

Lecture 7: February 8th

*Lecturer: Professor Alex Dimakis**Scribes: Justin Lewis, Dany Haddad*

This lecture's notes illustrate some uses of various L^AT_EX macros. Take a look at this and imitate.

Topics Covered

- Submodularity
- Feature selection (see [KG05-1])
- Nemhauser's proof for greedy maximization of submodular functions

7.1 Definitions**Entropy**

Given a set, S , of discrete random variables, define the set function $f_H(S) : 2^V \mapsto \mathbb{R}$

$$f_H(S) = H(X_S) = - \sum_{x_i \in S} p(x_i) \log p(x_i)$$

and for differential entropy:

$$f_H(S) = H(X_S) = - \int_{\mathcal{X}_S} p(x) \log p(x) dx$$

Mutual Information

Given random vectors Y and X_S , define the following as the mutual information between them $f_I(S) : 2^V \mapsto \mathbb{R}$

$$f_I(S) = I(Y; X_S) = H(Y) - H(Y|X_S)$$

7.2 Properties

Lemma 7.1. f_H is submodular.

Proof. Consider subsets A and B of random variables, \mathcal{X} , where $A \subseteq B$. Also consider a random variable $X_m \notin A \cup B$

$$f_H(A \cup \{m\}) - f_H(A) = H(X_A, X_m) - H(X_A) = H(X_m|X_A)$$

and similarly

$$f_H(B \cup \{m\}) - f_H(X_B) = H(X_m|X_B)$$

Since conditioning on a larger set of random variables cannot increase the entropy:

$$\begin{aligned} H(X_m|X_B) &\leq H(X_m|X_A) \\ f_H(B \cup \{m\}) - f_H(X_B) &\leq f_H(A \cup \{m\}) - f_H(X_A) \end{aligned}$$

□

In the discrete case we can show that f_H is also monotone. However, in the continuous case, this function is no longer monotone.¹ (TODO: counterexample).

Example 1. Consider X_1, \dots, X_n jointly gaussian random variables with pdf:

$$p(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

The differential entropy of a subset indexed by S is given by:

$$H(X_S) = \frac{1}{2} 2\pi e \log \det \Sigma_S$$

Where Σ_S denotes the submatrix of the covariance matrix Σ formed by taking only the variables indexed by S . So to choose the subset of k variables with the largest entropy, we must maximize the determinant of Σ_S . □

Proposition 7.2. Mutual information is, in general, not submodular.

Proof. Consider X, Y independent $Bernoulli(\frac{1}{2})$ random variables. Let $Z = X \oplus Y$. So:

$$\begin{aligned} H(Z) &= H(Z|X) = H(Z|Y) = 1 \text{ and } H(Z|X \cup Y) = 0 \\ \implies H(Z) - H(Z|X) &\leq H(Z|Y) - H(Z|X \cup Y) \end{aligned}$$

□

Claim 7.3. Mutual information is monotone. This follows immediately from the fact that conditioning does not increase entropy.

Proposition 7.4. TODO: fix this proof Given sets S and U of random variables such that the elements of S are independent of each other conditioned on X_U , then $f_I(A) = I(U; A)$ is submodular for all $A \subseteq S \cup U$.

Proof. Let $W = U \cup S$ and $C = A \cup B$ such that $A, B \subseteq W$. So:

$$\begin{aligned} H(U|C) &= H(U|A \cup B) \\ &= H(U \cup A) - H(A \cup B) \\ &= H(A|U) + H(U) - H(C) \\ &= H(U) - H(C) + \sum_{Y \in C \cap S} H(Y|U) \end{aligned}$$

Where the last step follows using the chain rule for the joint entropy. $H(U)$ is constant, $H(C)$ is submodular and $\sum_{Y \in C \cap S} H(Y|U)$ is linear in C on $U \cup S$ and hence on W . □

This claim holds if the distribution factorizes according to an undirected graphical model similar to 7.1. Recall the conditional independence properties we can infer from an undirected graphical model.

