1 Introduction

2 Problem Statement and Application

We first describe the application that motivates our research, and then we provide mathematical formalism to address a more general problem. In the last sub-section we derive our method in solving this problem.

2.1 Motivating application

We have two enzymes (Sfp from *Bacillus subtilis*, and PaAcpH from *Pseudomonas aeruginosa*), and a collection of peptides that can potentially act as a substrate for one or both of these enzymes. Our goal is to find a peptide that acts as a substrate for both of these enzymes, and is as short as possible.

To support this goal, we can do lab experiments, in which we synthesize a peptide and test, for each enzyme, whether it is a substrate or not. We need to find a policy that suggests which peptide to synthesize and test next, so as to reach our goal with as few experiments as possible.

Experiments have parallel setup, thus can be done with a batch of peptides at a time, and so the algorithm suggests a batch of peptides at a time, waiting for the results from the experiment before suggesting the next batch of peptides. A large collection of peptides would be considered by the algorithm for potential synthesis and testing, e.g., all peptides with length less than a given threshold. That is, we would consider more peptides than just those that are sub-peptides of peptides from the literature known to be substrates for one enzyme.

2.2 General Problem Statement

We now formalize and generalize our problem as an active learning problem, which includes but is not limited to our motivating application.

Let E be a generic search space of exemplars. In our motivating application, E is the space of peptides. Each element $x \in E$ has an unknown binary label $y(x) = \{0, 1\}$. A known deterministic function $f(x)$ measures the cost or disutility associated with x. Our goal is to perform experiments so as to find x such that it has positive label and its cost function $f(x)$ is minimum.

To obtain labels of exemplars, we can do a batch of experiments, which evaluate a subset $S \subseteq E$ and obtain labels at each time. We measure quality

of S by

$$f^*(S) = \begin{cases} \min\limits_{x \in S: y(x)=1} f(x), & \text{if } \{x \in S : y(x) = 1\} \neq \emptyset, \\ \infty, & \text{if } \{x \in S : y(x) = 1\} = \emptyset. \end{cases} \tag{1}$$

Let b be a target value and we wish to find $S \subseteq E$ such that $f^*(S)$ is, in some sense, better than b. Specifically, we consider the following two measures:

$$\begin{array}{ll} \text{Probability of Improvement:} & \mathrm{P}^*(S) = \mathbb{P}(f^*(S) < b) \\ \text{Expected Improvement:} & \mathrm{EI}(S) = \mathbb{E}[(b - f^*(S))^+] \end{array} \tag{2}$$

We wish to find S that maximize one of these two measures. Let $g(S)$ be either $\mathrm{P}^*(S)$ or $\mathrm{EI}(S)$ and let the cardinality of S be the only constraint on S. Our goal is then:

$$\max_{S \subseteq E: |S| < k} g(S) \tag{3}$$

3 Solution Method

We solve (3) using greedy heuristic, that is, starting with empty set $S = \emptyset$, find element $e = \arg\max_e g(S \cup \{e\}) - g(S)$ to include in S iteratively until $|S| = K$ for some chosen K. We show first the solution using greedy heuristic has a lower bound, and then present our method.

3.1 Lower bound of greedy algorithm

We claim that if objective function is probability of improvement (i.e $\mathrm{P}^*(S)$) or expected improvement (i.e $\mathrm{EI}(S)$), the greedy algorithm is guaranteed to achieve a factor $(1 - 1/e)(\approx 63\%)$ of the optimal value. This lower bound is obtained from a theorem stated in the following:

Theorem 1. *(Nemhauser, Wolsey, & Fisher (1978)) If $F(S)$ is submodular, nondecreasing and $F(\emptyset) = 0$, the greedy heuristic always produces a solution whose value is at least $1 - [(K - 1)/K]^K$ times the optimal value, where $|S| \leq K$. This bound can be achieved for each K and has a limiting value of $1 - 1/e$, where e is the base of the natural logarithm.*

If we can show our objective functions meet condition in Theorem 1, we find lower bound of the greedy solution.

