Sample Average Approximation for Optimizing Probability of Improvement and Expectation of Improvement

August 14, 2013

Let $\theta^{(k)}$ represent the full set of Naive Bayes parameters, $\theta_{y,i}^{(k)}(j)$, for $y = 0, 1$ being whether the peptide is a substrate for enzyme k, $i = -20, ..., -1, 1, ..., 20$ being position of amino-acids in the peptide and $j = 1, ..., 8$ denoting the class each amino-acid belongs to. Here k denotes different types of enzyme, and in our case $k = 1, 2$.

We have

$$P(Y(x,k) = 1|\theta^{(k)}) = \frac{P(Y(x,k) = 1) \prod_i \theta_{1,i}^{(k)}(x_i)}{P(Y(x,k) = 1) \prod_i \theta_{1,i}^{(k)}(x_i) + P(Y(x,k) = 0) \prod_i \theta_{0,i}^{(k)}(x_i)}$$

$$P(Y(x,k) = 0|\theta^{(k)}) = \frac{P(Y(x,k) = 0) \prod_i \theta_{0,i}^{(k)}(x_i)}{P(Y(x,k) = 1) \prod_i \theta_{1,i}^{(k)}(x_i) + P(Y(x,k) = 0) \prod_i \theta_{0,i}^{(k)}(x_i)}$$

Therefore,

$$\frac{P(Y(x,k) = 1|\theta^{(k)})}{P(Y(x,k) = 0|\theta^{(k)})} \propto \prod_i \frac{\theta_{1,i}^{(k)}(x_i)}{\theta_{0,i}^{(k)}(x_i)}$$

We let

$$\eta_i^{(k)}(x_i) = \frac{\theta_{1,i}^{(k)}(x_i)}{\theta_{0,i}^{(k)}(x_i)}$$

Then we have

$$P(Y(x,k) = 1|\theta^{(k)}) = \frac{\prod_i \eta_i^{(k)}(x_i)}{1 + \prod_i \eta_i^{(k)}(x_i)}$$

Now, suppose we're given a set of N peptides: $S = (X)_{n=1}^{n}$, then the probability that all peptides in S are not substrate for enzyme k is:

$$
\begin{aligned}
P(Y(X_1,k)=0,...,Y(X_N,k)=0) &= E_{\theta^{(k)}}[P(Y(X_1,k)=0,...,Y(X_N,k)=0|\theta^{(k)})] \\
&= E_{\theta^{(k)}}[\prod_n P(Y(X_n,k)=0|\theta^{(k)})] \\
&= E_{\theta^{(k)}}[\prod_n \frac{1}{1+\prod_i \eta_i^{(k)}(x_i)}]
\end{aligned}
$$

The second equality follows from that given $\theta^{(k)}$, the probability each peptide in S is a substrate for enzyme k is independent. To estimate the above expectation, we can use the sample average approximation approach. We first simulate a set of L different $\eta^{(k)}$ parameters: $\{\eta_{i,l}^{(k)}(j) : i = -20, ..., -1, 1, ...20, j = 1, ..8\}_{l=1}^{L}$, then we estimate the above expectation as

$$
\hat{E}_{\theta^{(k)}}[\prod_n \frac{1}{1+\prod_i \eta_i^{(k)}(x_i)}] = \frac{1}{L}\sum_{l=1}^{L} \frac{1}{\prod_n[1+\prod_i \eta_{i,l}^{(k)}(x_i)]}
$$

To increase the probability of improvement, we should minimize the above the expectation. If we introduce a new set of variables z:

$$
z_{i,j}^n = \begin{cases} 1 & \text{if } x_{n,i} = j \\ 0 & \text{otherwise} \end{cases}
$$

we're facing an optimization problem:

$$
\begin{aligned}
&\min_z \frac{1}{L}\sum_{l=1}^{L} \frac{1}{\prod_n[1+\prod_i(\sum_j \eta_{i,l}^{(k)}(j) z_{i,j}^n)]} \\
&\text{s.t.} \quad \sum_{j=1}^{8} z_{i,j}^n = 1 \quad i = -20, ..., -1, 1, ...20, n = 1, ..., N \\
&\qquad\ \ z_{i,j}^n \in \{0,1\} \quad\ \ i = -20, ..., -1, 1, ...20, j = 1, ..., 8, n = 1, ..., N
\end{aligned}
$$