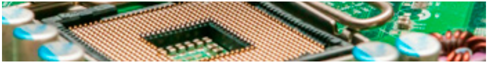
Máster Universitario en Ingeniería Informática

Sistemas Inteligentes

Unit 6. Maximum Entropy Models - Theory

2022/2023



Outline

1. Introduction to Maximum Entropy

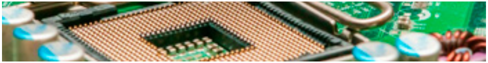
2. Probability estimation with Maximum Entropy

3. Basic concepts

4. Maximum Entropy principle

5. IIS Algorithm

6. References



Outline

1. Introduction to Maximum Entropy
2. Probability estimation with Maximum Entropy
3. Basic concepts
4. Maximum Entropy principle
5. IIS Algorithm
6. References

Entropy definitions

Entropy of a discrete variable $X \sim p(x)$

$$H(X) = -\sum_{x} p(x) \; \log p(x)$$

Also written as $H(p)$

If the log is to the base 2 then the entropy is measured in bits

Interpretation 1: average number of bits needed to describe X
Interpretation 2: *uncertainty* about the outcome of a stochatic variable X

Properties:

- Entropy is a concave function
- Entropy is maximum when p is uniform $\rightarrow$ maximun *uncertainty*

Introduction to Maximum Entropy

Example [Cover 1991]: Consider a dice with 8 faces (with numbers from 0 to 7), all of them with the same probability

A 3-bit string is necessary as label for each face

The entropy of this random variable is:

$$H(X) = -\sum_{i=1}^{8} p(i) \log p(i) = -\sum_{i=1}^{8} \frac{1}{8} \log \frac{1}{8} = \log 8 = 3 \text{ bits}$$

Intuition: let us supose that we send the winner face to a friend; the average number of bits we need to send to the friend to identify the face is 3 bits

Let us supose that:

x	0	1	2	3	4	5	6	7
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$

Then, the entropy is: $H(X) = 2$ bits

Intuition: the most probable outcomes need less bits to be coded

Consider the following coding for the faces:

$\{0, 10, 110, 1110, 111100, 111101, 111110, 111111\}$

$$\frac{1}{2} + 2\,\frac{1}{4} + 3\,\frac{1}{8} + 4\,\frac{1}{16} + 4 \cdot 6\,\frac{1}{64} = 2$$

Joint entropy of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ [Cover 1991]:

$$H(X, Y) = -\sum_x \sum_y p(x, y) \; \log p(x, y)$$

Expectation of $g(X)$ when $X \sim p(x)$

$$E[g(X)] = \sum_x g(x)p(x)$$

Note that $H(X) = E\left[\frac{1}{\log p(X)}\right]$

Conditional entropy of a pair of discrete random variables (X, Y):

$$
\begin{aligned}
H(Y|X) &= \sum_{x} p(x)\; H(Y|X=x) \\
&= -\sum_{x} p(x) \sum_{y} p(y|x)\;\log p(y|x) \\
&= -\sum_{x} \sum_{y} p(x,y)\;\log p(y|x)
\end{aligned}
$$

Theorem (Chain rule) [Cover 1991]

$$H(X,Y) = H(X) + H(Y|X)$$

Example [Cover 1991]: Let (X, Y) have the following joint distribution:

$Y \diagdown X$	1	2	3	4	$\sum$
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
$\sum$	1/2	1/4	1/8	1/8	1

Marginal of X: $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ Marginal of Y: $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

$$
\begin{aligned}
H(X|Y) &= \sum_{i=1}^{4} p(Y=i)\; H(X|Y=i) \\
&= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\
&= \frac{11}{8} = 1.375 \quad \text{bits}
\end{aligned}
$$

Outline

Probability estimation with Maximum Entropy

- **Problem**: estimate $p(y|x)$ where $x = (x_1, \ldots, x_D)$ is D-tuple of discrete (sometimes categoric) observations

$$p(y|x_1, \ldots, x_D)$$

- **Goal**: estimate p given data that follows an empirical distribucion $\tilde{p}$

Probability estimation with Maximum Entropy

Maximum Entropy solution:

- Choose p with maximum entropy (or “uncertainty”) subject to some constraints (given by $\tilde{p}$)
- **Entropy** is a mathematical measure of the uniformity (uncertainty) of a distribution $p(x, y)$

$$H(p) = -\sum_{x,y} p(x, y) \log p(x, y)$$

- For a conditional distribution $p(y|x)$ its conditional entropy is:

