10-704: Information Processing and Learning **Fall 2016**

Lecture 3: September 7

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

We continue our discussion of properties of information theoretic quantities by first discussing data processing inequality and then submodularity.

3.1 Data Processing Inequality

Intuitively, the data processing inequality says that no clever transformation of the received code (channel output) Y can give more information about the sent code (channel input) X than Y itself.

Theorem 1. *(Data processing inequality)*
Suppose we have a probability model described by the following (Markov Chain):

$$X \to Y \to Z$$

where $X \perp Z|Y$, then it must be that $I(X,Y) \geq I(X,Z)$.

Proof. Before we prove this, let us define conditional mutual information $I(X,Y|Z)$ between X and Y given Z as follows:

$$I(X,Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z).$$

Notice that this follows the unconditional property that $I(X,Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$. By the Chain rule for entropy, we can decompose I(X, (Y,Z)) in two ways:

$$I(X,(Y,Z)) = H(X) - H(X|Y,Z) = H(X) - H(X|Y) + H(X|Y) - H(X|Y,Z) = I(X,Y) + I(X,Z|Y)$$

Similarly,

$$I(X,(Y,Z)) = I(X,Z) + I(X,Y|Z)$$

Because $I(X,Z|Y) = 0$ since by assumption $(X \perp Z|Y)$, we have that $I(X,Z) + I(X,Y|Z) = I(X,Y)$. Since mutual information is always non-negative, we get that $I(X,Z) \leq I(X,Y)$. □

3.2 Submodularity

Submodularity introduces and describes a notion of diminishing returns when adding a new element to a set. Submodularity intuitively states that adding an element to a smaller set helps more than adding it to a larger set.

Definition 2. *Let Ω be a finite set. A set function $f : 2^{\Omega} \to \mathbb{R}$ is a submodularity function if any of the following hold (they are all equivalent):*

1. *Adding an element to a subset set has more value than adding the same element to a super set:*

$$\forall X, Y \subseteq \Omega, \quad X \subseteq Y, \quad \forall x \in \Omega \setminus Y, \quad f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$$

The intuition is that if f, for instance, is a utility function, it is submodular if the marginal utility from adding a new element $\{z\}$ to a set of goods X (this marginal utility is $f(X \cup \{z\}) - f(X)$) is larger than the one obtained when $\{z\}$ is added to a larger set $Y \supseteq X$ (this marginal utility is $f(Y \cup \{z\}) - f(Y)$). Submodularity in some way generalizes the idea of decreasing (positive) first derivative of an increasing function. In fact, submodularity is a useful property of functions in optimization problems.

2.
$$\forall S, T \subseteq \Omega \quad f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$$

3.
$$\forall X \subseteq \Omega \quad x_1, x_2 \in \Omega \backslash X \quad f(X \cup \{x_1\}) + f(X \cup \{x_2\}) \geq f(X \cup \{x_1, x_2\}) + f(X)$$

Examples:

1. Linear functions are submodular. A linear set-valued function is parametrized by a vector $w \in \mathbb{R}^{|\Omega|}$ and defined by
$$f(S) = \sum_{x \in S} w_x = w^T \mathbf{1}_S$$
where $\mathbf{1}_S \in \{0,1\}^{|\Omega|}$ is the characteristic (indicator) vector for the set S.

2. Entropy of a collection of random variables is submodular. Let $\Omega = \{X_1, X_2, ..., X_n\}$. Specifically, the function is the joint entropy of a subset of random variables.
$$f(S) = H(\{X_s | s \in S\}) = -\sum p(X_s) \log p(X_s)$$
We will prove this below and also show that mutual information between some sets of random variables is submodular.

Submodularity is a nice property in part because of greedy maximization of a submodular function. If we want to maximize a submodular function f over all sets of size at most K, then a natural algorithm is based on greedy maximization: Starting with the emptyset, we find the single element that leads to the largest increase in f and add it to the candidate solution, and repeat. This algorithm comes with the following approximation guarantee: If the submodular function f is non-decreasing with $f(\emptyset) = 0$ then

$$f(\hat{S}) \geq (1 - \frac{1}{e})(f(\text{OPT})$$

where $f(\hat{S})$ is a the greedy solution and OPT is the optimal solution.

Intuitively what this says is that performance of the greedy solution, i.e. adding the next element which results in the largest incremental benefit performs reasonably well in terms of the optimal solution. Note that finding the optimal solution in general requires exhaustive search over all $\binom{|\Omega|}{K}$ subsets.

3.3 Submodularity of Entropy and Mutual information

3.3.1 Entropy

In this case Ω is a set of random variables. We use capital letters to denote sets of random variables (i.e. $X \subset \Omega$) and lower case letters to denote individual random variables $x \in \Omega$. Note that this contrasts with

the usual notation of using capital letters for random variables and lower case letters for realizations of a random variable.

