36-350: Data Mining

Handout 4
September 8, 2003

Finding the important dimensions in data

Similarity searching tells you the local structure of a dataset. But we also want to know the global structure. For example: What are the words which distinguish the auto/moto groups? What are the colors which distinguish the sailing/racing groups?

An **important** dimension is very **informative** about what group an object belongs to. Intuitively, this means that each group tends to have a distinct value on that dimension. For example, a word is important if it is more common in one group of documents than another.

Let x be a dimension and $\mathtt{x}$ be a particular value. The value $\mathtt{x}$ is informative if it is more likely to occur in one group than another. Let c be the group, so that $c = 1$ is the first group and $c = 2$ the second group. Then $\mathtt{x}$ is informative if

$$p(x = \mathtt{x} \mid c = 1) \neq p(x = \mathtt{x} \mid c = 2)$$

The entire dimension x is informative if its values tend to be informative.

This intuition can be represented formally by Bayes' rule:

$$\begin{aligned} p(c = \mathtt{c} \mid x) &= \frac{p(x = \mathtt{x} \mid c = \mathtt{c})p(c = \mathtt{c})}{p(x = \mathtt{x})} \\ \frac{p(c = 1 \mid x = \mathtt{x})}{p(c = 2 \mid x = \mathtt{x})} &= \frac{p(x = \mathtt{x} \mid c = 1)}{p(x = \mathtt{x} \mid c = 2)} \frac{p(c = 1)}{p(c = 2)} \end{aligned}$$

$p(c)$ is our uncertainty about c before observing x, and $p(c|x)$ is our uncertainty after observing x. Thus the ratio $p(x = \mathtt{x} \mid c = 1)/p(x = \mathtt{x} \mid c = 2)$ describes what we learn about c from observing $x = \mathtt{x}$.

Measuring information

Entropy is a measure of uncertainty (in **bits**):

$$\mathcal{H}(c) = -\sum_{\mathsf{c}} p(c=\mathsf{c}) \log_2 p(c=\mathsf{c})$$

$$\mathcal{H}(c \mid x) = -\sum_{\mathsf{c}} p(c=\mathsf{c} \mid x) \log_2 p(c=\mathsf{c} \mid x)$$

Entropy is large when $p(c)$ is flat (totally uncertain) and is zero when $p(c)$ is concentrated on one value (totally certain). Examples:

p(c=1)	p(c=2)	H(c)
1/2	1/2	1 bit
1/3	2/3	0.918 bits
0	1	0 bits

Information is the **change in uncertainty**:

$$\mathcal{I}(c, x=\mathtt{x}) = \mathcal{H}(c) - \mathcal{H}(c|x=\mathtt{x}) \qquad \text{(actual information)}$$

This is the information in a particular value $\mathtt{x}$. The information in the entire dimension is the average information in each value:

$$\mathcal{I}(c, x) = \sum_{\mathtt{x}} p(x=\mathtt{x}) \mathcal{I}(c, x=\mathtt{x}) \qquad \text{(expected information)}$$

$$= \mathcal{H}(c) - \sum_{\mathtt{x}} p(x=\mathtt{x}) \mathcal{H}(c|x=\mathtt{x})$$

Entropy is called **self information** because $\mathcal{I}(c, c) = \mathcal{H}(c)$.

Uncertainty usually decreases, but can also increase. For example, on a random day, it is unlikely to rain, so the uncertainty of "rain" is low, say 0.1 bits. But if you read a weather report giving a 50% chance of rain, your uncertainty increases to 1 bit. In these cases, the actual information is negative.

For example, suppose we pick a random position in a random document. Let x be 1 if the word is "car", and 0 otherwise. The frequencies for the 10 auto/moto documents are

```
          x
c         car not car
auto       13     598
not auto    0     696
```