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

- Final form is a log-linear combination

Pending problems:

- Selection of features
- Smoothing: no closed-form solutions for optimal parameters

Probability estimation with Maximum Entropy

Overview of the process:

- Collect (x, y) pairs from training data:
  - y: thing to be predicted
  - x: the context
- Learn the probability p of each (x, y)
- Maximum Entropy strategy:

  Model all that is known and assume nothing about what is unknown
  - to satisfy a set of constraints
  - to assume the most “uniform” distribution

Example 1: MT

"We wish to model an expert translator's decisions concerning the proper French rendering of the English word **in**"

Let p our model of the translator's decisions and let (x, y) a set of samples: $\{(\textit{in}, \textit{dans}), (\textit{in}, \textit{en}), \ldots\}$

"Suppose that the expert translator always chooses among the following five French phrases: $\{\textit{dans}, \textit{en}, \textit{à}, \textit{au cours de}, \textit{pendant}\}$"

$$p(\textit{dans}) + p(\textit{en}) + p(\textit{à}\ ) + p(\textit{au cours de}) + p(\textit{pendant}) \quad = \quad 1$$

Probability estimation with Maximum Entropy

"Suppose we notice that the expert chose either *dans* or *en* 30% of the time"

$$\begin{aligned} p(\textit{dans}) + p(\textit{en}) &= 3/10 \\ p(\textit{dans}) + p(\textit{en}) + p(\textit{à}) + p(\textit{au cours de}) + p(\textit{pendant}) &= 1 \end{aligned}$$

"...in half the cases, the expert chose either *dans* or *à*"

$$\begin{aligned} p(\textit{dans}) + p(\textit{en}) &= 3/10 \\ p(\textit{dans}) + p(\textit{en}) + p(\textit{à}) + p(\textit{au cours de}) + p(\textit{pendant}) &= 1 \\ p(\textit{dans}) + p(\textit{à}) &= 1/2 \end{aligned}$$

Example 2: POS tagging

- Let say we have the following event space

NN	NNS	NNP	NNPS	VBZ	VBD

- Empirical data

3	5	11	13	3	1

- Maximize the entropy

$1/e$	$1/e$	$1/e$	$1/e$	$1/e$	$1/e$

- $E[\mathrm{NN}, \mathrm{NNS}, \mathrm{NNP}, \mathrm{NNPS}, V\mathrm{BZ}, \mathrm{VBD}] = 1$

$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

- N* are more common than V*. Add feature $f_N = \{\mathrm{NN}, \mathrm{NNS}, \mathrm{NNP}, \mathrm{NNPS}\}$ with $E[f_N] = 32/36$

NN	NNS	NNP	NNPS	VBZ	VBD
$8/36$	$8/36$	$8/36$	$8/36$	$2/36$	$2/36$

- Proper nouns are more frequent than common nouns. Add feature $f_P = \{\mathrm{NNP}, \mathrm{NNPS}\}$ with $E[f_P] = 24/36$

NN	NNS	NNP	NNPS	VBZ	VBD
$4/36$	$4/36$	$12/36$	$12/36$	$2/36$	$2/36$

Outline

Basic concepts

Problem: Construct a stochastic model that accurately represents the behaviour of a random process: $p(y|x)$

Training data

Given a training sample (x, y), its empirical probability distribution $\widetilde{p}$ is defined by:

$$\widetilde{p}(x, y) \equiv \frac{1}{N} \times \text{number of times that } (x, y) \text{ occurs in the sample}$$

Features

If *April* is the word following *in*, then the translation of *in* is *en* with frequency $9/10$.

$$f(x, y) = \begin{cases} 1 & \text{if } y = \textit{en} \text{ and } \textit{April} \text{ follows } \textit{in} \\ 0 & \text{otherwise} \end{cases}$$

Constraints

Expected value of f with respect to $\widetilde{p}(x, y)$:

$$\widetilde{p}(f) \equiv \sum_{x,y} \widetilde{p}(x, y) f(x, y)$$

Expected value of f with respect to the model $p(y|x)$:

$$p(f) \equiv \sum_{x,y} \widetilde{p}(x) p(y|x) f(x, y) = \sum_{x,y} \widetilde{p}(x, y) f(x, y)$$

Outline

1. Introduction to Maximum Entropy
2. Probability estimation with Maximum Entropy
3. Basic concepts
4. Maximum Entropy principle
5. IIS Algorithm
6. References

