Machine Learning 15:681

Assignment 7, Fall 1995

Due December 5

1. Consider the Minimum Description Length principle applied to the hypothesis space H consisting of conjunctions of up to n boolean attributes (e.g., $Sunny \wedge Warm$). Assume each hypothesis is encoded simply by listing the attributes present in the hypothesis, where the number of bits needed to encode any one of the n boolean attributes is $\log_2 n$. Suppose the encoding of an example given the hypothesis uses zero bits if the example is consistent with the hypothesis and uses $1 + \log_2 m$ bits otherwise (1 bit to indicate the correct classification, and $\log_2 m$ to indicate which of the m examples was misclassified).

   - Write down the expression for the quantity to be minimized according to the Minimum Description Length principle.
   - Is it possible to construct a set of training data such that a consistent hypothesis exists, but MDL chooses a less consistent hypothesis? If so, give such a training set. If not, explain why not.
   - (Optional extra credit) Can you specify probability distributions for $P(h)$ and $P(D|h)$ such that the above MDL algorithm outputs MAP hypotheses? Explain any difficulties, or assumptions you must make.

2. Write down the network that would be generated by the first step of the KBANN algorithm, given the following domain theory:

   pass_final $\leftarrow$ know_material, wide_awake
   know_material $\leftarrow$ studied
   wide_awake $\leftarrow$ slept_last_night

   Assume instances are described by the four boolean attributes $slept\_last\_night$, $studied$, $ate\_breakfast$, $passed\_midterm$.

3. Give an example of a search space in which you might expect the Genetic Algorithm crossover operator to be:

   - useful
   - detrimental

   In each case, explain why.

1 Minimum Description Length Principle

Recall from Chapter 4 the discussion of Occam's razor, a popular inductive bias that can be summarized as "choose the shortest explanation for the observed data." In that chapter we discussed several arguments in the long-standing debate regarding Occam's razor. Here we consider a Bayesian perspective on this issue, and a closely related principle called the Minimum Description Length (MDL) principle.

The Minimum Description Length principle is motivated by interpreting the definition of h_{MAP} in the light of basic concepts from information theory. Consider again the now familiar definition of h_{MAP}.

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

which can be equivalently expressed in terms of maximizing the $\log_2$

$$h_{MAP} = \arg\max_{h \in H} \ \log_2 P(D|h) + \log_2 P(h)$$

or alternatively, minimizing the negative of this quantity

$$h_{MAP} = \arg\min_{h \in H} \ -\log_2 P(D|h) - \log_2 P(h) \tag{1}$$

Somewhat surprisingly, Equation 1 can be interpreted as a statement that short hypotheses are prefered, assuming a particular representation scheme for encoding hypotheses and data. To explain this, let us introduce a basic result from information theory: Consider the problem of designing a code for transmitting messages drawn at random, and where the probability of encountering message i is p_i. We are interested here in the most compact code; that is, the code that minimizes the expected number of bits we must transmit in order to encode a message drawn at random. Clearly, to minimize the expected code length we should assign shorter codes to messages that are more probable. Shannon (1949?) showed that the optimal code (i.e., the code that minimizes the expected message length) assigns $-\log_2 p_i$ bits[1] to encode message i. We will refer to the number of bits required to encode message i using code C as the *description length of message i with respect to C*, which we denote by $L_C(i)$.

Let us interpret Equation 1 in the light of the above result from coding theory.

- $-\log_2 P(h)$ is the description length of h under the optimal encoding for the hypothesis space H. In other words, this is the size of the representation of hypothesis h using this optimal representation. In our notation, $L_{C_H}(h) = -\log_2 P(h)$, where C_H is the optimal code for hypothesis space H.

- $-\log_2 P(D|h)$ is the description length of the training data D given hypothesis h, under its optimal encoding. In our notation, $L_{C_{D|h}}(D|h) = -\log_2 P(D|h)$, where $C_{D|h}$ is the optimal code for describing data D assuming that both the sender and receiver know the hypothesis h.