For this table,

$$\mathcal{H}(c) = 0.997$$

$$\mathcal{H}(c|x = \text{``car''}) = 0$$

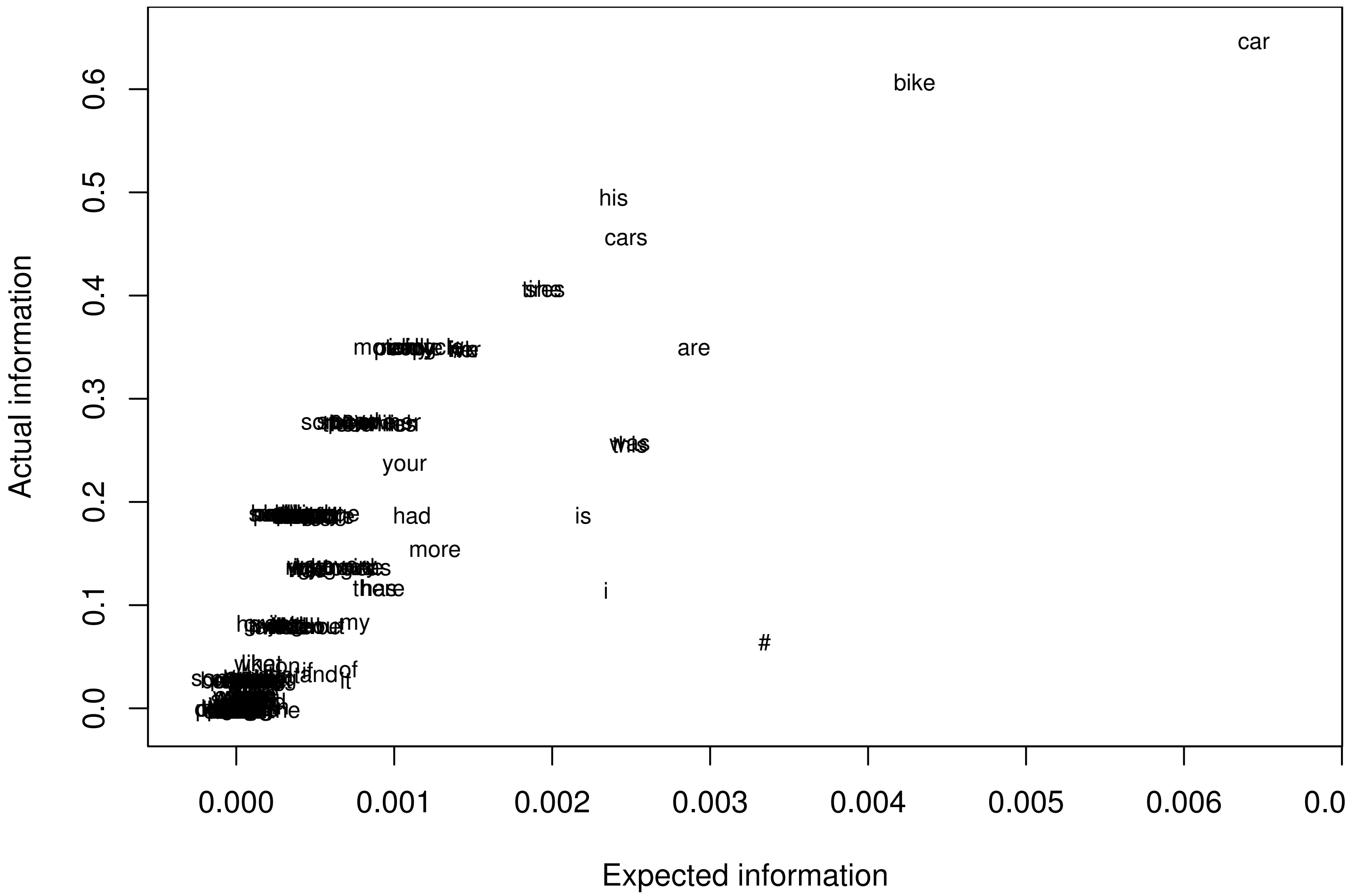
$$\mathcal{H}(c|x = \text{not ``car''}) = 0.996$$

$$p(x = \text{``car''}) = 0.01$$

$$\mathcal{I}(c, x) = 0.997 - (0.01 * 0) - (0.99 * 0.996) = 0.01$$

How to find the important words:

1. Collect counts for each class (only need prototypes)
2. For each word, collect a subtable of counts (no IDF or other weighting)
3. Compute the expected information in each subtable. Alternatively, compute the actual information for the word having occurred.



Results:

- Actual information is better at picking the words we intuitively think of as distinguishing the groups.
- Expected information tends to favor frequent words.
- Expected information is similar to performing a χ^2 independence test on each subtable.