Information and Interaction Among Features

36-350: Data Mining

September 12, 2008

READING: Aleks Jakulin and Ivan Bratko, "Quantifying and Visualizing Attribute Interactions", `http://arxiv.org/abs/cs.AI/0308002` (also on Blackboard)

The last handout set up a game in which we went to a random position in the document and tested for a particular word. This tends to give small expected information, since the answer is usually "no". A different question we could ask is "Is this word present anywhere in the document?" The expected information we get from this question can be computed from a table of word-presence counts. For example, this is the table for testing if the word "car" is present (in the larger collection of 200 documents):

```
      word
label  FALSE TRUE
  auto 47    53
  moto 95     5
```

This table tells you that "car" is present in 58 documents, 53 of which are in the "auto" group, and 5 of which are in the "moto" group. The other column describes the cases where "car" is not present. The expected information is 0.21 bits, compared to 0.0025 for the random-position test.

We can do this computation for all words and sort them, as shown in the figure. Interestingly, the most informative word is not "car", but "DoD", the abbreviation for a motorcycle club called "Denizens of Doom".

So far we've been computing the information content in a single word. Suppose we are allowed to ask another question about the document, after getting the answer to the first question. The best second question is not necessarily the second-best first question. For example, if you know "car" is present, it doesn't help as much to know that "cars" is also present. This effect is called **interaction**. Whereas correlation and information are properties between two variables, interaction is a property between three variables. Interaction is when measuring one variable changes the importance of another variable.

Conditional information is the information X gives about C when a third variable Y is already known:

$$I[C; X|Y = y] = H[C|Y = y] - H[C|X, Y = y] \qquad \text{(actual conditional information)}$$

$$I[C;X|Y] \quad = \quad \sum_y \Pr(Y=y)\, I[C;X|Y=y] \qquad \text{(expected conditional information)}$$

Interaction measures how Y changes the information in X:

$$\begin{aligned} I[C;X;Y=y] &= I[C;X|Y=y] - I[C;X] \qquad \text{(actual interaction)} \\ I[C;X;Y] &= I[C;X|Y] - I[C;X] \qquad \text{(expected interaction)} \end{aligned}$$

Expected interaction is symmetric in all three variables: $I[C;X;Y] = I[C;Y;X] = I[X;Y;C]$, etc. A **positive interaction** means that knowing Y makes X more informative about C. A **negative interaction** means that Y makes X less informative about C. For example, suppose C is tomorrow's weather and X and Y are weather reports from two different stations. Once you hear one weather report, the other going to be much less informative.

Relations of information and interaction can be conveniently visualized with **information graphs**.

Example: the interaction between the presence of "car" and "cars". When we look at documents without "cars", we get this table:

```
      car
label  FALSE TRUE
  auto 35    34
  moto 93     3
```

When we look at documents which contain "cars", the table is

```
      car
label  FALSE TRUE
  auto 12    19
  moto  2     2
```

Calculate:

$$
\begin{aligned}
I[label; car] &= 0.227 \\
I[label; car | cars = F] &= 0.233 \\
I[label; car; cars = F) &= 0.006 \quad \text{(positive interaction)} \\
I[label; car | cars = T] &= 0.004 \\
I[label; car; cars = T] &= -0.223 \quad \text{(negative interaction)} \\
\Pr(cars = T) &= 0.175 \\
I[label; car; cars] &= -0.034 \quad \text{(negative on average)}
\end{aligned}
$$

Note that the expected interaction can be zero, even though the actual interactions are nonzero (but of opposite sign).

