CS 189 - Decision trees and some basic information theory

Noah Golmant

March 16, 2017

1 Decision trees in general

Decision trees are a useful and intuitive classification technique. We receive a set of data points $\{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ with scalar labels $\{y_i\}_{i=1}^n$. Given a single x_i, our goal is to ask a series of informative questions about the elements of x_i to make a final classification.

A non-leaf node represents some feature. We ask some question about the value of that feature for the sample, and depending on the answer we take a branch to a new node. We repeat this process. A leaf node in the tree represents the value of our prediction once we reach the end of the tree after traversing from the root (our final label guess).

How do we automatically determine what question to ask? To answer this question, we'll need some background in basic information theory.

2 Deriving the information function

Suppose we have some probability space (Ω, Σ, P) where Ω is the sample space, Σ is the σ-algebra containing all groups of outcomes (events) from Ω, and P is a probability measure $P : \Sigma \to [0, 1]$.

We wish to construct a function $I : \Sigma \to \mathbb{R}$ that tells us the amount of information contained in an event $\sigma \in \Sigma$. We'll want it to satisfy a few natural properties.

Suppose I am sending a message to some receiver who does not necessarily know the content of the message. The message contains information exactly when this receiver is uncertain about the contents of the message. So we know that if $P(\sigma) = 1$, then $I(\sigma) = 0$. And since we are dealing with uncertainties, we know that $I(\sigma) = f(P(\sigma))$ for some function $f : [0, 1] \to \mathbb{R}$, $f(P(\sigma)) = I(\sigma)\ \forall \sigma \in \Sigma$.

We also see that f and I must be nonnegative. Negative information doesn't really make much sense in this content.

Suppose we now send two independent messages simultaneously. Then the information content of receiving both messages at once should be the sum of

their information contents. That is, for $A, B \in \Sigma$, A, B independent, $I(A \cap B) = I(A) + I(B)$. So it is also additive.

But we also know that $I(A \cap B) = f(P(A)P(B))$ since A, B are independent. So $I(A) + I(B) = f(P(A)) + f(P(B)) = f(P(A)P(B))$. The f that satisfies this property is $f(x) = K \log x$ for some constant K. Since the domain of f is $[0, 1]$, and $\log x \leq 0 \; \forall x \in (0, 1]$, we have that $K < 0$ to ensure that our measure is nonnegative. We choose $K = -1$ since the scale of our function doesn't really matter.

Then

$$I(\sigma) = f(P(\sigma)) = -\log P(\sigma)$$

This is known as the self-information function. It is often considered as the "surprise" of observing an event σ. Note for example that as $P(\sigma) \to 0$, $I(\sigma) \to \infty$.

3 Entropy

Let's say we have a random variable X defined on a distribution D in our original probability space. X can take on a variety of values with potentially differing probabilities. We want to know how much information we can "expect" to receive by observing X. Equivalently, we want to know how surprised we will be by the observation. This is just the expected information content, or entropy, which is denoted by

$$H(X) = \mathbb{E}_{x \sim D} - \log P(x)$$

.

4 Information gain

Now let's consider two random variables, X and Y, each with their own distributions. Our "state of the world" is X, e.g. it is our current belief about some value represented by X. How much knowledge do we gain by observing Y? Exactly the change in information that occurs when we take Y as a given. When we take Y as given, we get a new distribution $X|Y$ that represents our beliefs about X given that we know Y. So the information gain is:

$$IG(X, Y) = H(X) - H(X|Y)$$

5 Back to decision trees

How do we determine what questions to ask? Well consider asking about whether or not a sample x_i is in a class y_i. We can consider some feature

of x_i as the attribute that we will consider as a new given, and y_i is the unknown thing we would like information about. So we should ask the question that will give us the largest information gain, usually approximating the probabilities of these random variables using counts of feature values and class labels from the training data.