CS 101.2 - Active Learning
Problem Set 3

Handed out:	19	Feb 2009
Due:	5	Mar 2009

---

1 Value of Information

Consider the medical diagnosis problem discussed in lecture. Let Bernoulli random variable Y equal 1 if our patient is sick, or 0 if our patient is healthy. We wish to choose some course of action for the patient, which will be indicated by variable A. We can encode the patient's welfare using a utility function $U(a, y)$ that maps the possible actions and patient status to a real number (positive numbers indicate a good outcome, and negative numbers indicate a bad outcome).

We can not observe Y directly, so we will make our decision based on medical tests whose results are given by random variables $X_1, \ldots, X_n$ indexed by the set $V = \{1, \ldots, n\}$. We must determine which set of tests to perform in order to obtain a good outcome for the patient. Let $\mathbf{X}_B$ be the random vector of the variables indexed by a subset of indices $B \subseteq V$, which correspond to the outcomes of the tests specified by B, and let $P(\mathbf{X}_B)$ denote the joint distribution of that random vector.

One way to evaluate a set of tests is the Value of Information criterion:

$$VOI(B) = \mathbb{E}_{\mathbf{x}_B} \max_a \Big\{ \mathbb{E}_y \big[U(a, y) \mid \mathbf{x}_B \big] \Big\} = \sum_{\mathbf{x}_B} P(\mathbf{x}_B) \max_a \sum_y P(y \mid \mathbf{x}_B) U(a, y), \tag{1}$$

which captures the expected utility that results from choosing an action based on the outcome of test $\mathbf{X}_B$. In this problem, we will explore the properties of this criterion.

1. The value of observing $\mathbf{X}_B = \mathbf{x}_B$ is defined as

$$\text{Value}(\mathbf{X}_B = \mathbf{x}_B) = \max_a \Big\{ \mathbb{E}_y [U(a, y) | \mathbf{x}_B] \Big\}, \tag{2}$$

   and can be written as

$$\max_a \Big\{ p' U(a, 1) + (1 - p') U(a, 0) \Big\} \tag{3}$$

   where $p' = P(Y = 1 | \mathbf{X}_B = \mathbf{x}_B)$. Prove that $\text{Value}(\mathbf{X}_B = \mathbf{x}_B)$ is a convex function of p'.

2. Recall Jensen's inequality

$$\mathbb{E} F(x) \geq F(\mathbb{E} x), \tag{4}$$

which holds for any convex function $F(x)$. Use this along with the result from the above subproblem to prove that the value of information set function is monotonic. That is, if $A \subseteq B$ then

$$VOI(A) \leq VOI(B). \tag{5}$$

3. As a simple example, let's assume that there are only two possible experiments: X_1 and X_2. We will use the joint distribution $P(X_1, X_2, Y)$ specified on page 16 of the Bayesian Experimental Design lecture notes (with $\varepsilon = 0$): For either experiment, $X_i = Y$ with probability 0.5. Otherwise X_i is uniformly chosen from $\{0, 1\}$. There are three possible actions, and you will use the utility function given on the same slide. Compute $VOI(\{1\})$, $VOI(\{2\})$, and $VOI(\{1, 2\})$.

4. Show that the above example implies that Value of Information is in general not a submodular set function.

2 Submodularity of Entropy

Let $X_1, \ldots, X_n$ be random variables that take values in some finite set $\mathcal{X}$ and are indexed by the set $V = \{1, \ldots, n\}$. Let $\mathbf{X}_A$ be the random vector of the variables indexed by a subset of indices $A \subseteq V$. That is, if $A = \{i_1, i_2, \ldots, i_m\}$ then $\mathbf{X}_A = [X_{i_1}, X_{i_2}, \ldots, X_{i_m}]$. The entropy of the random vector is defined as

$$H(\mathbf{X}_A) = -\sum_{\mathbf{x} \in \mathcal{X}^m} P(\mathbf{X}_A = \mathbf{x}) \log P(\mathbf{X}_A = \mathbf{x}), \tag{6}$$

where

$$P(\mathbf{X}_A = \mathbf{x}) = P(X_{i_1} = x_1, \ldots, X_{i_m} = x_m) \tag{7}$$

is the joint distribution of the random vector $\mathbf{X}_A$.

The conditional entropy of random vector $\mathbf{X}_A$ given $\mathbf{X}_B$ (where $A = \{i_1, \ldots, i_m\}$ and $B = \{j_1, \ldots, j_l\}$ are disjoint subsets of indices) is defined as

$$H(\mathbf{X}_A | \mathbf{X}_B) = -\sum_{\mathbf{x} \in \mathcal{X}^m} \sum_{\mathbf{y} \in \mathcal{X}^l} P(\mathbf{X}_A = \mathbf{x}, \mathbf{X}_B = \mathbf{y}) \log P(\mathbf{X}_A = \mathbf{x} | \mathbf{X}_B = \mathbf{y}), \tag{8}$$

where

$$P(\mathbf{X}_A = \mathbf{x} | \mathbf{X}_B = \mathbf{y}) = P(X_{i_1} = x_1, \ldots, X_{i_m} = x_m | X_{j_1} = y_1, \ldots, X_{j_l} = y_l) \tag{9}$$

is the conditional distribution of $\mathbf{X}_A$ given $\mathbf{X}_B$.

1. Prove the chain rule for entropy:

$$H(\mathbf{X}_{A \cup B}) = H(\mathbf{X}_A) + H(\mathbf{X}_B | \mathbf{X}_A), \tag{10}$$

assuming that A and B are disjoint subsets of indices.

2. Prove that "information never hurts":

$$H(\mathbf{X}_A|\mathbf{X}_B) \leq H(\mathbf{X}_A), \tag{11}$$

which means that knowing the values of additional random variables $\mathbf{X}_B$ never increases the uncertainty about $\mathbf{X}_A$. You can prove this in two steps. First show that

$$H(\mathbf{X}_A|\mathbf{X}_B) - H(\mathbf{X}_A) = \sum_{\mathbf{x}\in\mathcal{X}^m} \sum_{\mathbf{y}\in\mathcal{X}^l} P(\mathbf{X}_A = \mathbf{x}, \mathbf{X}_B = \mathbf{y}) \log \frac{P(\mathbf{X}_A = \mathbf{x})P(\mathbf{X}_B = \mathbf{y})}{P(\mathbf{X}_A = \mathbf{x}, \mathbf{X}_B = \mathbf{y})}. \tag{12}$$

Then use Jensen's inequality (this time for concave functions) to show that

$$\sum_{\mathbf{x}\in\mathcal{X}^m} \sum_{\mathbf{y}\in\mathcal{X}^l} P(\mathbf{X}_A = \mathbf{x}, \mathbf{X}_B = \mathbf{y}) \log \frac{P(\mathbf{X}_A = \mathbf{x})P(\mathbf{X}_B = \mathbf{y})}{P(\mathbf{X}_A = \mathbf{x}, \mathbf{X}_B = \mathbf{y})} \leq 0 \tag{13}$$

3. Use the properties proved in subproblems 1 and 2 to show that entropy is monotonic, i.e., if $A \subseteq B$, then $H(\mathbf{X}_A) \leq H(\mathbf{X}_B)$.

4. Use the properties proved in subproblems 1 and 2 to show that entropy has the diminishing returns property:

$$H(\mathbf{X}_{A\cup\{s\}}) - H(\mathbf{X}_A) \geq H(\mathbf{X}_{B\cup\{s\}}) - H(\mathbf{X}_B) \tag{14}$$

when $A \subseteq B$ and $s \notin B$ is a single index. Therefore, entropy is a submodular set function.