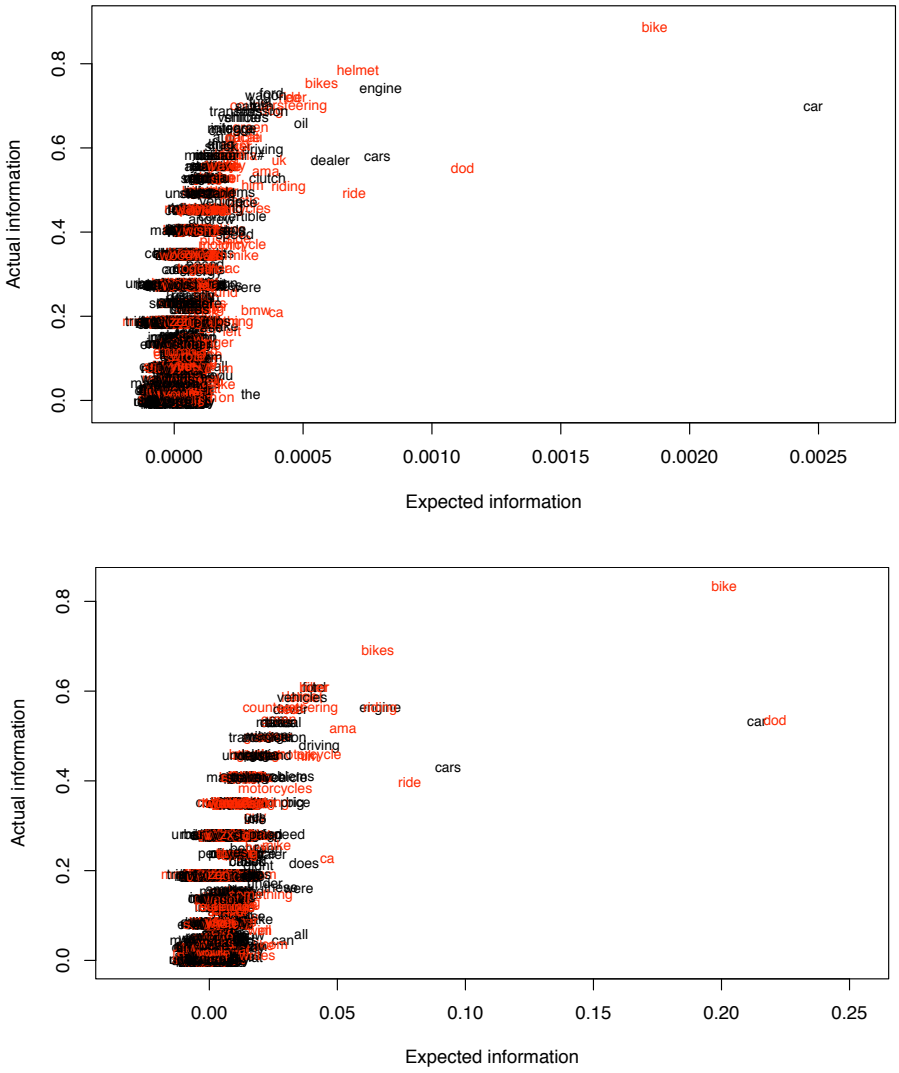

Figure 1: Important words for the large auto/moto collection. Black words indicate "auto", red words indicate "moto". *Top*: Results according to random position test we used last lecture. *Bottom*: Results according to the "is the word present at all?" test.