Theorem 2. *Probability of improvement $P^*(S)$ is submodular, nondecreasing and $P^*(\emptyset) = 0$.*

Proof.

- $\mathrm{P}^*(\emptyset) = \mathbb{P}(f^*(\emptyset) < b) = \mathbb{P}(\infty < b) = 0$.
- Suppose $A \subseteq B \subseteq E$ where E is a finite set.

$$\begin{aligned}
\mathrm{P}^*(B) &= \mathbb{P}(f^*(B) < b) \\
&= \mathbb{P}(f^*(B) < b | f^*(A) \geq b)\mathbb{P}(f^*(A) \geq b) + \mathbb{P}(f^*(B) < b | f^*(A) < b)\mathbb{P}(f^*(A) < b) \\
&= \mathbb{P}(f^*(B) < b | f^*(A) \geq b)\mathbb{P}(f^*(A) \geq b) + \mathbb{P}(f^*(A) < b) \\
&\geq \mathbb{P}(f^*(A) < b) \\
&= \mathrm{P}^*(A)
\end{aligned}$$

- For $e \in E \backslash B$,

$$\begin{aligned}
\mathrm{P}^*(A \cup \{e\}) - \mathrm{P}^*(A) &= \mathbb{P}(f^*(A \cup \{e\}) < b) - \mathbb{P}(f^*(A) < b) \\
&= \mathbb{P}(f^*(A \cup \{e\}) < b | f^*(A) < b)\mathbb{P}(f^*(A) < b) + \\
&\mathbb{P}(f^*(A \cup \{e\}) < b | f^*(A) \geq b)\mathbb{P}(f^*(A) \geq b) - \mathbb{P}(f^*(A) < b) \\
&= \mathbb{P}(f^*(A) < b) + \mathbb{P}(f^*(A \cup \{e\}) < b | f^*(A) \geq b)\mathbb{P}(f^*(A) \geq b) - \\
&\mathbb{P}(f^*(A) < b) \\
&= \mathbb{P}(f^*(A \cup \{e\}) < b | f^*(A) \geq b)\mathbb{P}(f^*(A) \geq b) \\
&= \mathbb{P}(f(e) < b, y(e) = 1 | f^*(A) \geq b)\mathbb{P}(f^*(A) \geq b) \\
&= \mathbb{P}(f(e) < b, y(e) = 1, f^*(A) \geq b)
\end{aligned}$$

Using similar argument,

$$\begin{aligned}
\mathrm{P}^*(B \cup \{e\}) - \mathrm{P}^*(B) &= \mathbb{P}(f(e) < b, y(e) = 1, f^*(B) \geq b) \\
&= \mathbb{P}(f(e) < b, y(e) = 1, f^*(A) \geq b, f^*(B \backslash A) \geq b)
\end{aligned}$$

Therefore, $\mathrm{P}^*(A \cup \{e\}) - \mathrm{P}^*(A) \geq \mathrm{P}^*(B \cup \{e\}) - \mathrm{P}^*(B)$, thus we conclude that $\mathrm{P}^*(S)$ is submodular.

□

Theorem 3. *Expected improvement* $\mathrm{EI}(S)$ *is submodular, nondecreasing and* $\mathrm{EI}(\emptyset) = 0$.

Proof.