Maximum Entropy principle

Given n features f_i, we would like p to lie in the subset $\mathcal{C}$ of $\mathcal{P}$ defined by

$$\mathcal{C} \equiv \{p \in \mathcal{P} \mid p(f_i) = \widetilde{p}(f_i) \text{ for } i \in \{1, 2, \ldots, k\}\}$$

Conditional entropy

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

To select a model from a set $\mathcal{C}$ of allowed probability distributions, choose the model $p_* \in \mathcal{C}$ with maximum entropy $H(p)$:

$$p_* = \arg\max_{p \in \mathcal{C}} H(p)$$

Solution to the primal problem:

http://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/node7.html#SECTION00024000000000000000

The Maximum Entropy solution p_* has the form:

$$\begin{aligned} p(y|x) &= \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x,y)\right) \\ Z(x) &= \sum_{y} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x,y)\right) \end{aligned}$$

where k is the number of features

Let Λ^* be

$$\Lambda^* = \arg\max_{\Lambda} \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x,y)\right)$$

Then $p_{\Lambda^*} = p_*$

Outline

Solution to the dual problem

http://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/node10.html#SECTION00030000000000000000

http://luthuli.cs.uiuc.edu/~daf/courses/optimization/papers/berger-iis.pdf

1. Start with some (arbitrary) value for each λ_i
2. Repeat until convergence:
   (a) Solve $\quad \dfrac{\partial \mathcal{B}(\delta)}{\partial \delta_i} = 0 \quad \text{for} \quad \delta_i$
   (b) Set $\lambda_i = \lambda_i + \delta_i$

where:

$$\frac{\partial \mathcal{B}(\delta)}{\partial \delta_i} = \overbrace{\sum_{x,y} \widetilde{p}(x,y) f_i(x,y)}^{\widetilde{p}(f_i)} - \overbrace{\sum_{x,y} \widetilde{p}(x) p_\lambda(y|x) f_i(x,y)}^{p_\lambda(f_i)} e^{\delta_i f^{\#}(x,y)}$$

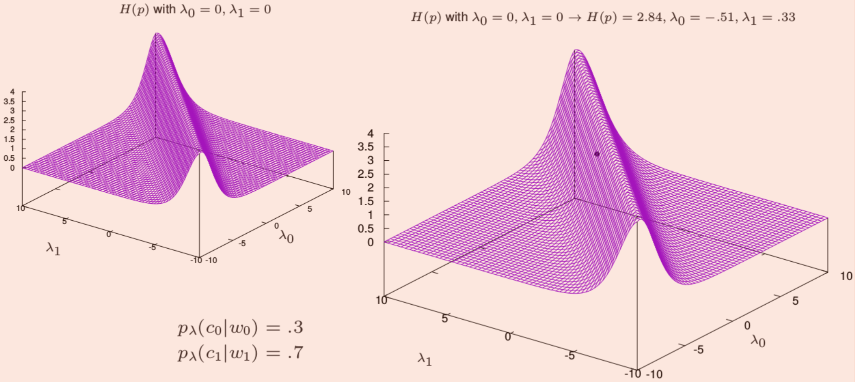
If $f^{\#}(x,y) = M$ then

$$\delta_i = \frac{1}{M} \log \frac{\widetilde{p}(f_i)}{p_\lambda(f_i)}$$

Example to illustrate the optimization problem (just two features)

$$C = \{(w_0,c_0),(w_0,c_0),(w_0,c_0),(w_1,c_1),(w_1,c_1),(w_1,c_1),(w_1,c_1),(w_1,c_1),(w_1,c_1),(w_1,c_1)\}$$

$$\lambda_0 = f(w_0,c_0) \quad \lambda_1 = f(w_1,c_1)$$



$p_\lambda(c_0|w_0) = .3$

$p_\lambda(c_1|w_1) = .7$

Outline

References

- A.L. Berger, V.J. Della Pietra, S. Della Pietra. *A Maximum Entropy Approach to Natural Language Processing.* Computational Linguistics, 22(1):39-71 , 1996.
- T.M. Cover, J.A. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.
- R. Malouf. *A comparison of algorithms for maximum entropy parameter estimation*, COLING, 1-7, 2002.
- A. Ratnaparkhi. *Learning to Parse Natural Language with Maximum Entropy Models.* Machine Learning, 34, 151-175, 1999.