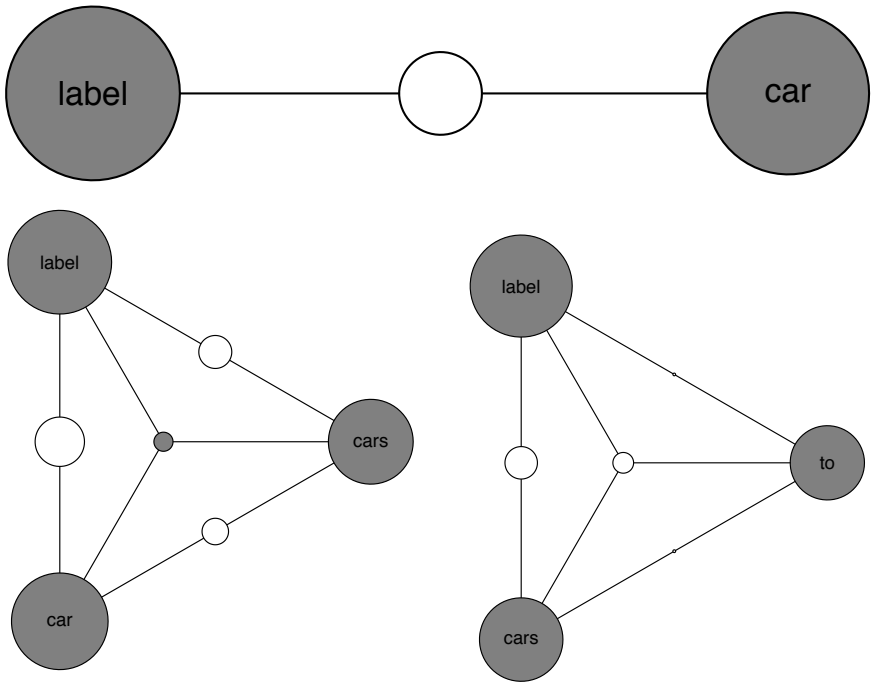

Figure 2: Information-interaction graphs. Grey stands for entropy, and white for information. The white circle attached to the line between "label" and "car" in the first graph shows that the presence of the word "car" contains information about the class label (automobile or motorcycle). In the bottom left figure, the white circles along the sides of the triangle show that the two terms "car" and "cars" convey information about each other's presence, and both contain information about the class label. However, their interaction is negative, as shown by the grey circle in the center of the triangle — this means the information they give about the class is redundant. On the bottom left, the term "to" contains almost no information about the class, or about the presence of the word "cars", but has a positive interaction with "cars" and the label.

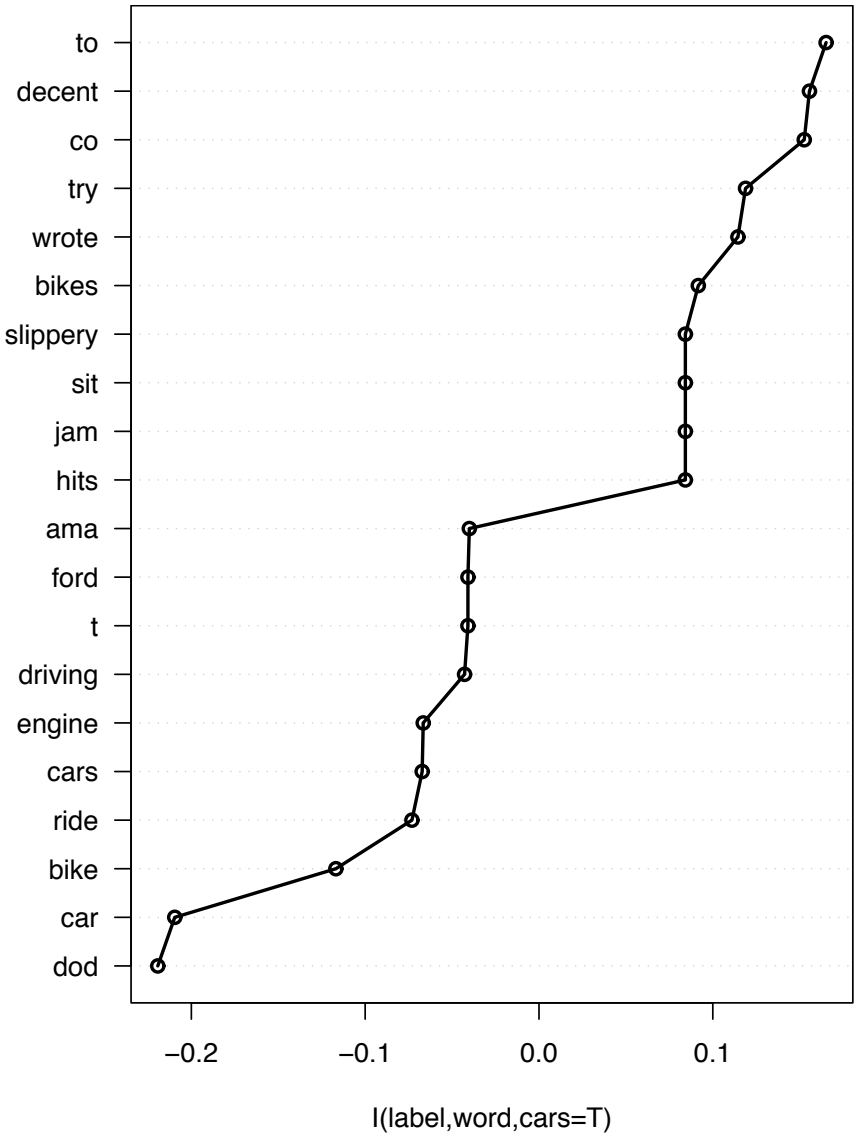

Figure 3: Interactions, positive and negative, of different words with the class labels, conditional on the document containing the word "cars".