Maximum entropy modeling

1 Introduction

The problem of maximum entropy modeling is:

Problem 1 (maximum entropy) *Find the probability distribution p that maximizes entropy*

$$H(p) = -\sum_{x \in X} p(x) \ln p(x) \tag{1}$$

given constraints:

$$\sum_{x \in X} p(x) = 1, \tag{2}$$

$$\sum_{x \in X} p(x) T_i(x) = \alpha_i, \quad 1 \leq 1 \leq m. \tag{3}$$

The solution of the maximum entropy problem is as follows:

Theorem 1 *If there exists distribution*

$$p^*(x) = \exp\left[\lambda_0^* + \sum_{i=1}^{m} \lambda_i^* T_i(x)\right], \tag{4}$$

where λ_i^* are chosen so that p^* satisfies conditions (2)–(3), then p^* maximizes entropy (1) on the space of probability distributions that satisfy (2)–(3).

Theorem 2 *Consider the distribution p_λ and the Lagrangian function $L(\lambda)$ defined*

$$p_\lambda(x) = \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x) - \ln Z(\lambda)\right], \tag{5}$$

$$L(\lambda) = \ln Z(\lambda) - \sum_{i=1}^{k} \lambda_i \alpha_i, \tag{6}$$

where the canonical sum is

$$Z(\lambda) = \sum_{x \in X} \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right].$$

Function $L(\lambda)$ has a single minimum and $\lambda^* = (\lambda_1^*, ..., \lambda_m^*)$ for which (5) satisfies conditions (3) is the solution of

$$\lambda^* = \arg\min_{\lambda} L(\lambda). \tag{7}$$

2 Improved iterative scaling algorithm

Let us assume that

$$T_i(x) \geq 0. \tag{8}$$

In this case we can find the minimum of the Lagrangian function $L(\lambda)$ via the improved iterative scaling algorithm.

Let $\delta = (\delta_1, ..., \delta_m)$. By inequality $\log y \leq y - 1$ we have

$$\begin{aligned} L(\lambda + \delta) - L(\lambda) \leq A(\lambda, \delta) &:= \frac{Z(\lambda + \delta)}{Z(\lambda)} - 1 - \sum_{i=1}^{m} \delta_i \alpha_i \qquad (9) \\ &= \sum_{x \in X} p_\lambda(x) \exp\left[\sum_{i=1}^{m} \delta_i T_i(x)\right] - 1 - \sum_{i=1}^{m} \delta_i \alpha_i. \qquad (10) \end{aligned}$$

Function $y \mapsto \exp y$ is convex, hence $\exp\left[\sum_{i=1}^{m} n_i g_i\right] \leq \sum_{i=1}^{m} n_i \exp g_i$ for $n_i \geq 0$, $\sum_{i=1}^{m} n_i = 1$ by the Jensen inequality. Then putting $T_+(x) = \sum_{i=1}^{m} T_i(x)$ and setting $n_i = T_i(x)/T_+(x)$ and $g_i = \delta_i T_+(x)$ we obtain

$$A(\lambda, \delta) \leq B(\lambda, \delta) = \sum_{x \in X} p_\lambda(x) \sum_{i=1}^{m} \frac{T_i(x)}{T_+(x)} \exp\left[\delta_i T_+(x)\right] - 1 - \sum_{i=1}^{m} \delta_i \alpha_i. \tag{11}$$

The derivatives of $B(\lambda, \delta)$ w.r.t δ are

$$\frac{\partial B(\lambda, \delta)}{\partial \delta_i} = B'(\lambda, \delta_i) := \sum_{x \in X} p_\lambda(x) T_i(x) \exp\left[\delta_i T_+(x)\right] - \alpha_i, \tag{12}$$

$$\frac{\partial^2 B(\lambda, \delta)}{\partial \delta_i^2} = B''(\lambda, \delta_i) := \sum_{x \in X} p_\lambda(x) T_i(x) T_+(x) \exp\left[\delta_i T_+(x)\right]. \tag{13}$$

In the improved iterative scaling algorithm, we approximate finding the minimum of $L(\lambda)$ via stepwise finding of the minima of $B(\lambda, \delta)$ using the Newton's method. The minimum of $B(\lambda, \delta)$ corresponds to condition

$$B'(\lambda, \delta_i) = 0 \tag{14}$$

for all i. In the Newton's method, the zero of the derivative of $B'(\lambda, \delta_i)$ can be found by iteration

$$\delta_i \leftarrow \delta_i - \frac{B'(\lambda, \delta_i)}{B''(\lambda, \delta_i)} \tag{15}$$

until sufficient convergence is observed. Hence the improved iterative scaling algorithm is as follows, cf. Berger (1997); Berger et al. (1996):

procedure Improved iterative scaling
  for $i \in \{1, ..., k\}$ **do**
    $\lambda_i \leftarrow 0$
  end for
  repeat
    for $i \in \{1, ..., k\}$ **do**
      $\delta_i \leftarrow 1$

while $\left|\frac{B'(\lambda,\delta_i)}{B''(\lambda,\delta_i)}\right| > \epsilon$ **do**

$\delta_i \leftarrow \delta_i - \frac{B'(\lambda,\delta_i)}{B''(\lambda,\delta_i)}$

end while

$\lambda_i \leftarrow \lambda_i + \delta_i$

end for

until $\max_{i\in\{1,...,k\}} |\delta_i| > \epsilon$

for $i \in \{1, ..., k\}$ **do**

return λ_i

end for

end procedure

3 Task

1. Download some texts, DNA sequences, or other discrete symbolic sequences (e.g., music in an appropriate format) in a sufficient amount (say, about 1MB) from the internet.

2. Let $x_1, x_2, ..., x_n$ be the consecutive bytes of the text and let X be the set of all possible bytes.

3. Consider the following features for $s = 1, ..., 8$ and $a \in X$:

$$T_s(x) = \begin{cases} 1, & \text{the } s\text{-th bit of byte } x \text{ is 1.} \\ 0, & \text{else.} \end{cases} \tag{16}$$

$$T_a(x) = \begin{cases} 1, & x \text{ is the character } a. \\ 0, & \text{else.} \end{cases} \tag{17}$$

4. Compute the averages

$$\alpha_i = \frac{1}{n}\sum_{k=1}^{n} T_i(x_k). \tag{18}$$

5. Find the maximum entropy model p^* for features $(T_a)_{a\in X}$. (This can be done without numerical minimization of the Langrangian function! Try to solve the problem analytically as far as possible.) Report $p^*(x)$ for all $x \in X$.

6. Find the maximum entropy model p^* for features $(T_s)_{s=1}^{8}$. (This requires numerical minimization of the Langrangian function.) Report $p^*(x)$ for all $x \in X$ and λ_s^* for $s = 1, ..., 8$.

7. Compute the entropy $H(p^*)$ and cross entropy $-\frac{1}{n}\sum_{k=1}^{n} \log p^*(x_k)$ for these two models.

8. Describe what you have obtained in a report, attach the used scripts, and send it to me (`ldebowsk@ipipan.waw.pl`).

References

Berger, A. L., 1997. The improved iterative scaling algorithm: A gentle introduction, Carnegie Mellon University.

Berger, A. L., Della Pietra, S. A., Della Pietra, V. J., 1996. A maximum entropy approach to natural language processing. Computational Linguistics 22, 39–71.