To show that $H : 2^{\Omega} \to [0, \infty)$ is submodular consider:

$$\underbrace{H(X, z) - H(X)}_{= \underbrace{H(z|X)}_{\text{cond.entropy}}} \geq \underbrace{H(Y, z) - H(Y)}_{= \underbrace{H(z|Y)}_{\text{cond.entropy}}} \tag{3.1}$$

where $H(z|Y) = H\left(z|X \cup (Y \setminus X)\right) \leq H(z|X)$, since conditioning cannot increase entropy.

3.3.2 Mutual information

Notice that

- $I(X, \Omega) = H(X) - H(X|\Omega) = H(X)$, which is therefore submodular (as function of X)
- $I(X, \Omega \setminus X) = H(X) + H(\Omega \setminus X) - H(\Omega)$ is submodular (as function of X). In fact we have:

$$\begin{aligned} A_X &\equiv I(X \cup \{z\}, \Omega \setminus (X \cup \{z\})) - I(X, \Omega \setminus X) \\ &= H(X \cup \{z\}) + H(\Omega \setminus (X \cup \{z\})) - H(X) - H(\Omega \setminus X) \\ &= [H(X \cup \{z\}) - H(X)] + [H(\Omega \setminus (X \cup \{z\})) - H(\Omega \setminus X)] \end{aligned}$$

  and similarly for A_Y. By submodularity of entropy, we have

$$H(X \cup \{z\}) - H(X) \geq H(Y \cup \{z\}) - H(Y)$$

  and

$$H(\Omega \setminus Y) - H(\Omega \setminus (Y \cup \{z\})) \geq H(\Omega \setminus X) - H(\Omega \setminus (X \cup \{z\}))$$

  since $\Omega \setminus (Y \cup \{z\}) \subseteq \Omega \setminus (X \cup \{z\})$. Therefore $A_X \geq A_Y$.

3.3.3 Application to Machine Learning: Sensor placement problem

Suppose we want to monitor the temperature in a room by using k sensors as studied in [krause2008]. Let Ω be a set of random variables corresponding to specific (feasible) locations in the room to place sensors We want to choose the subset $X_k \subseteq \Omega$ with $|X_k| = k$ that would collect information about the temperature of the room the best. Let us consider two possible ways to do it:

1. "Maximum entropy subset selection problem"

$$X_k^* \in \underset{X \subseteq \Omega,\ \text{card(X)=k}}{\operatorname{argmax}} I(X, \Omega) = \underset{X \subseteq \Omega,\ \text{card(X)=k}}{\operatorname{argmax}} H(X) \tag{3.2}$$

   where we find the optimal solution (sensors locations) by maximizing the mutual information of X_k and the total space of sensors. Thus X_k^* is the subset with the largest uncertainty. **Disadvantages**: NP-hard; tendency to place sensors near the boundary (walls of the room) where uncertainty is maximum. **Advantages**: submodularity of the objective function.

2. "Maximum mutual information subset selection problem"

$$X_k^* \in \underset{X\subseteq\Omega,\ \mathrm{card(X)=k}}{\operatorname{argmax}} I(X, \Omega \setminus X) \tag{3.3}$$

Solutions to (3.3) might be better than solutions to (3.2), since we try to put sensors on locations to be most informative about the unsensed locations $(\Omega \setminus X)$. **Disadvantages**: NP-hard. **Advantages**: submodularity of the objective function.

In the next section we deal with the actual maximization of submodular functions, useful to solve problems (3.2) and (3.3).

3.4 Maximizing submodular functions

Algorithm 1 Greedy-Algorithm(Ω, f, k)

Input: A set Ω, A set function $f : 2^\Omega \to \mathbb{R}$, Size of subset k.
Output: A subset $A_k \subset \Omega$ of size k.

$A_0 \leftarrow \emptyset$
For $i = 1, \ldots, k$

1. for $x \in \Omega \backslash A_{i-1}$, set $\delta_x \leftarrow f(A_{i-1} \cup \{x\}) - f(A_{i-1})$
2. $x_* \leftarrow \operatorname{argmax}_{x\in\Omega\backslash A_{i-1}} \delta_x$
3. $A_i \leftarrow A_{i-1} \cup \{x_*\}$

Theorem 3 (Nemhauser et al. (1978)). *Let f be a function such that:*

1. *f is submodular over finite set Ω*
2. *f is monotone, i.e. $\forall X \subseteq Y \subseteq \Omega$, we have $f(Y) \geq f(X)$*
3. *$f(\emptyset) = 0$*

Let $A_k \subseteq \Omega$ be the first k elements chosen by ***Greedy-Algorithm(Ω, f, k)*** *(see Algorithm 1). Then*

$$f(A_k) \geq \left(1 - \frac{1}{e}\right) f(A_{\mathrm{opt}})$$

where $A_{\mathrm{opt}} = \underset{A\subseteq\Omega,\ \mathrm{card(A)=k}}{\operatorname{argmax}} f(A)$.

We will prove it in next class.

References

Krause, Andreas, Singh, A., & Guestrin, C. (2008). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, *9*, 235-284.

Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functionsI. *Mathematical Programming*, *14.1*, 265-294.