Maximum Likelihood-Maximum Entropy Duality and its Application

Pushpak Bhattacharyya

ACM Summer School

IIT Kharagpur

June 1, 2017

Origin of the duality question

- Students of AI, ML, NLP, CV, Speech meet a few celebrated techniques
  - HMM- Hidden Markov Model
  - MEMM- Maximum Entropy Markov Model
  - CRF- Conditional Random Field
  - Also SVMs, NNs, Deep Learning
- The question they ask: when to choose which algorithm
- Other than HMM all are discriminative techniques
- HMM takes the Maximum Likelihood path
- MEMM and CRF, the maximum entropy path
- Do these 2 paths lead to the same destination, albeit with different amounts of effort?

Xerox Research Innovation Challenge (2015)

- *Patients admitted to hospitals are susceptible to many complications like various infections and acute conditions (like stroke, sepsis…). The aim of this challenge is to analyze past data of patients and design novel machine learning models and algorithms that can predict these complications before they occur. These models can be used to provide alerts to clinical staff and identify high—risk patients in hospitals in order to provide better care and save patients' lives.*

Formulation

- Rule based or Probability based
- Conditions like stroke, sepsis etc. are given Class Labels: C_1, C_2, C_3…
- Patients’ past data: feature engineering
  - Lab observations
  - Native data like age, gender, profession…
- Features: f_1, f_2, f_3…

Rule based classification

- If
  - f_i has value v_{ij} (for various *i* and *j*),
- then
  - class decision is C_k

- (hypothetical naïve example) If *pressure* continues above 170 systolic and 120 diastolic and there is persistent headache, there is high chance of *stroke*

Probabilistic classification

- $C^* = \mathrm{argmax}_c(P(C|<f_1,f_2,f_3...>))$
- Where C* is the winning class
- One can apply Bayes theorem and independence assumption to get
- $C^* = \mathrm{argmax}_c(P(C).P(<f_1,f_2,f_3...>|C))$

  $= \mathrm{argmax}_c(P(C).P(f_1|C).\ P(f_2|C).\ P(f_3|C)...)$

Count based probability calculation: MLE based

$$P(f_i \mid C) = \frac{\#< f_i, C)}{\sum_j \#< f_j, C >}$$

Main message (1/3)

- A large no. of problems in AI-ML-NLP need estimating a probability distribution
- It is known that the FORM of the distribution is very very difficult to learn!
- However, given the form of the distribution, it is possible to estimate the parameters.
- Sometimes there are hidden variables which enforce using Expectation Maximization.

Main Message (2/3)

- When we use EM, the expectation step updates values of hidden variables
- The maximization step updates values of parameters
- For example, estimating probabilities of Heads of TWO coins from observations of outcomes of their tosses, when any of the coins can be picked randomly

Main Message (3/3)

- Now, there are ONLY two significant ways to estimate the parameters
- First: by maximizing the likelihood of the observations, aka, DATA
- Second: by maximizing the entropy of the underlying probability distribution

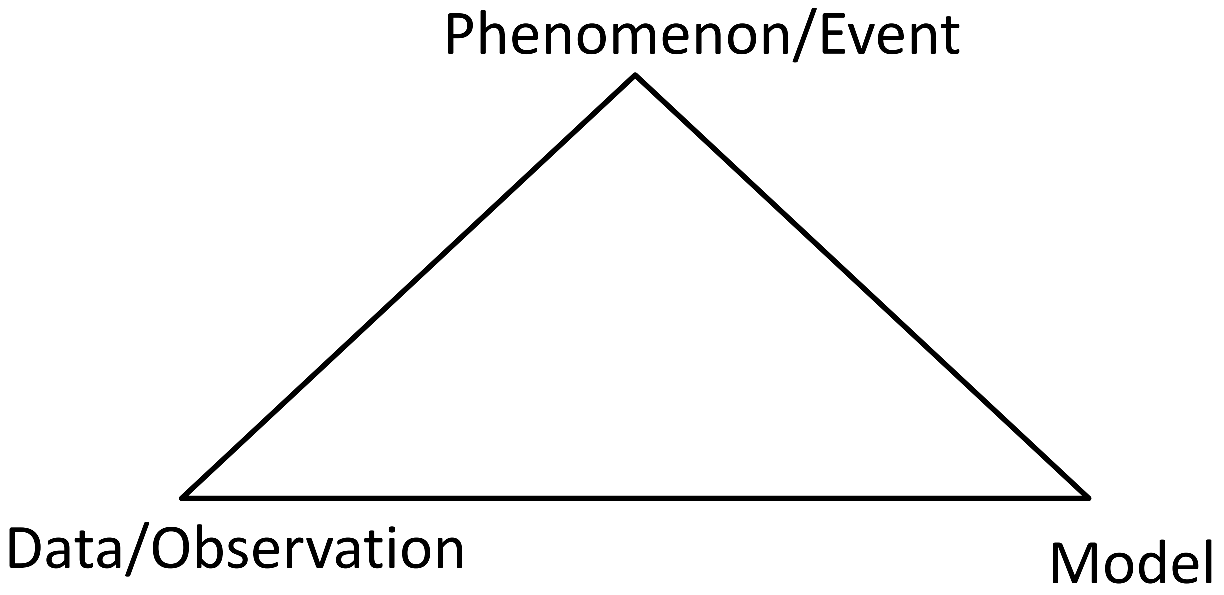
Probability Distributions: are they really needed?

