

CS573: Homework 2

Due date: Thursday September 30, start of class

1 Identifying Hypotheses (15 pts)

1. Identify five potentially controversial hypotheses, claims, or conjectures in news articles that can be analyzed with data. See some examples claims below. Summarize the articles and the claims.
2. For each of your identified claims, match it to one of the following tasks:
 - (a) Classification
 - (b) Regression
 - (c) Pattern discovery
 - (d) Clustering
 - (e) Anomaly detection

Restate the claim and describe, in one or more paragraphs, how the claim could be analyzed as an example of the task to which you have matched it. Consider carefully whether you could obtain the needed data from web-accessible sources or elsewhere, since you will need to find at least one data source to use in question 2 below.

Example claims

During the past few years, distracted driving has evolved from a dangerous practice to a deadly epidemic and pressing public-safety crisis.

The newly-declared end-date to the recession also confirms what many had suspected: The 2007-9 recession was the deepest on record since the Great Depression, at least in terms of job losses.

The reality of any green product is that they generally dont work as well... From hybrid cars to solar panels, environmentally friendly alternatives can cost more. They can be less convenient, like toting cloth sacks or canteens rather than plastic bags or bottled water. And they can prove less effective, like some of the new cleaning products.

Ever since 1980, when Ronald Reagan inspired more men than women, the difference in the way men and women vote has been a significant part of American politics.

What is unfolding this year is only the second known global bleaching of coral reefs. Scientists are holding out hope that this year will not be as bad, over all, as 1998, the hottest year in the historical record, when an estimated 16 percent of the worlds shallow-water reefs died.

The Princeton researchers say the experiments suggest that high-fructose corn syrup prompts more weight gain than sucrose, at least in rats, even when the animals eat the same number of calories over all.

2 Working with Data (10 pts)

1. Identify, obtain, and prepare a data set to support one of the claims you identified in question 1. Data.gov is one possible place to look for public data that relates to claims found in news articles.

Your data must have $D > 500$, where $D = NA - M$.

- N is the number of instances (rows)
- A is the number of attributes (columns)
- M is the total number of missing data points

Make sure that $N > 40$, and $A > 4$ and do not count identifiers (e.g., name, unique code number, etc) toward A . In addition, make sure that the data have at least three continuous attributes and two discrete attributes.

For example, if you analyzed the House races in 2008 by looking at the party of the winner, the margin of victory, the party split in the district, whether there was an opponent, and campaign spending, then $N = 435$, $A = 5$, and M might be as large as 100 (because of unavailable data). Thus, $D = 2075$.

2. Write a description of the data that includes:
 - (a) Information about the data, including at least the following:
 - Data representation
 - Population, sample, and sampling mechanism
 - Attribute description that includes semantics (i.e., what they record) , type (e.g., nominal, ordinal), and range
 - Which of your claims from question 1 the data is appropriate for
 - (b) A discussion of possible data quality issues and an outline as to what might be an appropriate response to address the issues.

3 Data Exploration (10 pts)

Using your dataset from question 2, load your data into R¹, and do the following analysis:

1. Find two attributes with interesting histograms. Plot them and describe briefly why they are interesting.
2. Find two sets of two attributes that have interesting scatter plots or box plots. Plot them and describe briefly why they are interesting.

¹The class resource page has links to the R software, a tutorial, and a cheatsheet with basic command. If you prefer to use Matlab or some other other statistical package, that's fine—but the TA will not be able to support you for packages other than R.

3. Consider three pairs of continuous attributes and three pairs of discrete attributes. (If you don't have enough discrete attributes for this, then first discretize a continuous attribute to pair up with a discrete attribute.) Determine whether there are any pairs with strong associations. For pairs of continuous attributes, calculate correlation; for pairs of discrete attributes calculate χ^2 scores from contingency tables. Include assessments of significance.
4. Transform all discrete attributes into numerical attributes, center the data to have zero means, and run PCA on the data. Plot the scree plot and determine whether the data can be reduced to a smaller number of dimensions. Discuss how to choose an appropriate number of reduced dimensions. Report the details of the first two principal components (i.e., the component attributes and their weights), as well as the proportion of variance that they explain. Discuss whether this gives any insight into the claim that has been made about the data in question 1.