CS3236 Project Report
Maximum Entropy

Xiang Pan
A0213836E
xiangpan.cs@gmail.com

04/09/2020

1 Introduction

This project is about maximum entropy. Firstly, we demonstrated the definition of the maximum entropy distribution and the maximum entropy principle. Then, we introduced applications of maximum entropy principle. Finally, we give an implementation of maximum entropy model and analyzed the result.

2 Maximum Entropy Principle

2.1 A Simple Example

In daily life, many things happen to show certain randomness, the results of the experiment are often uncertain, and we don't know the probability distribution that this random phenomenon obeys. Using some observed test samples or sample characteristics, how to make a reasonable inference without a prior distribution? For the simplest example(also the most used example), when we are asked what the probability of the dice showing 6 when we randomly throw a dice, we usually say 1/6. So why we can give such an answer? Just as stated above, we do not know the probability distribution of dice, we do not know any prior knowledge of such a probability system. Actually, we get the intuitive result with the usage of maximum entropy without notice.

2.2 General Understanding

The simplest way to illustrate maximum entropy principle is 'Model all that is known and assume nothing about that which is unknown'.

We know that entropy actually defines an uncertain when the entropy is the largest, it means that the random variable is the most uncertain, in other words, the random variable is the most random, and its behavior is accurate Prediction is the most difficult. In this sense, the essence of the principle of maximum entropy is that, given the knowledge of some knowledge, about The most reasonable inference for an unknown distribution is the most uncertain or random inference that fits known knowledge. This is the only reason we can make an unbiased choice, any other choice means that we add other constraints and assumptions, these constraints and falsely suppose which cannot be made based on the information we have.

3 Maximum Entropy Distribution

3.1 Definition

Formulaly, we have the entropy

$$H(X) = -\sum_{x\epsilon X} p(x) lg p(x) \tag{1}$$

And the objective function of maximum entropy distribution is:

$$max(p\epsilon P) H(X) = -\sum_{(x)} p(y) log p(x) \tag{2}$$

where the $P = \{p | p \in the\ distributions\ satisfiy\ the\ constraints\}$

3.2 Solution

We need to find the optimized the distribution p (or f demonstrated in Thomas Cover book [1]). We can use Lagrange multiplier because the differential entropy $h(p)$is a concave function over a convex set. (The book choose the continuous form, thus we choose the discrete form)

3.2.1 Solution Example

For a general example, we do not have additional constraints and we choose the discrete form. Thus the only constraint is:

$$g(p_1, p_2, ..., p_n) = \sum_{i=1}^{n} p_i = 1 \tag{3}$$

$$F(p_1, p_2, ..., p_n) = h(p_1, p_2, ..., p_n) + \lambda * (g(p_1, p_2, ..., p_n) - 1) \tag{4}$$

We can do deviation for every p_i:

$$\frac{\partial}{\partial p_i}\left(-\sum_{i=1}^{n} p_i \log_2 p_i + \lambda\left(\sum_{i=1}^{n} p_i - 1\right)\right) = 0 \tag{5}$$

we can get

$$-\left(\frac{1}{\ln 2} + \log_2 p_i\right) + \lambda = 0 \tag{6}$$

$$p_i = \frac{1}{n} \tag{7}$$

The result shows when there is no additional constraint, the average distribution is the maximum entropy distribution for discrete form. Such a result can also answer the mice question above.

3.2.2 The proof of maximum

In the Thomas Cover Book, they form a optimization problem, and a solution, Then proof the solution is indeed the maximum entropy distribution. We can use the information inequality below to prove the optimization.

$$0 \leq D\left(p \| p^*\right) = -h(p) + h\left(p^*\right) \tag{8}$$

Where p^* is the result.

For the Lagrange multiplier one, it is harder to generalize but it directly solves the maximum entropy objective function. (It is just another expression of the calculus method in the book, but it may be more directly)

3.3 Trivial case only with the constraint of entropy

Every probability distribution is trivially a maximum entropy probability distribution under the constraint that the distribution has its own entropy.

To proof this, rewire the $p(x)$

$$p(x) = \exp\left(\ln p(x)\right) \tag{9}$$

For the continuous situation,by choosing the $\ln p(x) \rightarrow f(x)$, we can get

$$\int \exp\left(f(x)\right) f(x) dx = -H \tag{10}$$

3.4 With Constraints Case

In a similar way, we can do the Lagrange multiplier for continuous form, such a process is also demonstrated in the Thomas Cover Book.