- Probability is needed
- Partial information is the norm, and not an exception
- Example: placing useful labels on sequences
- E.g., putting parts of speech tags on sentences
  - *Flying (NN/JJ) can (NN/VX) be (VB) dangerous (JJ)*
- The POS uncertainties are inevitable in this case
- Only more context can resolve completely
- But one CAN say which is more probable: NN/JJ from the appearances of such sentences in the corpus
- NN/VX uncertainty is almost COMPLETELY resolvable; since ‘be’ follows ‘can’

Probability is needed and is useful

- Instead of throwing up one’s arms and giving up, one can go to the next stage of processing:
- In this case, the next stages of processing are parsing, semantic role determination, sense disambiguation and eventually meaning.
- So, take the labels WITH probability values and launch the next stage of processing.
- Since the input has associated probability values, output too will have probabilities
- Thus parsing will be done with trees ranked with probabilities

Essence of AI



- Phenomenon/Event could be a linguistic process such as POS tagging or sentiment prediction.
- Model uses data in order to “predict” future observations w.r.t. a phenomenon

Notations

X : x_1, x_2, x_3.... x_m (m observations)

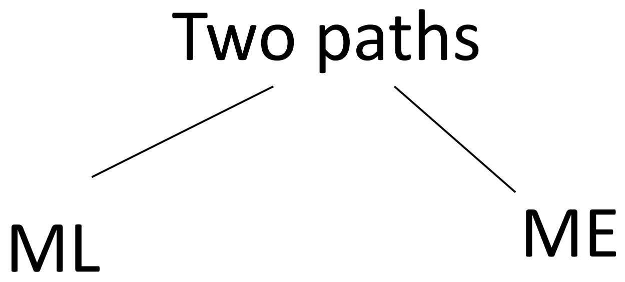
A: Random variable with n possible outcomes such as a_1, a_2, a_3... a_n

e.g. One coin throw : a_1= 0, a_2=1

One dice throw: a_1=1, a_2=2, a_3=3, a_4=4, a_5=5, a_6=6

Goal

- Goal: Estimate $P(a_i) = P_i$



Two paths

ML

ME

Are they equivalent?

Statement of Maximum Likelihood Principle (from Wolfram Mathworld)

- Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes.
- Maximum likelihood, also called the maximum likelihood method, is the procedure of finding the value of one or more parameters for a given statistic which makes the *known* likelihood distribution a maximum.

Statement of Maximum Entropy Principle (Wikipedia)

- Entropy= $-\sum_{i=1,N} p_i \log p_i$

- Subject to precisely stated prior data (such as a proposition that expresses testable information), the probability distribution which best represents the current state of knowledge is the one with largest entropy.

Calculating probability from data

Suppose in X: x_1, x_2, x_3.... x_m (m observations),

a_i occurs $f(a_i) = f_i$ times

e.g. Dice: If outcomes are 1 1 2 3 1 5 3 4 2 1

F(1) = 4, f(2) = 2, f(3) = 2, f(4) = 1, f(5) = 1, f(6)=0 and m = 10

Hence, $P_1 = 4/10$, $P_2=2/10$, $P_3=2/10$, $P_4=1/10$, $P_5=1/10$, $P_6=1/10$

In general, the task is...

Task: Get θ : the probability vector <P(θi)>

MLE

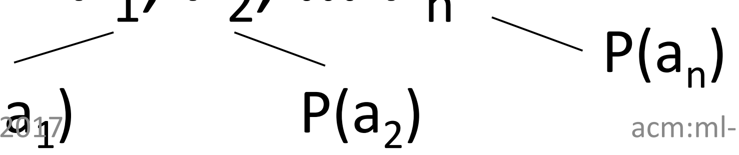
- **MLE**: $\theta^* = \underset{\theta}{\text{argmax}} \Pr (X; \theta)$

With i.i.d. (identical independence) assumption,

$$\theta^* = \underset{\theta}{\text{argmax}} \overset{m}{\underset{i=1}{\pi}} \Pr (X_i; \theta)$$

Where,

$\theta : <P_1, P_2, \ldots P_n>$



$P(a_1)$

$P(a_2)$

$P(a_n)$

What is known about: θ : $<P_1, P_2, \ldots P_n>$

$$\sum_{i=1}^{n} P_i = 1$$

$P_i >= 0$ for all i

Introducing Entropy:

$$H(\theta) = - \underbrace{\sum_{i=1}^{n} P_i \log P_i}_{\text{Entropy of distribution } <P_1, P_2, \ldots P_n>}$$

Some intuition

Example with dice

Outcomes = 1,2,3,4,5,6

$P_1 + P_2+P_3+...P_6 = 1$

Entropy(Dice) = $H(\theta)= - \sum_{i=1}^{6} P_i \ln P_i$

Now, there is a principle called Laplace's Principle of Unbiased reasoning

The best estimate for the dice

$P_1 = P_2 = \ldots P_6 = 1/6$

We will now prove it assuming:

NO KNOWLEDGE about the dice except that it has six outcomes, each with probability >= 0 and $\Sigma P_i = 1$

What does “best” mean?

