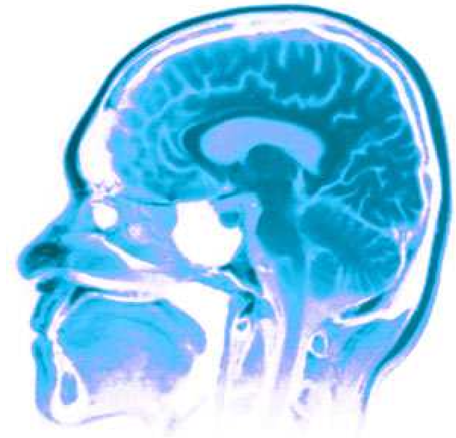
CPSC340



Entropy and maximum likelihood

Nando de Freitas
September, 2012
University of British Columbia

Outline of the lecture

This lecture introduces to our first strategy for learning: **Maximum Likelihood**. The goal is for you to learn:

- ❑ Definition of the maximum likelihood learning strategy.
- ❑ How to apply maximum likelihood to Bernoulli r.v.s.
- ❑ Understand the concepts of **information** and **entropy**.
- ❑ Derive the connection between maximum likelihood and differential entropy.
- ❑ Understand maximum likelihood as a contrasting principle (the world vs. the the hallucinations of the mind).

Frequentist learning

Frequentist learning *assumes that there exists a true model, say with parameters* $\boxed{\theta_o}$.

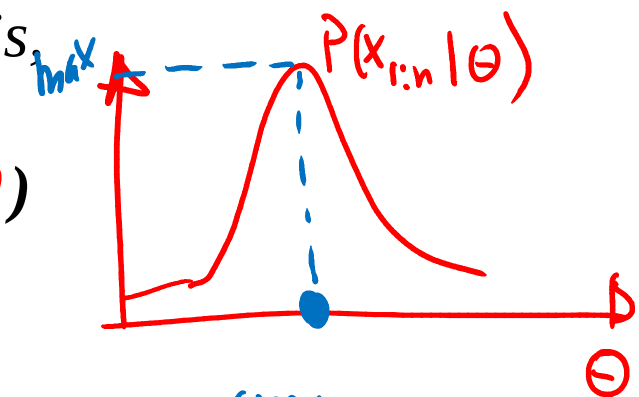
The estimate (learned value) will be denoted $\boxed{\hat{\theta}}$.

Given $\mathbf{n}$ *data,* $x_{1:n} = \{x_1, x_2, \ldots, x_n\}$*, we choose the value of* θ *that has more probability of generating the data. That is,*

$$\hat{\theta} = \arg\max_{\theta} \; p(x_{1:n} \mid \theta)$$

argument

maximizes



arg max

$= \hat{\Theta}$

$\theta = P(x=1)$

$1-\theta = P(x=0)$

Frequentist learning

$n=6$

<u>Example:</u> *Suppose we observe the data,* $\mathbf{x}_{1:n} = \{\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}\}$*, where each* $\mathbf{x}_i$ *comes from the same Bernoulli distribution (i.e. it is independent and identically distributed (**iid**)). What is a good guess of* θ*?*

$$P(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i} \qquad x_i \in \{0, 1\}$$

$$= \begin{cases} \theta & x_i = 1 \\ 1-\theta & x_i = 0 \end{cases}$$

$$\hat{\theta}_{ML} = \frac{\#1\text{'s}}{\#\text{flips}} = 1$$

$\hat{\theta}_1 = 0.99$ ✓ $\qquad P(x=1|\hat{\theta}_1) = 0.99$

$\hat{\theta}_2 = 0.5 \qquad P(x=1|\hat{\theta}_2) = 0.5$

Maximum Likelihood procedure

Step 1: *Given* $\mathbf{n}$ *data,* $x_{1:n} = \{x_1, x_2, \ldots, x_n\}$*, write down the expression for the joint distribution of the data:*

$$\log(AB) = \log A + \log B$$

$$p(x_{1:n} \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta)$$

Step 2: *Compute the log-likelihood.*

$$\mathcal{L}(\theta) = \log P(x_{1:n} \mid \theta) = \sum_{i=1}^{n} \log P(x_i \mid \theta)$$

Step 3: *Differentiate and equate to zero to find the estimate of* θ.

