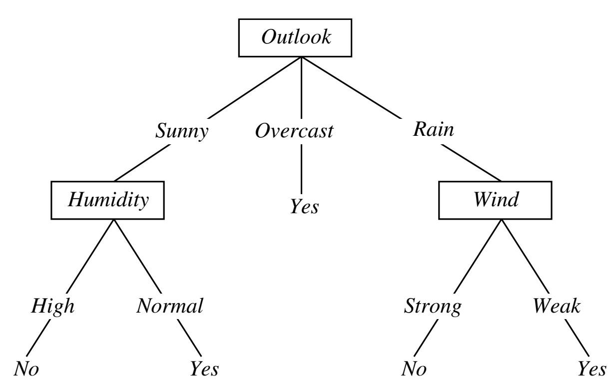
10-701/15-781 Machine Learning: Assignment 1

- The assignment is due **September 27, 2005** at the beginning of class.
- Write your name in the top right-hand corner of each page submitted. No paperclips, folders, etc.
- If you have any questions, email `questions-10701@autonlab.org`.
- This assignment consists of five questions totalling 100 points.
- Each student must hand in an writeup. See the web page for the collaboration policy.

Q1 Decision Trees and Overfitting [20 pts]

Consider the following training data and the following decision tree learned from this data using the ID3 algorithm (without any postpruning).

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



1. Show that the choice of the *Wind* attribute at the second level of the tree is correct, by showing that its information gain is superior to the alternative choices.

2. Add one new example to the above data set, so that the learned tree will contain additional nodes.

3. Is it possible to add new examples to the above training set, which are *consistent with the above tree*, to produce a larger training set such that ID3 will now learn a tree whose root node is not *Outlook*?. (We say an example is *consistent* with the above tree if the tree classifies the example correctly). Justify your answer by explaining informally why this is impossible, or explaining the new data you would add.

4. Any decision tree can be re-expressed as a set of rules, with one rule for each leaf (Mitchell pp.71-72). The preconditions of the rule correspond to the sequence of attribute tests along the path from the tree root to the leaf. For example, the leftmost leaf in the above tree corresponds to the rule

   ```
   IF (Outlook = sunny  AND Humidity=High)    THEN PlayTennis=No
   ```

   Write the rules for the remaining leaves. Note this set of rules produces classifications that are identical to the above tree, over the training data and over any other possible instance.

5. It is possible to translate any tree into a set of rules that represents an equivalent classifier. Is it possible to translate any set of rules into an equivalent tree? Explain how or give a counterexample.

6. In class we discussed post-pruning the tree using reduced-error pruning over a validation data set, to avoid overfitting. Consider the alternative of converting the tree to its equivalent rules, then pruning the rules. In particular, let's consider the pruning each rule independently, by iterating the following steps:

   (a) Determine which rule precondition would, if removed, produce the greatest accuracy improvement over the validation set. (the "accuracy" of the rule is the number of correct predictions it makes, divided by the number of predictions it makes).

   (b) If removing this precondition improves validation set accuracy, then remove it and iterate, else stop the pruning of this rule

   Consider the two pruning strategies (pruning the tree, versus pruning the equivalent rule set). Is it the case that these two pruning strategies produce equivalent pruned classifiers (i.e., do the pruned tree and the pruned rules make equivalent classifications over all possible instances?). Explain why or when not.

Q2 Entropy, Conditional Entropy, and Information Gain [25 pts]

Properties of Entropy

Given two discrete random variables X and Y which take on values $\{v_1, \ldots, v_n\}$ the entropy is

$$H(X) = -\sum_{i=1}^{n} p(X = v_i) \log p(X = v_i)$$

The specific conditional entropy is

$$H(Y|X = v_k) = -\sum_{i=1}^{n} p(Y = v_i|X = v_k) \log p(Y = v_i|X = v_k)$$

and can be viewed as the expected number of bits needed to encode a sample from Y given that we know $X = v_k$. The average conditional entropy is

$$H(Y|X) = \sum_{k=1}^{n} p(X = v_k) H(Y|X = v_k)$$

and can be viewed as the expected number of bits needed to encode a sample from Y given X. The joint entropy

$$H(X,Y) = -\sum_{i}\sum_{j} p(x,y)\log p(x,y)$$

and can be viewed as the expected number of bits needed to jointly encode a sample from (X,Y). An important property of joint entropy is the chain rule

$$H(X,Y) = H(X) + H(X|Y) = H(Y) + H(Y|X)$$

The information gain is

$$\begin{aligned} IG(Y|X) &= H(Y) - H(Y|X) \\ IG(X|Y) &= H(X) - H(X|Y) \end{aligned}$$

and can be viewed as the expected number of bits saved when encoding a sample from Y given that both sender and receiver already have X. An important property to remember is that $IG(X|Y) = IG(Y|X)$.