- “BEST” means most consistent with observations, if observations are given.

OR

- “Best” means that these P_i values should be such that they maximize the entropy.

Optimization formulation

- Max $-\sum_{i=1}^{6} P_i \log P_i$

Subject to: $$\sum_{i=1}^{6} P_i = 1$$

$$P_i >= 0 \text{ for } i = 1 \text{ to } 6$$

Solving the optimization (1/2)

Using Lagrangian multipliers, the optimization can be written as:

$$Q = - \sum_{i=1}^{6} P_i \log P_i - \lambda \left(\sum_{i=1}^{6} P_i - 1 \right) - \sum_{i=1}^{6} \beta_i P_i$$

We will ignore the last term for now

Solving the optimization (2/2)

Differentiating Q w.r.t. P(i), we get

$\delta Q/\delta P(i) = -\log(P_i) - 1 - \lambda$

Equating to zero,
$\log P_i + 1 + \lambda = 0$
$\log P_i = -(1+ \lambda)$

$P_i = e^{-(1+ \lambda)}$

This means that to maximize entropy, every P(i) must be equal.

This shows that $P_1 = P_2 = \ldots P_6$

But,

$P_1 + P_2 + .. + P_6 = 1$

Therefore $P_1 = P_2 = \ldots P_6 = 1/6$

Introducing data in the notion of entropy

Now, we introduce data:

X : x_1, x_2, x_3.... x_m (m observations)

A: a_1, a_2, a_3... a_n (n outcomes)

e.g. For a coin, in absence of data: P(H) = P(T) = 1/2

However, if data X is observed as follows:

Obs-1: H H T H T T H H H (m=10) (n=2)

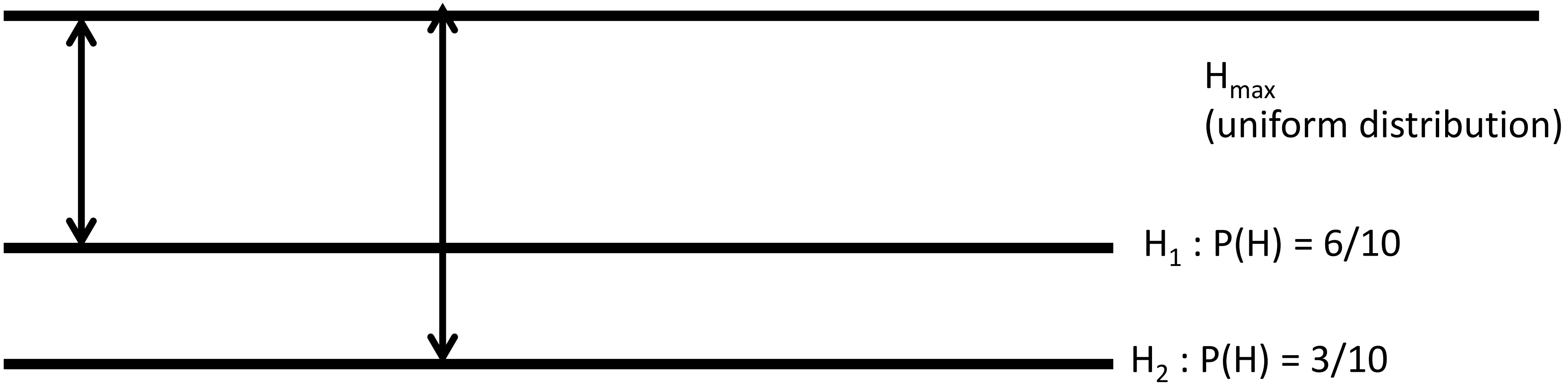
P(H) = 6/10, P(T) = 4/10

Obs-2: T T H T H T H T T T (m=10) (n=2)

P(H) = 3/10, P(T) = 7/10

WHY and Which of these is a valid estimate?

