

Statistical Methods for Artificial Intelligence, Autumn 2006
Midterm Exam

Problem 1. (Entropy of functions of a random variable). Let X be a discrete random variable. Let $f(X)$ be a function of the outcome X . Through a few steps, this problem will ask you to prove that the entropy of X is always greater than or equal to a function of X , i.e.

$$H(X) \geq H(f(X))$$

a) First, let us prove a useful property of the entropy. Recall that the entropy of two joint random variables X and Y , jointly distributed as $\Pr(X, Y)$, is

$$H(X, Y) = \sum_{x,y} \Pr(x, y) \log \frac{1}{\Pr(x, y)}$$

A useful definition is that the *conditional entropy* with respect to a distribution $\Pr(X, Y)$ is

$$H(Y|X) = \sum_{x,y} \Pr(x, y) \log \frac{1}{\Pr(y|x)}$$

Prove the “chain rule” of conditional probabilities. Namely, that

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

b) Justify the following steps:

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

c) Also, justify:

$$H(X, f(X)) = H(f(X)) + H(X|f(X)) \geq H(f(X))$$

Thus, $H(X) \geq H(f(X))$

Problem 2. Consider a run of an HMM with observed values a_1, \dots, a_T . The forward and backward probabilities for an HMM which can be computed using the following equations (respectively).

$$\begin{aligned}
 P(X_{i+1} = x, O_1 = a_1, \dots, O_i = a_i) &= \sum_y P(X_i = y, O_1 = a_1, \dots, O_{i-1} = a_{i-1}) P(O_i = a_i | X_i = y) P(X_{i+1} = x | X_i = y) \\
 P(O_i = a_i, \dots, O_T = a_T | X_i = x) &= \sum_y P(O_i = a_i | X_i = x) \sum_y P(X_{i+1} = y | X_i = x) P(O_{i+1} = a_{i+1}, \dots, O_T = a_T | X_{i+1} = y)
 \end{aligned}$$

Give an expression for $P(X_i = x, X_{i+1} = y | O_1 = a_1, \dots, O_T = a_T)$ in terms of inside and outside probabilities (probabilities of the form of the expressions on the left hand side of the above equations).

Problem 3. Consider a graph each node is an n bit bit string (so that there are at most 2^n nodes) and having the property that if two nodes are connected by an edge then the two nodes differ in only one bit. Let 0 be the all zero bit vector and let g be the bit string of $n/2$ zeros followed by $n/2$ ones (assume n is even). For bit strings x and y let $d(x, y)$ be the number of places (bits) at which x and y are different. This is called the Hamming distance from x to y . For any node x we have that $d(x, g)$ is a monotone heuristic that is admissible for goal g . Let m be a value such that every node y on the shortest path from 0 to g satisfies $d(y, g) \leq m$ (we can take m to be the maximum of $d(y, g)$ over the nodes y in the shortest path). Give an upper bound as a function of m on the number of nodes removed from the queue in going from 0 to g when A^* is run with this heuristic. Tighter bounds will receive higher grades as long they are correct. Explain why your bound is correct.

Problem 4. Consider the Bayesian network $X \rightarrow Z \leftarrow Y \rightarrow W$. Show that X and W are independent. Give an instance of this network (particular conditional probability tables) where X and W are not independent given Z .

Problem 5. A Dynamical Bayesian network (DBN) is a special case of an HMM where each hidden state variable is replaced by a set of hidden variables

at that point in time. We let n be the number of hidden variables at one point in time. In a DBN the state transition probability is given by a Bayesian network (see the figure on the white board). We can “unroll” a DBN for T time steps into one large Bayesian network with nT hidden variables and T observed variables. We also assume that each observation variable for time i has all n state variables for time i as parents. This unrolled DBN defines a hypergraph whose nodes are the state and observation variables and where there is one hyperedge for each variable consisting of that variable and its parents.

a) Give an upper bound on the tree width of the unrolled DBN as a function of n and T .

b) Assuming that the Bayesian network defining $P(X_{i_1} = \sigma | X_i = \gamma)$ has width w , give the width of the unrolled network as a function of n , w and T .

Problem 6. For the optimal β in a linear regression problem, the normal equation states that

$$\mathbb{E} [x_i(y - \hat{y})] = 0$$

where $\hat{y} = \beta^T x = \sum_i \beta_i x_i$. Using this, show that:

$$\mathbb{E} [\hat{y}(y - \hat{y})] = 0$$

Note that the above equation states that the optimal predictions \hat{y} are uncorrelated with the prediction errors.