Finding Informative Features

36-350: Data Mining

September 11, 2006

Readings: David P. Feldman, "Introduction to Information Theory", chapter 1 (on Blackboard)
Textbook, sections 10.1, 10.2, 10.6 and 10.8

As I mentioned last time, everything we have learned how to do so far — similarity searching, nearest-neighbor and prototype classification, multidimensional scaling — relies on our having a vector of **features** or **attributes** for each object in data set. (These are also called **dimensions**, because the number of dimensions in the vector space is the number of features.) The success of our procedures depends crucially on our choosing good features, but I've said very little about how to do this, aside from qualitative considerations about invariance. This week we'll look at one way of picking out presumably-useful features, using information theory.

The basic idea, remember, is that the features are the aspects of the data which show up in our representation. However, they're not what we *really* care about, which is rather something we can't directly observe or represent, like the **class** of the object (is it a post about cars or motorcycles? a picture of a flower or a tiger?). We use the observable features to make a guess (formally, an **inference**) about the unobservable thing, like the class. Good features are ones which let us make better guesses — ones which reduce our **uncertainty** about the unobserved class.

Good features are therefore **informative**, **discriminative** or **uncertainty-reducing**. This means that they need to *differ* across the different classes, at least statistically. I said before that the number of occurrences of the word "the" in an English document isn't a useful feature, because it occurs about as often in all kinds of text. This means that looking at that count leaves us exactly as uncertain about which class of document we've seen as we were before. Similarly, the word "narthex" is going to be equally *rare* whether the topic is cars or motorcycles, so it's also uninformative. On the other hand, the word "seatbelt" is going to be a lot more common in posts about cars than in ones about motorcycles, so counting its occurrences *is* going to reduce our uncertainty. The important thing is that the distribution of the feature *differ* across the classes.

Let's try and make these ideas about uncertainty, discrimination, and reduction in uncertainty precise. (This is where the information theory comes in.)

Let X be some feature of the data in our representation, and x a particular value. How uncertain are we about X? Well, one way to measure this is the **entropy** of X:

$$H[X] = -\sum_{x} \Pr(X=x) \log_2 \Pr(X=x)$$

The entropy, in bits, equals the average number of yes-or-no questions we'd have to ask to figure out the value of X. (This is also the number of bits of computer memory needed to store the value of X.) If there are n possible values for X, and they are all equally likely, then our uncertainty is maximal, and $H[X] = \log_2 n$, the maximum possible value. If X can take only one value, we have no uncertainty, and $H[X] = 0$.

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

Similarly, our uncertainty about the class C, in the absence of any other information, is just the entropy of C:

$$H[C] = -\sum_{c} \Pr(C=c) \log_2 \Pr(C=c)$$

Now suppose we observe the value of the feature X. This will, in general, change our distribution for C, since we can use Bayes's Rule:

$$\Pr(C=c|X=x) = \frac{\Pr(C=c, X=x)}{\Pr(X=x)} = \frac{\Pr(X=x|C=c)}{\Pr(X=x)}\Pr(C=c)$$

$\Pr(X=x)$ tells us the frequency of the value x is over the whole population. $\Pr(X=x|C=c)$ tells us the frequency of that value is when the class is c. If the two frequencies are not equal, we should change our estimate of the class, making it larger if that feature is more common in c, and making it smaller if that feature is rarer. Generally, our uncertainty about C is going to change, and be given by the **conditional entropy**:

$$H[C|X=x] = -\sum_{c} \Pr(C=c|X=x) \log_2 \Pr(C=c|X=x)$$

The difference in entropies, $H[C]-H[C|X=x]$, is how much our uncertainty about C has changed, conditional on seeing $X=x$. This change in uncertainty is **information**:

$$I[C;X=x] = H[C] - H[C|X=x]$$

Notice that this information can be negative. For a simple example, suppose that C is "it will rain today", and that it normally rains only one day out of seven. Then $H[C] = 0.59$ bits. If however we look at the weather forecast, and it tells us that it will rain with 50% probability, $H[C|X=x] = 1$ bit, so our uncertainty has increased by 0.41 bits.

We can also look at the **expected information** information a dimension gives us about the class:

$$I[C;X] = H[C] - H[C|X] = H[C] - \sum_{x} \Pr\left(X = x\right) H[C|X = x]$$

The expected information is never negative. In fact, it's not hard to show that the only way it can be zero is if X and C are **statistically independent** — if the distribution of X is the same for all classes c,

$$\Pr\left(X|C = c\right) = \Pr\left(X\right)$$

It's also called the **mutual information**, because it turns out that $H[C] - H[C|X] = H[X] - H[X|C]$. (You might want to try to prove this to yourself, using Bayes's rule and the definitions.)

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,

$$\begin{aligned}
H[C] &= 0.997 \\
H[C|X = \text{``car''}] &= 0 \\
H[C|X = \text{not ``car''}] &= 0.996 \\
\Pr\left(X = \text{``car''}\right) &= 0.01 \\
I[C;X] &= 0.997 - (0.01 * 0) - (0.99 * 0.996) = 0.01
\end{aligned}$$

Finding the Important Words

Here's our information-theoretic procedure for finding the important words. (This is how the words used in the first lecture were chosen.)

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.

This is just looking at whether the presence or absence on the word is informative; we could also look at whether how often it appears is informative, but we'd need bigger tables.

Results:

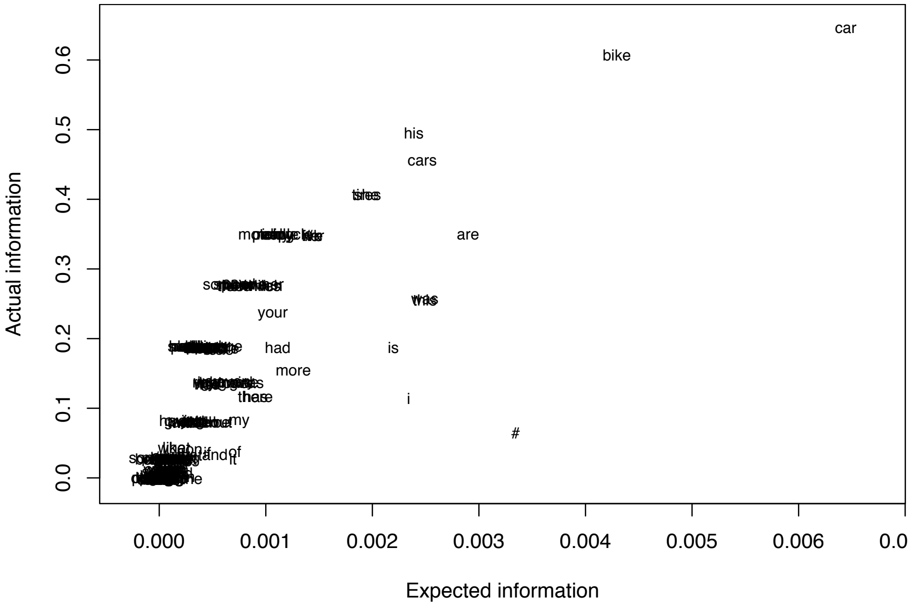

Figure 1: Reduction in uncertainty about the class (automboile/motorcycle) for the newsgroups example when a word is present ("actual information", vertical axis) versus the average reduction in uncertainty from checking for that word ("expected information", vertical axis). Note the difference in scales.

- 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. (Remember that expected information is 0 if and only if the class and the feature are statistically independent.)

All of this is just looking at one feature at a time, so it ignores the possibility that certain *combinations* of features are useful, or that some features are *redundant* given others. We will look at this sort of **interaction** in the next lecture.