Change in entropy



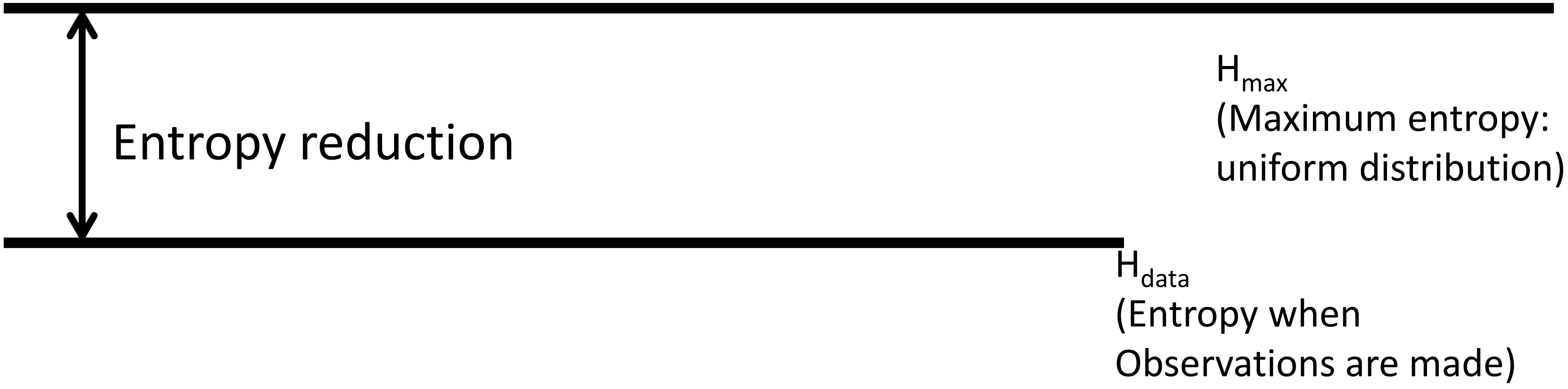
H_{max}
(uniform distribution)

H_1 : P(H) = 6/10

H_2 : P(H) = 3/10

Entropy reduces as data is observed!

Start of Duality


Entropy reduction
H_{max}
(Maximum entropy: uniform distribution)
H_{data}
(Entropy when Observations are made)

- Maximizing entropy in this situation is same as minimizing the `entropy reduction’, i.e.
  - Minimizing “relative entropy”

Recap

Recap (1/2)

- Whatever we can show with MLE, (in most cases), we can through ME as well

- **Laplace's unbiased reasoning principle**

Make only the most limited assumptions.

- We assume data:

X: x_1, x_2.... x_m m: number of observations

Every x_i is the outcome of the values of a random variable

Recap (2/2)

- E.g. x_i = {1, 2, ...6} for dice.

- If no data then by Laplace's unbiased reasoning principle, uniform distribution is the best estimate of the probability of each outcome.

In presence of data

- When we have the data, Pj = ?

- Answer: Pj = f j / m

WHY?

Can we arrive at this value through (a) MLE, or (b) ME ??

Taking the MLE route

MLE: We maximize data likelihood

$P(X; \theta)$ where $\theta = <P_1, P_2, \ldots P_n>$

Under i.i.d.,

$$P(X; \theta) = \mathop{\pi}\limits_{i=1}^{m} P(x_i; \theta)$$

$$= \mathop{\pi}\limits_{i=1}^{m} P_j^{fj}$$

Where, $$\sum_{j=1}^{n} f_j = m$$

e.g. 1 1 2 3 4 5 4

P1 P1 P2 P3 P4 P5 P4

... $P_1^2 \, P_2^1 \, P_3^1 \, P_4^2 \, P_5^1$

Maximization

MLE demands
maximize P(X; θ)

Subject to :

$$\sum_{j=1}^{n} P_j = 1$$

P_j>= 0 for all j

$$\equiv$$

Maximize ln (P(X; θ))

$$\sum_{i=1}^{n} P_j = 1$$

P_j >= 0 for all j

Evaluating the parameter P_j (1/2)

$$Q = \sum_{j=1}^{n} f_j \ln P_j - \lambda(\sum_{j=1}^{n} P_j - 1)$$

Taking derivative w.r.t. P_j

$$\delta Q/\delta P_j = f_j / P_j - \lambda$$

Equating to zero,

$$f_j / P_j - \lambda = 0$$

$$P_j = f_j / \lambda \quad (1)$$

Taking derivative w.r.t. λ

$$\delta Q/\delta P_j = \Sigma \; P_j - 1$$

Equating to zero,

$$\Sigma \; P_j = 1 \qquad (2)$$

Evaluating the parameter P_j (2/2)

From (1) and (2),

$$\sum_{j=1}^{n} f_j / \lambda = 1$$

$$\sum_{j=1}^{n} f_j = \lambda$$

$$\lambda = m$$

$$P_j = f_j / m$$

Proved!

Summary

1) No observation: Pj = 1/n

Entropy
(as seen before)

2) Observation: Pj = fj/m ; $\Sigma\ f_j = m$

MLE
(as shown now)

ME??

Does entropy change

- Before we move on to discussion on ME in case of observed data, we first see how entropy gets affected in presence of data.

- In context of cases given in the previous slide, we wish to see:
  - Is it true that Entropy (Case 2) <= Entropy (Case 1)??
- Lets verify.

