EE 381V: Special Topics on Unsupervised Learning **Spring 2018**

Lecture 7: February 8th

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

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 $f_H(S) : 2^{\mathcal{X}} \mapsto \mathbb{R}$

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

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 $f_I(S) : 2^{\mathcal{X}} \mapsto \mathbb{R}$

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

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 $X_m \notin A \cup B$

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

and similarly

$$f_H(X_B, X_{\{m\}}) - f_H(X_B) = H(X_m|X_B)$$

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

$$H(X_m|X_B) \leq H(X_m|X_A)$$
$$f_H(X_B, X_{\{m\}}) - f_H(X_B) \leq f_H(X_A, X_{\{m\}}) - f_H(X_A)$$

□

In the discrete case we can show that f_H is also monotone. However, in the continuous case, this function is no longer monotone, in general[KG14].

Example 7.2. Consider $X_1, ..., X_n$ jointly gaussian random variables with pdf:

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

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.

Consider the covariance matrix:

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

$$\det(\Sigma_{\{0\}}) = 1$$
$$\det(\Sigma_{\{0,1\}}) = 2$$
$$\det(\Sigma_{\{0,1,2\}}) = 0.2$$
$$\det(\Sigma_{\{0,1,2,3\}}) = 0.6$$

So $H(X_{\mathcal{X}})$ is not monotone in this case. □

Note 7.3. In the above example, to choose the subset of k variables with the largest entropy, we must maximize the determinant of Σ_S.

Proposition 7.4. Mutual information is, in general, not submodular.

Proof. Recall the set function for mutual information as defined earlier, $f_I(S) = I(Y; X_S)$

Consider X_1, X_2 independent $Bernoulli(\frac{1}{2})$ random variables. Let $Y = X_1 \oplus X_2$. So:

$$H(Y) = H(Y|X_1) = H(Y|X_2) = 1 \text{ and } H(Y|X_1, X_2) = 0$$
$$\implies H(Y) - H(Y|X_1) \leq H(Y) - H(Y|X_1, X_2)$$
$$\implies f_I(\{1\}|\emptyset) < f_I(\{1\}|\{2\})$$

□

Claim 7.5. Mutual information is monotone. This follows immediately from the fact that conditioning does not increase entropy.

Proposition 7.6. Given sets S and U of random variables such that the elements of S are independent of each other conditioned on U, then $f_I(A) = I(U; A)$ is submodular for all $A \subseteq S \cup U$.

Proof. Let $A \subseteq S \cup U$ and $S_1 \perp\!\!\!\perp S_2$ conditioned on U $\forall S_1, S_2 \subseteq S$.

$$\begin{aligned} I(U;A) &= H(U) - H(U|A) \\ &= H(U) - (H(U \cup A) - H(A)) \\ &= H(U) - (H(A|U) + H(U) - H(A)) \\ &= -\sum_{y \in A \cap S} H(y|U) + H(A) \end{aligned} \tag{7.1}$$

Where the last step follows by conditional independence the elements of S conditioned on U. The first term in equation 7.1 is modular in A and the second is submodular. □

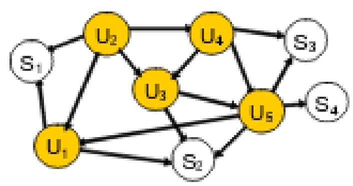

Figure 7.1: An directed graphical model where the elemets of S are independent conditioned on U

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

7.3 Optimization

The sensor selection problem addressed in [KG05-1] is framed as a joint entropy maximization problem. Intuitively, we can think of this problem equivalently as minimizing the uncertainty of the sensors that are not selected (since they will not be observed). One might suggest a PCA type solution to this problem, but unfortunately, these are physical sensors that we are selecting, so we cannot take a linear combination of them! Gram-Schmidt will not save you here!

Consider the chain rule of entropy:

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

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

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

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, in general, NP-Hard.

Problem 1. Given ground set S_g, subset $S \subseteq S_g$, and submodular set function $f(\cdot)$:

$$\max_{S \subseteq S_g} f(S)$$
$$\text{subject to } |S| \leq k$$

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

Define *ground set S_g, subset $S \subseteq S_g$, set function $f(\cdot)$, and cardinality constraint $|S| \leq k$*;
Description *greedily add to S at iteration i the element with the largest discrete derivative*;
Result: S_{greedy}
$S_0 = \emptyset$;
$i = 0$;
while $|S_i| \leq k - 1$ **do**
  $S_{i+1} = S_i \cup \underset{s \in \{S_g \setminus S_i\}}{\arg\max} \left\{ \Delta_f(s \,|\, S_i) \right\}$;
  $i = i + 1$;
end
Return S_{i+1};