Bernoulli MLE

$m=2$ $\quad x=(1\ 1\ 0)$ $\quad n=3$

$$\prod_i A^{x_i} = A^1 A^1 A^0 = A^2$$

Step 1: *Write down the specific distribution for each datum (Bernoulli in our case):*

$$\boxed{p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}}$$

$$\prod_i B^{(1-x_i)} = B^0 B^0 B^1 = B^{3-2}$$

$$p(x_{1:n} \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

m = # of 1's

$$= \theta^{m}(1-\theta)^{n-m}$$

$n-m$ = # zeros

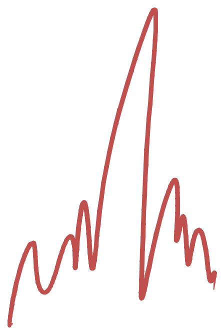
n = # coin flips

Step 2: *Compute the log-likelihood.*

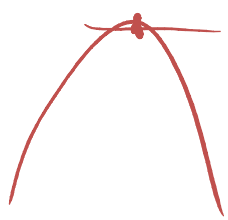
$$\mathcal{L}(\theta) = \log P(x_{1:n} \mid \theta) = m \log \theta + (n-m)\log(1-\theta)$$

Bernoulli MLE

Step 3: *Differentiate and equate to zero to find the estimate of θ:*



$$\frac{d\mathcal{L}(\theta)}{d\theta} = \frac{d}{d\theta}\Big(m \log\theta + (n-m)\log(1-\theta)\Big)$$



$$= \frac{m}{\theta} + (n-m)(-1)\frac{1}{1-\theta}$$

$$= \frac{(1-\theta)m + (m-n)\theta}{\theta(1-\theta)} = 0$$

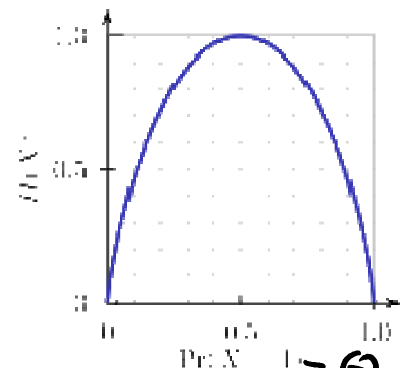
$$m - \cancel{\theta m} + \cancel{\theta m} - \theta n = 0 \Rightarrow \theta = m/n$$

Entropy

*In information theory, entropy **H** is a measure of the uncertainty associated with a random variable. It is defined as:*

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

<u>*Example:*</u> *For a Bernoulli variable **X**, the entropy is:*



$= \theta$

Wikipedia

$$H = -\sum_{x=0}^{1} \theta^{x}(1-\theta)^{1-x} \log\left[\theta^{x}(1-\theta)^{1-x}\right]$$

$$= -\theta \log\theta - (1-\theta)\log(1-\theta)$$

MLE - advanced

We begin with an example. Suppose you observe the binary sequence $X_{1:4} = \{1110\}$

Suppose too that such data was produced by a Bernoulli process with $\theta_0 = 0.9$.

That is,

$$P(x_i|\theta_0) = (0.9)^{x_i} (0.1)^{1-x_i}$$

$$P(X_{1:4}|\theta_0) = \theta_0^3 (1-\theta_0)^1 = (0.9)^3 (0.1)^1 = 0.0729$$

Assume we don't know θ_0. Can we use $X_{1:4}$ to guess what θ_0 was?