Situation 1: No observed data

- Entropy(Case1): $-\sum_{j=1}^{n} P_j \log P_j$

$$= - \sum_{j=1}^{n} (1/n) \log (1/n)$$

$$= \log (1/n) = \log (n)$$

Situation 2: Observed Data (1/2)

- Entropy(Case2): $- \sum_{j=1}^{n} P_j \log P_j$

$$= - \sum_{j=1}^{n} (f_j/m) \log (f_j/m)$$

$$= -1/m \sum_{j=1}^{n} f_j \log (f_j/m)$$

$$= -1/m \sum_{j=1}^{n} f_j [\log (f_j) - \log(m)]$$

Change in entropy

No data, 1/n, H_u

Data, fj/m, H_d

Entropy should reduce as data is observed. To prove.

Recap

Recap

$x_1, x_2, x_3 .... x_N$: Data

$a_1, a_2, a_3 ... a_M$: Outcomes

Goal: Estimate $P(a_i) = P_i$

In absence of any other information (not even data),

$$P_i = 1/M$$

This was obtained using ME

Max. $$-\sum_{i=1}^{m} P_i \log P_i$$

$$\text{s.t.} \sum_{i=1}^{m} P_j = 1$$

When data is observed,

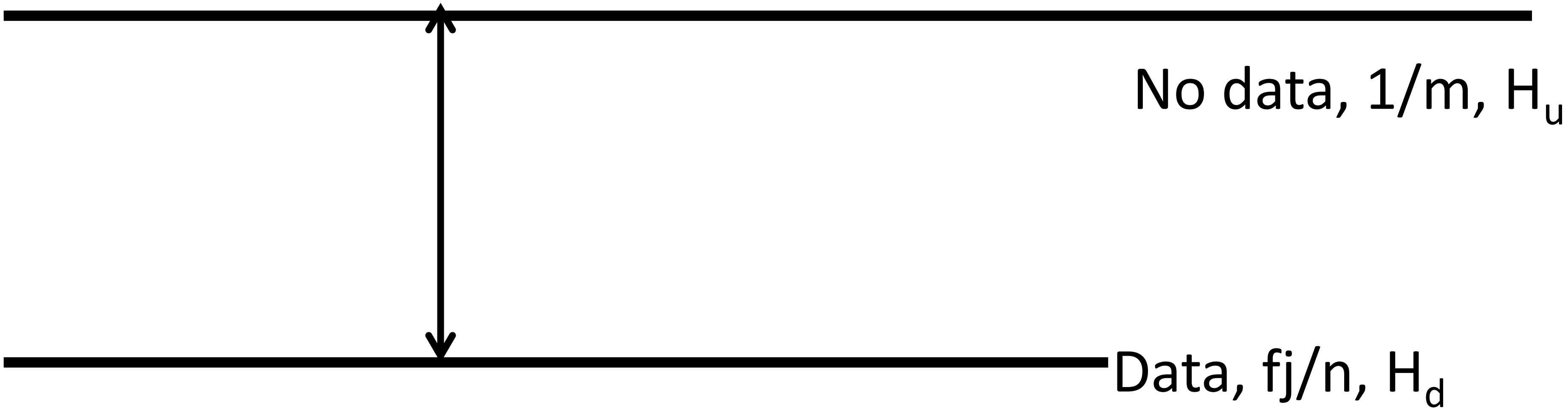
$Pi = f_i/N$ where $f_i = freq(a_i)$

This was obtained using MLE Max.

$$-\sum_{i=1}^{m} f_i \log P_i$$

$$\text{s.t.} \sum_{i=1}^{m} P_i = 1 \quad \& \quad \sum_{i=1}^{m} f_i = N$$

Change in entropy



No data, 1/m, H_u

Data, fj/n, H_d

Entropy changes as data is observed.

Today, we show that the entropy is “reduced”.

i.e. $E_u > E_d$

An intermediate goal

Now, we will show that the entropy is “reduced”.

i.e. $H_u > H_d$

- Two proofs

Proof 1 (1/4)

- Suppose without loss of generality,

P_1 -> $P + \varepsilon$ P_2 -> $P - \varepsilon$

Thus, in case of uniform distribution,
$P_1 = P_2 = \ldots P = 1/M$, and $\varepsilon = 0$

$$H_u = -\sum_{i=1}^{m} P_i \log P_i$$

$$H_d = -(P+\varepsilon)\log(P+\varepsilon) - (P-\varepsilon)\log(P-\varepsilon) - \sum_{i=3}^{m} P_i \log P_i$$

Proof 1 (2/4)

$$H_u – H_d = -P \log P + (P+\varepsilon) \log(P+\varepsilon)$$

$$-P \log P + (P-\varepsilon) \log(P-\varepsilon)$$

$$= P \log ((P+\varepsilon)/P) + \varepsilon (\log P + \log (1+\varepsilon/P))$$

$$+ P \log ((P-\varepsilon)/P) - \varepsilon (\log P + \log (1-\varepsilon/P))$$

$$= P \ [\log (1+\varepsilon/P)+ \log (1-\varepsilon/P)]$$

$$+ \varepsilon [\log (1+\varepsilon/P) - \log (1-\varepsilon/P)]$$

Proof 1 (3/4)

$$H_u – H_d = P\ [\log(1+\varepsilon/P)+ \log(1-\varepsilon/P)]$$

$$+ \varepsilon\ [\log(1+\varepsilon/P) - \log(1-\varepsilon/P)]$$

$$= P\ [\log(1 – \varepsilon^2/P^2)]$$

$$+ \varepsilon\ [\log((1+\varepsilon/P)/(1-\varepsilon/P))]$$

$$= \quad P\ [\log(1 – y^2)]$$

$$+ \varepsilon\ [\log((1+y)/(1-y))]$$

Where, $y= \varepsilon/P$

Proof 1 (4/4)