Q2.1 Entropy Inequalities [10 pts]

1. Given a random variable X and any deterministic function $g(X)$ we claim that $H(g(X)) \leq H(X)$ with the following justification:

$$\begin{aligned} H(X, g(X)) &= H(X) + H(g(X)|X) \\ &= H(X) \qquad (1) \\ H(X, g(X)) &= H(g(X)) + H(X|g(X)) \\ &\geq H(g(X)) \qquad (2) \end{aligned}$$

   Thus $H(X) = H(X, g(X)) \geq H(g(X))$.

   (a) Is the claim correct ?

   (b) If the claim is not correct, explain the mistake in steps 1 and 2. If the claim is correct, explain why. (Answer should be no more than two sentences long. You do not have to prove anything given as a property of entropy).

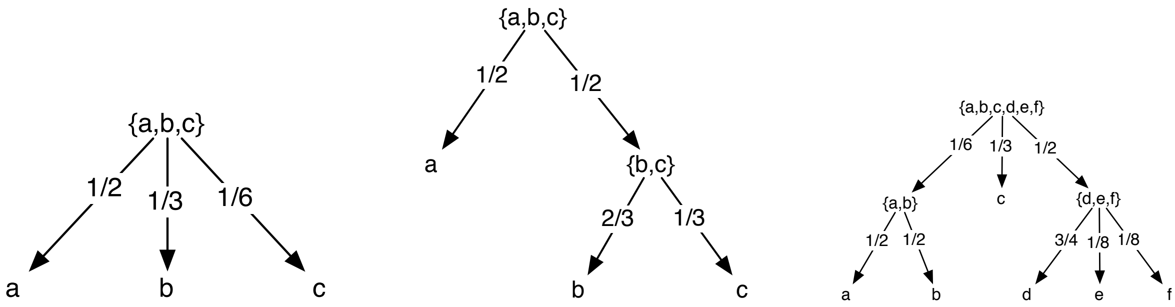
2. For which of the following functions does $H(g(X)) = H(X)$, where X is stochastic, discrete, and strictly positive ?

   (a) $g(X) = aX + b$ assuming $a, b > 0$

   (b) $g(X) = \log(X)$

   (c) $g(X) = 0$

   (d) $g(X) = 2.23$



(a) Encoding 1 (b) Encoding 2 (c) Encoding 3

Figure 1: Encoding 1 and 2 represent the same underlying distribution, encoding 3 does not

Q2.2 Derivation of Entropy [15 pts]

The equation for entropy

$$H_n(p_1, \dots, p_n) = -\sum_i p_i \log p_i$$

Is a rather non-obvious definition for entropy. In this question we will propose some desirable properties for entropy as a measure of uncertainty, and then show that entropy $m$ust be within a multiplicative factor of $-\sum_i \log p_i \log p_i$.

A1. H_n should be continuous in p_i and symmetric in its arguments.

A2. If $p_i = 1/n$ then H_n should be a monotonically increasing function of n. If all events are equally likely, then having more events means being more uncertain.

A3. If a choice among N events is broken down into successive choices, then entropy should be the weighted sum of the entropy at each stage. For example, In encoding 1, we draw from (a, b, c) with probabilities $(1/2, 1/3, 1/6)$ respectively. In encoding 2, we first draw from $\{a, \{b, c\}\}$. If the set $\{b, c\}$ is chosen we then select from b or c with probabilities $2/3$ and $1/3$.

Let $\psi_n(p_i, \dots, p_n)$ be any function which respects axioms A1-A3. Let $A(n) = \psi_n(1/n, \dots, 1/n)$.

1. The entropy of the distribution represented by encoding 1 is

$$H(1/2, 1/3, 1/6) = -\left[\tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{3}\log\tfrac{1}{3} + \tfrac{1}{6}\log\tfrac{1}{6}\right] = 1.46$$

   The entropy of distribution represented by encoding 2 is

$$H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2}H(\tfrac{2}{3}, \tfrac{1}{3}) = 1.46$$

   What is the entropy of the encoding 3, as shown in figure 1(c) ? (Use axiom A3, like we did for encoding 2) ?

2. Prove that $A(s^m) = mA(s)$ [Hint: We are dealing with s^m equiprobable events, use A3]

We choose another set of t^n equiprobable events where n is chosen so that $s^m \leq t^n \leq s^{m+1}$. By taking logarithms and rearranging

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \implies \left|\frac{m}{n} - \frac{\log t}{\log s}\right| \leq \frac{1}{n} = \epsilon \tag{3}$$

3. Explain why $A(s^m) \leq A(t^n) \leq A(s^{m+1})$. (Answer should be one or two sentences).

By using the results of questions 2 and 3 it is easy to show that

$$A(s^m) \leq A(t^n) \leq A(s^{m+1}) \implies mA(s) \leq nA(t) \leq (m+1)A(s) \implies \left|\frac{m}{n} - \frac{A(t)}{A(s)}\right| < \frac{1}{n} = \epsilon \tag{4}$$