MLE - advanced

The maximum likelihood approach to this problem is to find the θ that maximises $P(x_{1:4} | \theta)$. We call such θ: $\hat{\theta}_{ML}$. In math:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ P(x_{1:4} | \theta)$$

Now, we know that $\hat{\theta}_{ML} = \frac{\#\ 1\text{'s}}{\#\ \text{flips}} = \frac{3}{4} = 0.75$

So $\hat{\theta}_{ML} = 0.75$ and the truth is $\theta_0 = 0.9$

MLE - advanced

If we knew θ_0, we would conclude that the error is:

$$|\theta_0 - \hat{\theta}_{ML}| = 0.9 - 0.75 = 0.15$$

However, we don't know θ_0.

Assume instead that we can use our model to hallucinate data $\tilde{x}_{1:4}$

MLE - advanced

We hallucinate data as follows:

u = a uniform random number in $[0, 1]$

If $u < \hat{\Theta}_{ML} = 0.75$

Set $\tilde{x}_i = 1$

Else

set $\tilde{x}_i = 0$

For short, we say that $\tilde{x}_i \sim p(\tilde{x}_i | \hat{\Theta}_{ML})$.

Suppose we do this 4 times and produce $\tilde{x}_{1:4} = \{0\ 1\ 1\ 1\}$

MLE - advanced

If we compare $x_{1:4} = \{1110\}$ and $\tilde{x}_{1:4} = \{0111\}$ we see that they are different. However, they both have similar statistics. e.g. they have the same number of 1's.

If the hallucinations $\tilde{X}$ and the data X have the same statistics, we expect $\hat{\theta}_{ML} \approx \theta_0$.

MLE - advanced

That is, we can't compare Θ to Θ_0, but we can compare x to $\tilde{x}$.

Incidentally had we chosen $\Theta = 0.2$, then the sequence might be $\tilde{x}_{1:4} = \{0000\}$, wich seems worth than the sequence produced with $\Theta = \hat{\Theta}_{ML} = 0.75$

Why is $\hat{\Theta}_{ML}$ so good?

MLE - advanced

The next derivation shows that $\hat{\Theta}_{ML}$ is good because it tries to produce a sequence $\tilde{x}$ that has the same information as x.

First,

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} P(x_{1:N} | \Theta)$$

$$= \arg\max_{\Theta} \prod_{i=1}^{N} P(x_i | \Theta)$$

MLE - advanced

But since log() is monotonically increasing:

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \sum_{i=1}^{N} \log P(x_i | \Theta)$$

Moreover

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \left[\sum_{i=1}^{N} \log P(x_i|\Theta) - \underbrace{\sum_{i=1}^{N} \log P(x_i|\Theta_0)}_{\text{const}} \right]$$

since subtracting a constant doesn't change the location of the maximum

MLE - advanced

Since $\arg\max_{\Theta} f(\Theta) = \arg\min_{\Theta}[-f(\Theta)]$, we have

$$\hat{\Theta}_{ML} = \arg\min_{\Theta} \Big\{ \underbrace{\frac{1}{N}\sum_{i=1}^{N} \log P(x_i|\Theta_0)}_{\text{the world}} - \underbrace{\frac{1}{N}\sum_{i=1}^{N} \log P(x_i|\Theta)}_{\text{our model}} \Big\}$$

So $\hat{\Theta}_{ML}$ is the Θ that minimizes the difference between the true average log probability and the average log probability of our model.

MLE - advanced

As $N \to \infty$, the averages become expectations, and

$$\hat{\Theta}_{ML} = \arg\min_{\Theta} \left\{ \overbrace{\int \log P(x|\Theta_0)\, P(x|\Theta_0)\, dx}^{\text{information} \,=\, -\text{Entropy}} - \int \log P(x|\Theta)\, P(x|\Theta_0)\, dx \right\}$$

But this is more advanced.

If you got the derivation up to the previous page, that is all that matters for this course ∎

Next lecture

In the next lecture, we introduce Bayesian learning.