H_u-H_d = P [log $(1 – y^2)$] + ε [log ((1+y)/(1-y))]

log (1+x) = $x – x^2/2 + x^3/3 – x^4/4+$ ..
log ((1+x)/(1-x)) = $2x +2x^3/3+2x^5/5+2x^7/7+$..

= P [$-y^2 –y^4/2 –y^6/3 – y^8/4$...] + ε[2y +
$2y^3/3+2y^5/5+2y^7/7$...]

= $-Py^2$ [1+ $y^2/2+ y^4/3+ y^6/4$...]
+ Py^2 [2 + $2y^2/3+ 2y^4/5+ 2y^6/7$...]

...substitute ε = Py

When we compare the above statement term by term, (1..2), (1/2.. 2/3), (1/3...2/5).. Etc., we see that the above value is > 0. i.e. $H_u – H_d > 0$. ***i.e. Eu > Ed***

Proof 2 (1/2)

Identity: for $x,y > 0$

$$y - y\log y <= x - y\log x$$

Proof: $\log t <= (t - 1)$

Put $t = (x/y)$. $\log (x/y) <= (x/y) - 1$

i.e. $\log x - \log y <= (x-y)/y$

i.e. $y \log x - y \log y <= x - y$

i.e. $y - y \log y <= x - y \log x$

Proof 2 (2/2)

$p_1, p_2 .... p_m = 1/M$ .... Uniform

$q_1, q_2 ... q_m$ .... Perturbed distribution

We put x = p, y= q in y – ylog y <= x – ylog x

$q_i - q_i \log q_i <= p_i - q_i \log p_i$

take sum over i = 1....m,

$\Sigma q_i - \Sigma q_i \log q_i <= \Sigma p_i - \Sigma q_i \log p_i$

But, $\Sigma q_i = 1$ *&* $\Sigma p_i = 1$

$1 - \Sigma q_i \log q_i <= 1 + \log M \Sigma q_i$

$1 + H_d <= 1 + \log M$

$\cancel{1} + H_d <= \cancel{1} + H_u$

$\mathbf{H_u >= H_d}$

Therefore, established that

- There is a “reduction” in entropy when data is observed.

- We showed two proofs: the first is more intuitive but long; the second is simpler.

Recap

A uniform distribution and any p.d. P

Uniform, 1/n
P

Let us now compute relative entropy / KLD between these two vectors.

Relative entropy between two non-negative vectors

P: $<p_1\ p_2\ \ldots\ p_n>$, Q: $<q_1\ q_2\ \ldots\ q_n>$

KL Divergence or relative entropy between the two vectors is defined as:

$$D(P||Q) = \sum_{i=1}^{n} p_i \log (p_i / q_i) - \sum_{i=1}^{n} p_i + \sum_{i=1}^{n} q_i$$

& $D(P||Q) >= 0$

& $D(P||Q) = 0$ occurs, when $p_i = q_i$

Relative entropy between uniform distribution and any p.d. p

$$D(P||1/n) = \sum_{i=1}^{n} p_i \log (p_i /(1/n)) - \sum_{i=1}^{n} p_i + \sum_{i=1}^{n} (1/n)$$

$$= \Sigma\ p_i \log p_i + \Sigma\ p_i \log n - \cancel{1} + \cancel{1}$$

$$= - H_p + \log n$$

$$\boxed{D(P\ ||\ 1/n\ ) = - H_p + \log n}$$

Relation between entropy of a p.d. and its relative entropy with uniform

$$\boxed{D(P \,||\, 1/n) = -H_p + \log n}$$

$$\boxed{\underset{P}{\operatorname{argmin}}\, D(P \,||\, 1/n) = \underset{P}{\operatorname{argmax}}\, H_p}$$

This shows that, if we maximize entropy P, we minimize relative entropy with respect to uniform distribution.

Conclusion: The probability distribution found by maximizing entropy is the distribution with least KLD (relative entropy) from uniform distribution.

A uniform distribution and any pd with data

Uniform, 1/n

f_1, f_2, ...f_n : F

Now, when we have data (n-ary) given by frequencies (f_1, f_2... f_n) as shown above.

Let us compute relative entropy / KLD between these two vectors.