¹see Krause & Golovnia survey: <https://las.inf.ethz.ch/files/krause12survey.pdf>

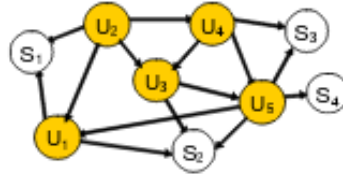


Figure 7.1: An undirected graphical model where the elements of S are independent conditioned on U

7.3 Optimization

Consider the chain rule of entropy:

$$H(X_1, \dots, X_n) = H(X_S) + H(X_{S^c} | X_S)$$

Since $H(X_1, \dots, X_n)$ has no dependence on S , maximizing the entropy of the subset S , is equivalent to minimizing the uncertainty of the unobserved set, S^c :

$$\max_{s: |s| \leq k} H(X_S) = \min_{s: |s| \leq k} H(X_{S^c} | X_S)$$

This requires us to maximize a monotone submodular function. The greedy algorithm selects the element with the largest discrete derivative at iteration i .

7.4 Approx. Submodular Function Maximization

It is well known that maximizing an arbitrary submodular function with given constraint set is NP-Hard.

Finding the optimal solution may be intractable; however, an approximate solution known as the Greedy Algorithm can achieve fair results. More specifically:

Algorithm 1: Greedy Algorithm

Description *blah blah blah*;

Result: Write here the result

initialization;

while *While condition* **do**

 instructions;

if *condition* **then**

 instructions1;

 instructions2;

else

 instructions3;

end

end

Theorem 7.5. Let S^* denote the optimal subset, and S_{greedy} as the Greedy Algorithm selection

Given set function f which is submodular, monotone, non-negative, and $f(\emptyset) = 0$

$$f(S_{\text{greedy}}) \geq [1 - 1/e] \cdot f(S^*) \approx 0.63 \cdot f(S^*)$$

7.4.1 latex reference

Here is an itemized list:

- this is the first item;
- this is the second item.

Here is an enumerated list:

1. this is the first item;
2. this is the second item.

Here is an exercise:

Exercise: Show that $P \neq NP$.

Here is how to define things in the proper mathematical style. Let f_k be the *AND – OR* function, defined by

$$f_k(x_1, x_2, \dots, x_{2^k}) = \begin{cases} x_1 & \text{if } k = 0; \\ AND(f_{k-1}(x_1, \dots, x_{2^{k-1}}), f_{k-1}(x_{2^{k-1}+1}, \dots, x_{2^k})) & \text{if } k \text{ is even;} \\ OR(f_{k-1}(x_1, \dots, x_{2^{k-1}}), f_{k-1}(x_{2^{k-1}+1}, \dots, x_{2^k})) & \text{otherwise.} \end{cases}$$

Theorem 7.6. This is the first theorem.

Proof. This is the proof of the first theorem. We show how to write pseudo-code now.

Consider a comparison between x and y :

```

if  $x$  or  $y$  or both are in  $S$  then
    answer accordingly
else
    Make the element with the larger score (say  $x$ ) win the comparison
    if  $F(x) + F(y) < \frac{n}{t-1}$  then
         $F(x) \leftarrow F(x) + F(y)$ 
         $F(y) \leftarrow 0$ 
    else
         $S \leftarrow S \cup \{x\}$ 
         $r \leftarrow r + 1$ 
    endif
endif

```

This concludes the proof. □

7.5 Next topic

Here is a citation, just for fun [CW87].

References

- [KG05-1] KRAUSE, ANDREAS AND GUESTRIN, CARLOS, “Near-optimal sensor placements,” *Proceedings of the fifth international conference on Information processing in sensor networks - IPSN 06*, 2005.
- [KG05-2] KRAUSE, ANDREAS AND GUESTRIN, CARLOS, “Near-optimal Nonmyopic Value of Information in Graphical Models,” *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005.