- $\mathrm{EI}(\emptyset) = \mathbb{E}[(b - f^*(\emptyset))^+] = \mathbb{E}[0] = 0$.
- Suppose $A \subseteq B \subseteq E$ where E is a finite set. Since $f^*(B) \leq f^*(A)$, $b - f^*(B) \geq b - f^*(A)$, and $(b - f^*(B))^+ \geq (b - f^*(A))^+$, therefore, $\mathbb{E}[(b - f^*(B))^+] \geq \mathbb{E}[(b - f^*(A))^+]$.
- For $e \in E \backslash B$, consider $\mathbb{E}[(b - f^*(A \cup \{e\}))^+] - \mathbb{E}[(b - f^*(A))^+]$. We can write

$$(b - f^*(A \cup \{e\}))^+ = \begin{cases} (b - f^*(A))^+ & \text{if } y(e) = 0 \\ (b - \min\{f(e), f^*(A)\})^+ & \text{if } y(e) = 1 \end{cases}$$

Then

$$\begin{aligned} &\mathbb{E}[(b - f^*(A \cup \{e\}))^+] - \mathbb{E}[(b - f^*(A))^+] \\ &= \mathbb{P}(y(e) = 1)\mathbb{E}[(b - \min\{f(e), f^*(A)\})^+ - (b - f^*(A))^+ | y(e) = 1] \\ &= \mathbb{P}(y(e) = 1)\mathbb{P}(f(e) < f^*(A) | y(e) = 1)\mathbb{E}[(b - e)^+ - (b - f^*(A))^+ | y(e) = 1, f(e) < f^*(A)] \\ &= \mathbb{E}[\mathbb{1}_{y(e)=1, f(e)<f^*(A)}((b - e)^+ - (b - f^*(A))^+)] \end{aligned}$$

Since $f^*(A) \geq f^*(B)$, $\mathbb{1}_{y(e)=1, f(e)<f^*(A)}((b - e)^+ - (b - f^*(A))^+)) \geq \mathbb{1}_{y(e)=1, f(e)<f^*(B)}((b - e)^+ - (b - f^*(B))^+))$, thus

$$\mathrm{EI}(A \cup \{e\}) - \mathrm{EI}(A) \geq \mathrm{EI}(B \cup \{e\}) - \mathrm{EI}(B)$$

$\mathrm{EI}(S)$ is submodular.

□

3.2 Greedy Algorithm

Suppose we have chosen $S = \{x_1, x_2, \ldots, x_n\}$ as a batch of points we are going to evaluate next, and if we want to incorporate one more point e, which is distinct from $x_1, x_2, \ldots, x_n$, such that the objective function increases most, we use the following criterion to find e:

$$\arg\max_{e \in E \backslash S} g(S \cup \{e\}) \tag{4}$$

3.2.1 Probability of Improvement

In the case that objective function is P^*, we rewrite (4) as

$$\arg\max_{e\in E\setminus S} \mathrm{P}^*(S\cup\{e\}). \tag{5}$$

Since

$$\begin{aligned}\mathrm{P}^*(S\cup\{e\}) &= \mathbb{P}(f^*(S\cup\{e\})<b)\\ &= \mathbb{P}(f^*(S)<b)+\mathbb{P}(f^*(S)\geq b)\mathbb{P}(f(e)<b, y(e)=1|f^*(S)\geq b),\end{aligned}$$

we can rewrite (5) as

$$\arg\max_{e\in E\setminus S} \mathbb{P}(f(e)<b, y(e)=1|f^*(S)\geq b). \tag{6}$$

Thus we use (6) as search criterion for our greedy approach. Note that when $f(e)\geq b$, $\mathbb{P}(f(e)<b, y(e)=1|f^*(S)\geq b)=0$, thus our algorithm will always propose e such that $f(e)<b$. Therefore, it is reasonable to assume that $f(x)<b$ for $\forall x\in S$, and $f^*(S)\geq b$ means $y(x)=0$ for $\forall x\in S$. Now we can write (6) as

$$\arg\max_{e\in E\setminus S, f(e)<b} \mathbb{P}(y(e)=1|y(x)=0, \forall x\in S). \tag{7}$$

3.2.2 Expected Improvement

If objective function is EI, rewrite (4) as

$$\arg\max_{e\in E\setminus S} \mathbb{E}\left[(b-f^*(S\cup\{e\}))^+\right]. \tag{8}$$