Frequency distribution F and a p.d P

Suppose F is the frequency of outcomes:

F : f_1, f_2.... f_n m : total no. of observations

Let P be “some” probability distribution

Let us now find the KLD between F and P

Relative entropy between frequency distribution and any p.d. p

$$D(F||P) = \sum_{i=1}^{n} f_i \log (f_i / p_i) - \sum_{i=1}^{n} f_i + \sum_{i=1}^{n} p_i$$

$$= \Sigma\ f_i \log f_i - \Sigma\ \log p_i^{\ fi} - m + 1$$

= constant – log-likelihood - constant

$$\boxed{D(F||P) = \text{constant} – \text{log-likelihood}}$$

Relation between entropy of P and its relative entropy with freq. dist.

$$D(F||P) = \text{constant} - \text{log-likelihood}$$

$$\underset{P}{\operatorname{argmin}}\, D(F||P) = \underset{P}{\operatorname{argmax}}\, LL(P)$$

This shows that, if we maximize log-likelihood of P, we minimize relative entropy with respect to frequency distribution

Conclusion: The probability distribution found by maximizing log-likelihood is the distribution with least KLD (relative entropy) from frequency distribution.

Observations

1) The P found by maximizing entropy is the one with least RE w.r.t. uniform distribution
2) The P found by maximizing likelihood is the one with least RE w.r.t. frequency distribution

Work to be done:

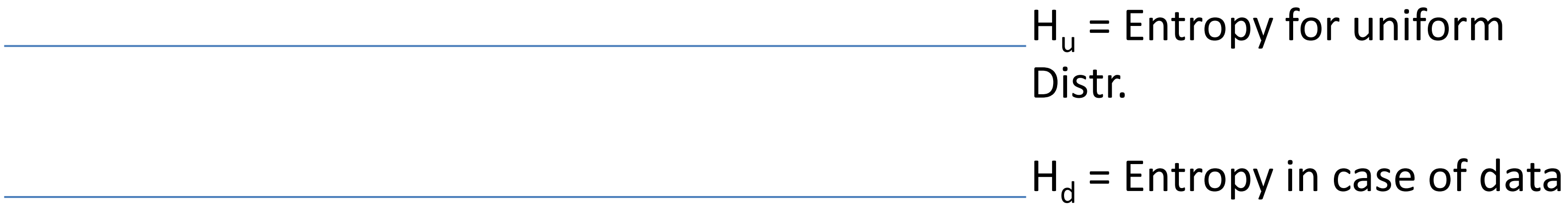
A) Bring in fs into (1) above

B) Bring in 1/n into (2) above

Recap

Recap: Reduction in entropy

When we have data,



H_u = Entropy for uniform Distr.

H_d = Entropy in case of data

We have shown in two ways that: Hu > Hd. i.e. If we perturb a distribution 1/n, 1/n .... To 1/n+k, 1/n-k, 1/n ..... , the entropy decreases

Recap: Relative Entropy

P: $p_1, p_2, p_3, \dots p_n$ $\quad \sum_{i=1}^{n} P_i = 1 \quad \sum_{i=1}^{n} q_i = 1$