By combining equations 3 and 4

$$\left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| < 2\epsilon \implies A(t) = K \log t, \quad (K > 0 \text{ because of A2}) \tag{5}$$

Now consider a set of N equally likely events. For any partition of these events, $S_1, S_2, \ldots, S_k$ let $p_i = |S_i|/N$. We propose the following two level encoding. First, choose a subset according to the probabilities $(p_1, \ldots, p_k)$. Then sample uniformly from the chosen set. By equation 5 we know that $A(N) = K \log N$. By axiom A3 and the two stage encoding we also know that $A(N) = \psi_k(p_1, \ldots, p_k) + K \sum_i p_i \log n_i$. Therefore

$$\begin{aligned}
K \log N &= \psi_k(p_1, \ldots, p_k) + K \sum_i p_i \log |S_i| \\
\psi_k(p_1, \ldots, p_k) &= K \left[\log N - \sum_i p_i \log |S_i|\right] \\
&= K \left[\log N \sum_i p_i - \sum_i p_i \log |S_i|\right] \\
&= -K \sum_i p_i \log \frac{|S_i|}{N} \\
&= -K \sum_i p_i \log p_i
\end{aligned}$$

Q3 Probability [25 pts]

1. Using only the axioms of probability show that $p(A|B, A) = 1$.
2. Using only the axioms of probability show that $p(A, B|C) = p(A|C)p(B|C)$ only holds when A or B is marginally independent of C, *i.e.* $P(A|B, C) = P(A|C)$ or $P(B|A, C) = P(B|C)$.
3. You have two bags. The first bag contains 11 white marbles and 4 black marbles. The second bag contains 8 white marbles and 5 black marbles. One bag is chosen uniformly at random, and then a marble is selected from the bag. The marble selected is white. What is the probability that it came from the first bag ?
4. We uniformly choose X between 0 and 1. Then we choose Y between 0 and X. Thus the expression for $p(x)$ is

$$p(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Give the expression for $p(y|x)$
   (b) Give the expression for $p(y)$
   (c) Give the expression for $p(x|y)$
   (d) If the values chosen for X and Y are used to form a rectangle, what is the expected area of that rectangle ?

Q4 Gaussians [20 pts]

Given the following model for IQ and test scores S

$$IQ \sim N(100, 15^2)$$
$$S|IQ \sim N(IQ, 10^2)$$
$$s_1 = 130$$

we derived the posterior density $IQ|S = s_1$ in class. The snobbish parent, upon discovering their daughter is seen as slightly less intelligent by Bayesians, decides to make her take the test again.

1. Let the daughter's score on the first and second tests be S_1 and S_2 respectively. What is the posterior density of her score on the second test, given the evidence of her score on the first test. Formally, calculate the mean and variance of the distribution $p(S_2|S_1 = 130)$.
2. The snobbish parent bets his neighbor \$100 that his daughter will improve her score on the second test. How much money can he expect to win (or lose) ?

Q5 Maximum Likelihood and Maximum a Posteriori [10 pts]

A food inspection officer at a chocolate factory is asked to measure the number of insect parts in chocolate bars. Fifteen chocolate bars are independently chosen and analyzed. The number of insect parts measured in each bar are $\{x_1 \dots x_{15}\} = \{2, 1, 0, 1, 0, 0, 1, 0, 2, 0, 1, 1, 0, 2, 1\}$. The Poisson distribution is used to model the number of insect parts in a chocolate bar, X.

$$p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

For this question, do not refer to textbooks other than Mitchell. Everything you need to know about the Poisson and Gamma distributions to solve this question has been provided.

1. Write down the log-likelihood of the data as a function of λ.
2. The maximum likelihood estimator of λ is

   $$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

   Compute $\hat{\lambda}_{MLE}$ given the data.
3. Assume a Gamma$(r = 2, \alpha = 8)$ prior distribution on λ

   $$p(\lambda) = \frac{\alpha^r}{\Gamma(r)}\lambda^{r-1}e^{-\alpha\lambda} \quad \lambda > 0$$

   If a random variable is Gamma(r, α) distributed then its expected value is r/α. If $r > 1$ the mode is $(r-1)/\alpha$.

   (a) How many insect parts do we expect to see in a chocolate bar before collecting the data ?

   (b) What is the posterior PDF of λ ? Hint:

   $$\int_0^\infty \lambda^{\sum x_i + r - 1} e^{-\lambda(n+\alpha)}\, d\lambda$$

   Looks a lot like the PDF of a Gamma distribution.

   (c) Compute the maximum a posteriori estimate of λ given the data.

   (d) If instead we assumed prior distribution $\lambda \sim$ Gamma$(r = 4, \alpha = 16)$, what would the MAP estimate of λ be ?