10-701/15-781, Machine Learning: Homework 1

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on January 12, 2010

1 Decision Trees and Information Theory [35+5 pt, Ni Lao]

Suppose a discrete variable X has n categories $1...n$. It's entropy is defined as

$$H(X) = -\sum_i P(X=i)\log P(X=i).$$

Suppose another variable Y has distribution $P(Y=j)$, and joint distribution $P(X=i, Y=j)$. Then their mutual information is

$$I(X;Y) = \sum_{i,j} P(X=i,Y=j)\log\frac{P(X=i,Y=j)}{P(X=i)P(Y=j)} = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

A *K-nary code* C for X is a mapping from X to a string (code word) $C(X)$ of which each character can have K values. A *prefix code* is a code for which no code word is a prefix of other code word. The average code length is defined as $L(C) = \sum_i P(X=i)l_i$ where l_i is the length of $C(i)$. It can be proved that $H(X)$ is the minimum average code length needed to encode X, and the minimum is reached if and only if $P(X=i) = K^{-l_i}$.

1. **The 9 ball problem [10 pt]** There are 9 metal balls. One of then is heavier than the others. Please design a strategy given a scale to find out which one is heavier with the least number of expected trials. Then show that it is optimal from information point of view. **Hint:** connect the average code length with the expected number of tests.

2. **Information gain and optimal encoding [10 pt]** show that, when using K-outcome tests, if each test Y_t has optimal mutual information $I(Y_t; X|Y_1...Y_{t-1}) = 1$ (in base K numeral system) conditioned on all previous tests, then the expected number of tests is optimal. Assume that the mapping from X to Y_t are deterministic ($H(Y_t|X) = 0$). **Hint:** $H(Y) \geq I(X;Y)$.

3. **The ID3 algorithm [10 pt]** If we run the ID3 algorithm on the 9 metal ball problem, will it generate the optimal decision tree? **Hint:** treat the outcome of comparing any two sets of balls as a feature, and compare their information gains.

4. **The number guessing problem [5 pt]** Alice has a favorite four digit number (e.g. 0123), and she wants Bob to guess it. At each iteration, Bob say a four digit number (e.g. 3210), and then Alice tell him how many digit is correct (four in this example), and how many of them have correct position (zero in this example). From information point of view, what is the minimum expected number of guesses Bob has to make in order to identify the number?

5. **[Extra 5 pt]** Try the number guessing game a few times with your friend. Did you achieve your estimated lower bound? Why?