For different constraints, we can get different maximum entropy distribution. We only need to change the Lagrange multiplier part and do the optimization.

3.5 Cases Examples

We only show the most common case of constraints. Those maximum entropy distribution can get from the above process.

3.5.1 Discrete Distribution

Maximum Entropy Constraint	Maximum Entropy Distribution
Given Interval	Uniform Distribution
Given Mean	Uniform Distribution
Given Interval and standard deviation (variance)	Normal Distribution
$\mathbf{E}(x) = \mu, f \in \mathbf{n}-$ generalized binomial distribution	$f(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$
$\mathbf{E}(x) = \lambda, f \in \infty$ -generalized binomial distribution	$f(k) = \frac{\lambda^k \exp(-\lambda)}{k!}$

3.5.2 Constraints Distribution

Maximum Entropy Constraint	Maximum Entropy Distribution
$\mathrm{E}(x) = \frac{1}{\lambda}$	$f(x) = \lambda \exp(-\lambda x)$
$\mathbf{E}(x) = \mu, \mathbf{E}\left((x-\mu)^2\right) = \sigma^2$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
$\mathbf{E}(\ln(x)) = \psi(\alpha) - \psi(\alpha+\beta)$ $\mathbf{E}(\ln(1-x)) = \psi(\beta) - \psi(\alpha+\beta)$	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ for $0 \leq x \leq 1$

4 Applications

There are usually four applications using the principle of maximum entropy[2] to inferential problems:

(i) Prior Probabilities(Naive Bayes)

(ii) Posterior Probabilities

(iii) Maximum Entropy Models

(iv) Probability Density Estimation

(v) Maximum Entropy on Markov Chain

4.1 Maximum Entropy Models

For Maximum Entropy Models [3], we can modify the equation 2 to conditional one to present the dependency between input and output:

$$max(p\epsilon P)H(Y|X) = -\sum_{(x,y)} p(x,y)logp(y|x) \tag{11}$$

$$H(p) = -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x) \tag{12}$$

Similarly, n optimization can be formulated as

$$\begin{array}{ll} \max_{p\in P} & H(p)=-\sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x) \\ \text{s.t.} & E_p\left(f_i\right)=E_{\tilde{p}}\left(f_i\right), \quad i=1,2,\cdots,d \\ & \sum_y p(y|x)=1 \end{array} \tag{13}$$

The using Lagrange multiplier:

$$L(p,\lambda)=\sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x)+\lambda_0(1\sum_y p(y|x))+\sum_{i=1}^{d}\lambda_i\left(\sum_{x,y}\tilde{p}(x,y)f_i(x,y)\sum_{x,y}\tilde{p}(x)p(y|x)f_i(x,y)\right) \tag{14}$$

Then, do the partial derivative:

$$\begin{aligned} \frac{\partial L(p,\lambda)}{\partial p(y|x)} &= \sum_{x,y}\tilde{p}(x)(\log p(y|x)+1)-\sum_y\lambda_0-\sum_{i=1}^{d}\lambda_i\left(\sum_{x,y}\tilde{p}(x)f_i(x,y)\right) \\ &= \sum_{x,y}\tilde{p}(x)(\log p(y|x)+1)-\sum_x\tilde{p}(x)\sum_y\lambda_0-\sum_{x,y}\tilde{p}(x)\sum_{i=1}^{d}\lambda_i f_i(x,y) \\ &= \sum_{x,y}\tilde{p}(x)(\log p(y|x)+1)-\sum_{x,y}\tilde{p}(x)\lambda_0-\sum_{x,y}\tilde{p}(x)\sum_{i=1}^{d}\lambda_i f_i(x,y) \\ &= \sum_{x,y}\tilde{p}(x)\left(\log p(y|x)+1-\lambda_0-\sum_{i=1}^{d}\lambda_i f_i(x,y)\right) \end{aligned} \tag{15}$$

Then the extremum can be got when the partial derivative equals zero.

We can get the:

$$p_\lambda=p(y|x)=\frac{1}{Z_\lambda(x)}e^{\sum_{i=1}^{d}\lambda_i f_i(x,y)}where, Z_\lambda(x)=\sum_y e^{\sum_{i=1}^{d}\lambda_i f_i(x,y)} \tag{16}$$

This p_λ is the solution of the maximum entropy model, which has an exponential form, $f_i(x,y)$ is the feature function, λ_i is the weight of the feature, the larger the λ_i, the more important the feature.