Recall: For set function $f : 2^V \to \mathbb{R}, S \subseteq V$, and $e \in V$
let $\Delta_f(e \,|\, S) := f(S \cup \{e\}) - f(S)$

Theorem 7.7. Let S^* denote the optimal subset , and $S_{greedy,\ell}$ as the Greedy Algorithm selection after ℓ iterations. Given set function f which is submodular, monotone, non-negative, and $f(\emptyset) = 0$:

$f(S_{greedy,\ell}) \geq [1 - \exp[-\frac{\ell}{k}]] \cdot f(S^*)$

$f(S_{greedy,\ell=k}) \geq [1 - \frac{1}{e}] \cdot f(S^*) \approx 0.63 \cdot f(S^*)$

Proof. [NW78]

Let S_i denote the Greedy algorithm selection after the i-th iteration

$f(S^*) \leq f(S^* \cup S_i)$ (monotonicity)

Claim 7.8. $f(S^* \cup S_i) = f(S_i) + \sum_{j=1}^{k} \Delta_f\left(v_j^* \,\middle|\, \{S_i \cup \{v_1^*, v_2^*, \ldots, v_{j-1}^*\}\}\right)$

Subproof. Expand:

$$\begin{aligned}
f(S^* \cup S_i) &= f(S_i) + \Delta_f\big(v_1^* \,\big|\, S_i\big) + \Delta_f\left(v_2^* \,\middle|\, \{S_i \cup v_1^*\}\right) + \cdots + \Delta_f\left(v_k^* \,\middle|\, \{S_i \cup \{v_1^*, v_2^*, \ldots, v_{k-1}^*\}\}\right) \\
&= \cancel{f(S_i)} + \cancel{f(S_i \cup v_1^*)} - \cancel{f(S_i)} + \cdots + f(S_i \cup \{v_1^*, v_2^*, \ldots, v_k^*\}) - \cancel{f(S_i \cup \{v_1^*, v_2^*, \ldots, v_{k-1}^*\})} \\
&= f(S^* \cup S_i)
\end{aligned}$$

The telescoping sum leaves only the desired term.

■

With claim above it follows:

$$\begin{aligned}
f(S^*) \le f(S^* \cup S_i) &= f(S_i) + \sum_{j=1}^{k} \Delta_f\left(v_j^* \,\middle|\, \{S_i \cup \{v_1^*, v_2^*, \ldots, v_{j-1}^*\}\}\right) \\
&\le f(S_i) + \sum_{j=1}^{k} \Delta_f(v_j^* \mid S_i) && \text{(by submodularity)} \\
&\le f(S_i) + \sum_{j=1}^{k} [f(S_{i+1}) - f(S_i)] && \text{(by Greedy selection)} \\
&= f(S_i) + k \cdot [f(S_{i+1}) - f(S_i)]
\end{aligned}$$

$$\begin{aligned}
\Rightarrow f(S^*) - f(S_i) &\le k \cdot [f(S_{i+1}) - f(S_i)] \\
\delta_i &\le k \cdot [\delta_i - \delta_{i+1}] && (\delta_i \triangleq f(S^*) - f(S_i)) \\
\delta_{i+1} &\le (1 - \frac{1}{k}) \cdot \delta_i \\
\delta_\ell &\le (1 - \frac{1}{k})^\ell \cdot \delta_0
\end{aligned}$$

$$\begin{aligned}
f(S_\ell) &\ge (1 - (1 - \frac{1}{k})^\ell) \cdot f(S^*) && (\delta_0 = f(S^*) - f(\emptyset) = f(S^*)) \\
f(S_\ell) &\ge (1 - \exp(-\frac{\ell}{k})) \cdot f(S^*) && (1 - x \le \mathrm{e}^{-x}\ \forall x)
\end{aligned}$$

□

7.5 Next Time: Feature Selection

The feature selection problem given by:

$$\max_{\beta} ||X\beta - y||_2^2$$
$$\text{subject to } ||\beta||_0 \leq k$$

can be relaxed by changing the ℓ_0 norm for an ℓ_1 norm. Bringing the constraint into the objective and setting λ as a hyperparameter, we arrive at the LASSO problem:

$$\max_{\beta} ||X\beta - y||_2^2 + \lambda||\beta||_1$$

Alternatively, we can frame the objective as a set function $f(S) = ||X_S \beta_S||_2^2$ and show that if the columns of X are orthogonal, then f is submodular. Otherwise, f is weakly submodular if X satisfies the restricted isometry property (RIP).

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.

[KG14] Krause, Andreas, and Daniel Golovin, "Submodular function maximization," 2014.

[NW78] Nemhauser, George L., Laurence A. Wolsey, and Marshall L. Fisher, "An analysis of approximations for maximizing submodular set functions," *Mathematical Programming*, 1978.