Since choosing e such that $f(e)\geq b$ has no contribution to the objective function, by using similar argument as dealing with probability of improvement, we argue that $f(x)<b$ for $\forall x\in S$. Thus

$$f^*(S)\begin{cases}=\infty & \text{if } y(x)=0 \text{ for } \forall x\in S,\\ <b & \text{else.}\end{cases}$$

Now objective function we want to maximize becomes

$$\begin{aligned}&\mathbb{E}\left[(b-f^*(S\cup\{e\}))^+\right]\\ &=\mathbb{E}[(b-f(e))^+\mathbb{1}_{f^*(S)=\infty, y(e)=1}]+\mathbb{E}[(b-f^*(S\cup\{e\}))^+\mathbb{1}_{f^*(S)<b}]\\ &=\mathbb{E}[(b-f(e))^+\mathbb{1}_{f^*(S)=\infty, y(e)=1}]+\mathbb{E}[(b-f^*(S))\mathbb{1}_{f^*(S)<b}]+\mathbb{E}[(f^*(S)-f(e))\mathbb{1}_{y(e)=1, f(e)<f^*(S)<b}].\end{aligned}$$

Equation (8) is equivalent to

$$\arg\max_{e\in E\backslash S, f(e)<b} \mathbb{E}[(b-f(e))\mathbb{1}_{f^*(S)=\infty, y(e)=1}] + \mathbb{E}[(f^*(S)-f(e))\mathbb{1}_{y(e)=1, f(e)<f^*(S)<b}]. \tag{9}$$

For $e \in E\backslash S, f(e) < b$,

$$\begin{aligned}
&\mathbb{E}[(b-f(e))\mathbb{1}_{f^*(S)=\infty, y(e)=1}] = (b-f(e))\mathbb{P}(y(e)=1, y(x)=0, \forall x \in S) \\
&\mathbb{E}[(f^*(S)-f(e))\mathbb{1}_{y(e)=1, f(e)<f^*(S)<b}] \\
&= \mathbb{E}[\mathbb{E}[(f^*(S)-f(e))\mathbb{1}_{y(e)=1, f(e)<f^*(S)<b}]|f^*(S)=l] \\
&= \sum_{l\in L, f(e)<l} \mathbb{P}(y(e)=1|f^*(S)=l)(l-f(e))\mathbb{P}(f^*(S)=l),
\end{aligned}$$

where $L = \{f(x) : x \in S\}$. If we rank elements in S such that $f(x_i) \leq f(x_j), \forall i < j, x_i, x_j \in S$, we can write equation above as

$$\sum_{i=1}^{|S|} \mathbb{P}(y(e)=1, y(x_i)=1, y(x_j)=0, \forall j<i, x_i, x_j \in S)(f(x_i)-f(e))^+$$

Since $\mathbb{P}(y(e)=1, \mathcal{F}(x_1, \ldots, x_{|S|}) \propto \mathbb{P}(y(e)=1|\mathcal{F}(x_1, \ldots, x_{|S|})$, and coefficient is known given S, we can write our criterion for greedy algorithm as

$$\arg\max_{e\in E\backslash S} c_0\mathbb{P}_0(e)(b-f(e))^+ + \sum_{i=1}^{|S|} c_i\mathbb{P}_i(e)(f(x_i)-f(e))^+, \tag{10}$$

where

$$\begin{aligned}
&\mathbb{P}_0(e) = \mathbb{P}(y(e)=1|y(x)=0, \forall x \in S) \\
&\mathbb{P}_i(e) = \mathbb{P}(y(e)=1|y(x_i)=1, y(x_j)=0, \forall j<i, x_i, x_j \in S),
\end{aligned}$$

and $c_i, i = 0, \ldots, |S|$ are known coefficients.

4 Application

4.1 Statistical Method

something really brief on how we set up Naive Bayes, and how we select the hyperparameters. [Pu, please add something here]