Through such a process, we can accomplish the feature selection.

We have a solution, then just need to maximize the objective function:

$$\begin{aligned}
\psi(\lambda) &= \sum_{x,y} \tilde{p}(x) p_\lambda \log p_\lambda + \sum_{i=1}^{d} \lambda_i \left(\sum_{x,y} \tilde{p}(x,y) f_i(x,y) - \sum_{x,y} \tilde{p}(x) p_\lambda f_i(x,y) \right) \\
&= \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^{d} \lambda_i f_i(x,y) + \sum_{x,y} \tilde{p}(x) p_\lambda \left(\log p_\lambda - \sum_{i=1}^{d} \lambda_i f_i(x,y) \right) \\
&= \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^{d} \lambda_i f_i(x,y) - \sum_{x,y} \tilde{p}(x) p_\lambda \log Z_\lambda(x) \\
&= \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^{d} \lambda_i f_i(x,y) - \sum_{x} \tilde{p}(x) \log Z_\lambda(x)
\end{aligned} \tag{17}$$

The Maximum Entropy's solution contains the exponent form. In fact, such a form can be expressed uniformly in exponential families of distributions.

4.2 Burg's Maximum Entropy Theorem

4.2.1 Max Entropy Rate Stochastic processes

$$\begin{aligned}
H\left(X_1, \ldots, X_n\right) &= \sum_{i=1}^{n} H\left(X_i \mid X_{i-1} \ldots X_1\right) \\
&\leq \sum_{i=1}^{n} H\left(X_i\right) = nH(X)
\end{aligned} \tag{18}$$

For a stochastic process with arbitrary dependencies, we have:

$$H(\mathcal{X}) := \lim_{n \to \infty} \frac{H\left(X_1, \ldots, X_n\right)}{n} \tag{19}$$

Every symbols' entropy is limited, thus we have :

$$H'(\mathcal{X}) := \lim_{n \to \infty} H\left(X_n \mid X_{n-1}, \ldots, X_1\right) \tag{20}$$

Stationary stochastic process: satisfy the shift properties:

$$\begin{aligned}
&p\left(X_1, \ldots, X_n\right) = p\left(X_{1+l}, \ldots, X_{n+l}\right) \\
&\forall n
\end{aligned} \tag{21}$$

Combine the 4.2.1 and 4.2.1, we can formulate the first order Markov process:

$$H(\mathcal{X}) = \lim_{n \to \infty} H\left(X_n \mid X_{n-1}\right) = H\left(X_2 \mid X_1\right) \tag{22}$$

Then, for the max entropy rate stochastic process with the constraints:

$$E\left[X_i X_{i+k}\right] = \alpha_k \quad \text{for } k = 0, 1 \ldots p \quad \forall i \tag{23}$$

We can formulate the whole process as :

$$X_i = -\sum_{i=1}^{p} a_k X_{i-k} + Z_i \quad Z_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \tag{24}$$

4.2.2 Proof

We can use the chain rule to get the upper bound of $H\left(X_1, \ldots, X_n\right)$ to prove the process formula:

$$\begin{aligned} H\left(X_1, \ldots, X_n\right) &\leq H\left(Z_1, \ldots, Z_n\right) \\ &= H\left(Z_1, \ldots, Z_p\right) + \sum_{i=p+1}^{n} H\left(Z_i | Z_{i-1}, \ldots, Z_1\right) \text{ (chain rule)} \\ &\leq H\left(Z_1, \ldots, Z_p\right) + \sum_{i=p+1}^{n} H\left(Z_i | Z_{i-1}, \ldots, Z_{i-p}\right) \text{ (conditioning doesn't increase entropy)} \\ &= H\left(Z'_1, \ldots, Z'_p\right) + \sum_{i=p+1}^{n} H\left(Z'_i | Z'_{i-1}, \ldots, Z'_{i-p}\right) \\ &= H\left(Z'_1, \ldots, Z'_n\right) \end{aligned} \tag{25}$$

We can also check the form satisfied the constraint by using Yule-Walker equations.

4.2.3 Application of Maximum Entropy on Markov Model(MEMM)

Just as the maximum entropy model is the application of the maximum entropy principle in a single variable. The MEMM [4] is the application of maximum entropy theory on Markov Chain related to arbitrary dependent variables' sequence.