- Therefore we can rewrite Equation 1 to show that h_{MAP} is the hypothesis h that minimizes the sum given by the description length of the hypothesis plus the description length of the data given the hypothesis.

$$h_{MAP} = \arg\min_h L_{C_H}(h) + L_{C_{D|h}}(D|h)$$

where C_H and $C_{D|h}$ are the optimal encodings for H and for D given h, respectively.

[1]Notice the expected length for transmitting one message is therefore $\sum_i -p_i \log_2 p_i$, the formula for the *entropy* (see Chapter 4) of the distribution of messages!

The Minimum Description Length (MDL) principle recommends choosing the hypothesis that minimizes the sum of these two description lengths. Of course to apply this principle in practice one must choose specific encodings or representations appropriate for the given learning task. Assuming we use the codes C_1 and C_2 to represent the hypothesis and the data given the hypothesis, we can state the MDL principle as the principle "choose hypothesis h_{MDL}" where

Minimum Description Length principle: Choose h_{MDL} where

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D|h) \tag{2}$$

The above analysis shows that if we choose C_1 to be the optimal encoding of hypotheses C_H, and if we choose C_2 to be the optimal encoding $C_{D|h}$, then $h_{MDL} = h_{MAP}$.

Intuitively, we can think of the MDL principle as recommending the shortest method for re-encoding the training data, where we count both the size of the hypothesis, and any additional cost of encoding the data given this hypothesis.

Let us consider an example. Suppose we wish to apply the MDL principle to the problem of learning decision trees from some training data. What should we choose for the representations C_1 and C_2 of hypotheses and data? For C_1 we might naturally choose some obvious encoding of decision trees, in which the description length grows with the number of nodes in the tree and with the number of edges. How shall we choose the encoding C_2 of the data given a particular decision tree hypothesis? To keep things simple, suppose that the sequence of instances $\langle x_1 \ldots x_m \rangle$ is already known to both the transmitter and receiver, so that we need only transmit the classifications $\langle f(x_1) \ldots f(x_m) \rangle$, where $f(x)$ is the classification of x according to the target function f. The cost of transmitting the instances themselves is independent of the correct hypothesis, so it does not affect the selection of h_{MDL} in any case. Now if the training classifications $\langle f(x_1) \ldots f(x_m) \rangle$ are identical to the predictions of the hypothesis, then there is no need to transmit any information about these examples (the receiver can compute these values because it knows the hypothesis). The description length of the classifications given the hypothesis in this case is therefore zero. In the case where some examples are misclassified by h, then for each misclassification we need to transmit a message that identifies which example is misclassified (which can be done using at most $\log_2 m$ bits) as well as its correct classification (which can be done using at most $\log_2 k$ bits, where k is the number of possible classifications). The hypothesis h_{MDL} under the encodings C_1 and C_2 is just the h from H that minimizes the sum of these description lengths.

Thus the MDL principle provides a way of trading off hypothesis complexity for the number of errors committed by the hypothesis. It might select a shorter hypothesis that makes a few errors, over a longer hypothesis that perfectly classifies the training data. Viewed in this light, it provides one method for dealing with the issue of *overfitting* the data (see the discussion of decision tree learning or neural network learning).

Quinlan and Rivest (1989) describe experiments applying the MDL principle to choose the best size for a decision tree. They report that the MDL-based method produced learned trees whose accuracy was comparable to that of the standard tree pruning methods discussed in Chapter 4. Mehta et al. (1995) describe an alternative MDL-based approach to decision tree pruning, and also provide experimental results showing results comparable to standard tree pruning methods.

What shall we conclude from this analysis of the Minimum Description Length principle? Does this prove once and for all that short hypotheses are best? No. What we have shown is only that *if* a representation of hypotheses is chosen so that the size of hypothesis h is $-\log_2 P(h)$, *then* the MDL principle produces MAP hypotheses. However, to show that we have such a representation we must know all the prior probabilities $P(h)$. There is no reason to believe that the MDL hypothesis relative to *arbitrary* encodings C_1 and C_2 should be prefered. As a practical matter it might sometimes be easier for a human designer to specify a representation that captures knowledge about the relative probabilities of hypotheses than it is to fully specify the probability of each hypothesis. Descriptions in the literature regarding the application of MDL to practical learning problems often include arguments justifying the encodings chosen for C_1 and C_2.