can smooth histogram using a variety of techniques
  – E.g., kernel density estimation

• Histograms effectively only work with 1 variable at a time
  – Difficult to extend to 2 dimensions, not possible for >2
  – So histograms tell us nothing about the relationships among variables
Right Skewness Example
Exploratory Data Analysis

Tools for Displaying Pairs of Variables
BoxPlots

Y-axis: real-valued or integer variable (e.g., income)
X-axis: categorical variable (e.g., job category)

• For each group, the boxplot shows
  – median
  – interquartile range (25 to 75%)
  – “whiskers” (most extreme points not considered to be outliers)
  – Outliers, e.g., points > Q3 + W (Q3 – Q1),  W = 1.5 by default
    (about plus/minus 2.7 sigma, or 99.3 % of the data for Normally distributed data)

Does not generalize to more than 2 variables, although there is a two-dimensional analog for 2 real-valued variables: “bagplot”
Car Gas Mileage by Country of Manufacture

Type “help boxplot” in MATLAB to find out how to generate this plot
Pima Indians Data

Box contains middle 50% of data

Q3-Q1

Median

Plots all data outside whiskers

Up to 1.5 x Q3-Q1 (or shorter, if no data that far above Q3)
2D Scatter Plots

- standard tool to display relation between 2 variables
  - e.g. y-axis = response, x-axis = suspected indicator
- useful to answer:
  - x,y related?
    - no
    - linearly
    - nonlinearly
  - variance(y) depend on x?
  - outliers present?
- MATLAB:
  - plot(X(:,1),X(:,2),'.');

Two variables related to credit card repayment, each point is a card customer.
Scatter Plot: No apparent relationship
Scatter Plot: Linear relationship
Scatter Plot: Quadratic relationship
Scatter plot: Homoscedastic

Variation of Y Does Not Depend on X
Scatter plot: Heteroscedastic

variation in $Y$ differs depending on the value of $X$
eq. $Y = \text{annual tax paid}, \quad X = \text{income}$
Problems with Scatter Plots of Large Data

scatter plot degrades into black smudge ...
other techniques can be used (e.g. contour plots)
**SCATTER PLOT OF HUMIDITY VERSUS RAINFALL**

**BOX PLOT OF HUMIDITY VERSUS DISCRETIZED RAINFALL**

- **RAINFALL AMOUNT**
- **RELATIVE HUMIDITY**
- **DISCRETIZED RAINFALL AMOUNT**
- **RELATIVE HUMIDITY**
Exploratory Data Analysis

Tools for Displaying More than 2 Variables
Multivariate Visualization

• Multivariate -> multiple variables

• 2 variables: scatter plots, etc

• 3 variables:
  – 3-dimensional plots
  – Look impressive, but often not used
  – Can be cognitively challenging to interpret
  – Alternatives: overlay color-coding (e.g., categorical data) on 2d scatter plot

• 4 variables:
  – 3d with color or time
  – Can be effective in certain situations, but tricky

• Higher dimensions
  – Generally difficult
  – Scatter plots, icon plots, parallel coordinates: all have weaknesses
  – Alternative: “map” data to lower dimensions, e.g., PCA or multidimensional scaling
  – Main problem: high-dimensional structure may not be apparent in low-dimensional views
Scatter Plot Matrix

For interactive visualization the concept of “linked plots” is generally useful
Trellis Plot

http://netlib.bell-labs.com/cm/ms/departments/sia/project/trellis/
Using Icons to Encode Information, e.g., Star Plots

- Each star represents a single observation. Star plots are used to examine the relative values for a single data point.
- The star plot consists of a sequence of equi-angular spokes, called radii, with each spoke representing one of the variables.
- Useful for small data sets with up to 10 or so variables.
- Limitations?
  - Small data sets, small dimensions
  - Ordering of variables may affect perception

| 1 Price | 5 Headroom |
| 2 Mileage (MPG) | 6 Rear Seat Room |
| 3 1978 Repair Record (1 = Worst, 5 = Best) | 7 Trunk Space |
| 4 1977 Repair Record (1 = Worst, 5 = Best) | 8 Weight |
|         | 9 Length  |
Parallel Coordinates

(dimensions (possibly all p of them!))
often (re)ordered to better distinguish among interesting subsets of n total cases

interactive “brushing” is useful for seeing such distinctions

(epileptic seizure data from text)

1 (of n) cases
(this case is a “brushed” one, with a darker line, to stand out from the n-1 other cases)
More elaborate parallel coordinates example (from E. Wegman, 1999). 12,000 bank customers with 8 variables. Additional “dependent” variable is profit (green for positive, red for negative).
Example of displaying 4d categorical data, e.g., as used in OLAP/databases
Other aspects (not discussed)

• Cognitive and human-factors aspects of visualization
  – In creating visualizations of data it is important to be aware of how the human brain perceives visual information
  – E.g., “Rules and principles of scientific data visualization”
    • [http://www.siggraph.org/education/materials/HyperVis/percept/visrules.htm](http://www.siggraph.org/education/materials/HyperVis/percept/visrules.htm)

• Artistic aspects of visualization

• Visualization of other data
  – 2d, 3d, 4d “volume” data (fluid flow, brain images, etc)
  – Network/graph data
    • Issues: graph layout/drawing, issues of graph size
  – Many others….., e.g.,
    • CHI conference, etc
Visualization of weather states for Kenya

Daily data from 20 year history clustered into 3 different weather “states”

Mean image for each state
- wind direction (arrows)
- wind intensity (size of arrows)
- rainfall (size of circles)
- pressure (contours)

Multidimensional Scaling (MDS)

- Map 150D vectors into 2D
- Tries to preserve vector-vector dissimilarities in the lower dimension

“Topics” of UCI/UCSD research papers [Gretarsson, et al]
Summary

• EDA and Visualization
  – Can be very useful for
    • data checking
    • getting a general sense of individual or pairs of variables
  – But…
    • do not necessarily reveal structure in high dimensions

• In HW1, you will perform some EDA on a data set, using Matlab