We can also modify the formula to conditioning one, where x is the observed value, and y is the label of the sequence. In other words, the X is the observed distribution, the Y is the true distribution. We need to inference the true distribution from what we observed.

$$P\left(y_1, \ldots, y_n | x_1, \ldots, x_n\right) = \prod_{t=1}^{n} P\left(y_t | y_{t-1}, x_t\right) \tag{26}$$

$$P\left(y | y', x\right) = P_{y'}(y|x) = \frac{1}{Z\left(x, y'\right)} \exp\left(\sum_a \lambda_a f_a(x, y)\right) \tag{27}$$

Here, the f is the feature function, and the λ is the feature weight. The f can be seen as the connection measure between input and output. Those can give the more entropy will give more weight. Similarly, such a process can be seen as a feature selection.

5 Experiments of the Maximum Entropy Model

5.1 Task

For the Maximum Entropy Model, we want to find a suitable input-output transformation function. For the x as the input data or feature. y as the label for optimization. Specifically, We do the sentiment classification based on Maximum Entropy Model Classifier. Thus the x is the text feature, and y is the classification label.

5.2 Dataset

We use the nltk movie reviews dataset, we changed the training samples' number to analyze the learning ability of Maximum Entropy on Model.

5.3 Optimization

The objective function is without an analytical solution, however, the objective function is always convex function. Hence, we can use convex optimization methods to get the final solution.

5.4 Results

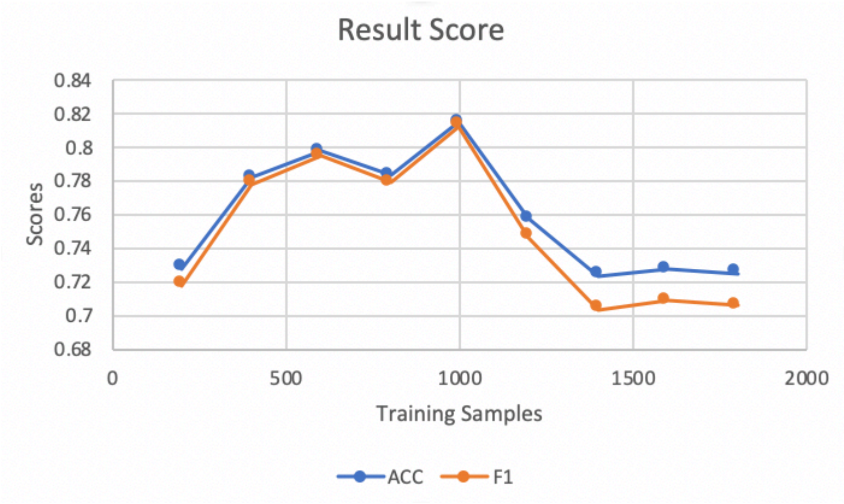

Figure 1: The Maximum Entropy Model Results

5.5 Analysis

5.5.1 Pros

(i) Only needs to concentrate on selecting features, without spending effort consider how to use these features.

(ii) Generally does not require the independent assumptions often used in other methods of modeling.

(iii) Slippage can be considered through feature selection, without the need to consider it separately using conventional smoothing algorithms.

(iv) The contribution of the probability distribution is determined by the parameters, which can be obtained by iterative training by a certain algorithm.

(v) The constraints can be set flexibly, and the degree of constraints can adjust the model's adaptability to unknown data and the degree of fitting to known data

5.5.2 Cons

The relationship between the number of constraint functions and the number of samples leads to high time complexity.

6 Conclusion

In this report, we demonstrate the Maximum Entropy Principle and analyzed its various forms in discrete, continuous variable and variable sequence on Markov Chains. Furthermore, we give a detailed formal definition and formula derivation in its application-Maximum Entropy Model and Maximum Entropy Markov Model. Finally, we implemented a test case of Maximum Entropy Model on the movie review dataset, analyzed the result and learning ability of Maximum Entropy Model. In addition, we provide the pros and cons of the MEM algorithm.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[2] Wikipedia. Principle of maximum entropy — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Principle%20of%20maximum%20entropy&oldid=943854349`, 2020.

[3] David J. C. MacKay. *Information Theory, Inference Learning Algorithms*. Cambridge University Press, USA, 2002.

[4] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598, 2000.