Q: $q_1, q_2, q_3, \dots q_n$

$$D(p||q) = \sum_{i=1}^{n} p_i \log (p_{i\,/}\, q_i\ )$$

Recap: Relative entropy wrt uniform distribution

For q: uniform distribution,

$$D(p||q) = \sum_{i=1}^{n} p_i \log (p_{i\,/}\, q_i\ )$$

$$=\sum_{i=1}^{n} p_i \log (p_{i\,/}\, (1/n)\ )$$

$$= \sum_{i=1}^{n} p_i \log p_i - \log (1/n)$$

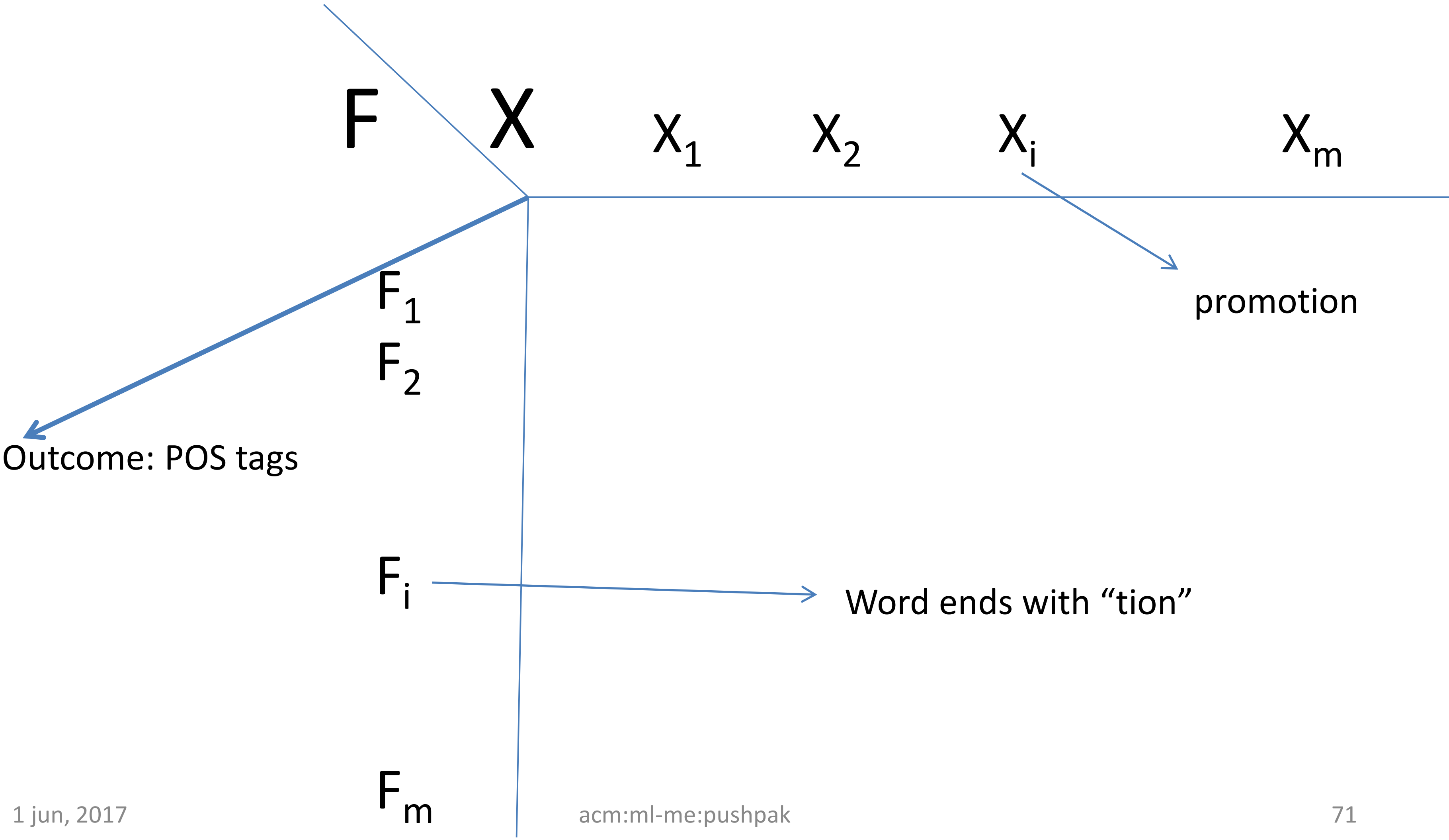
$$= - H_p + \log n = H_u - H_p$$

PIVOT: Relative entropy and absolute entropy difference are same for uniform distribution.

Approach to prove duality

- Theorem: MLE-ME converge when:
  - Distance measure = Relative entropy
  - Distributions belong to exponential family

Data v/s feature matrix



F X

X_1 X_2 X_i X_m

promotion

F_1

F_2

Outcome: POS tags

F_i

Word ends with “tion”

F_m

Exponential & Constrained distributions

$$P_i = c_i \overset{k}{\underset{j=1}{\pi}} \lambda_j^{\,fji} = c_i\, e^{\Sigma\lambda j\ (fji)}$$

Expected value

$$P = \{ P \text{ s.t. } E_{P\,F} = E_{\overline{P}\,F} \}$$

Distribution intended

Expected value s.t. Feature Value is F under distribution P

Distribution Obtained from data

Equivalence between constrained and exponential

Now, let:

P: set of Constrained distributions

Q: set of exponential distributions

And P* = P intersection Q

By ME, we will show that: H(P*) >= H(q) for all q

By MLE, we will show that: LL(P*) >= LL(p) for all p

This is done using concepts like:
(a) Pythagorean distance for relative entropy,
(b) LL in terms of P

Familiar example: POS tagging

- Vocabulary, V : V_1, V_2, V_3 ... $V_{|V|}$
- Tag set, T : T_1, T_2, T_3 ..... $T_{|T|}$

- Call |V|X|T| = A

We wish to estimate

P: P_1 P_2 P_3 ..... P_A

where

P_i = P (V_v, T_t) e.g. P("play", "NN")

Example

- “_“ The_DT mechanisms_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, ”_” said_VBD Mr._NNP Benton_NNP ._.

Argmax computation (1/2)

Best tag sequence

$= T^*$

$= \text{argmax}_T\ P(T|W)$

Dataset and features

- X: (w_1,t_1), (w_2,t_2), (w_3,t_3), ... (w_n, t_n)

- We introduce features, F: F_1, F_2, F_3 ... $F_{|F|}$

Binary

- example:

F_f = 1 if V_v ends with "-tion"

= 0 otherwise

Binding features with V X T

- example: *<promotion, N>*

Features = ends with *–tion*, *has 3 syllables*, etc.

F̃ is a matrix with |F| rows and |V| X |T| columns

P : P_1, P_2, P_3 …. $P_{|V||T|}$

Expected value of a feature

F VXT	$<v_1,t_1>$	$<v_1,t_2>$	.....	$<v_i, t_j>$ ......
F_1				
F_2				
F_3				
.				
.				
.				
.				
$F_{\|F\|}$				

Introducing two distributions

Constrained distribution: (w.r.t training data)

$$R : \{ r(y) : E_r(F) = E_{\tilde{P}} (F) \}$$

where y belongs to V X T

Exponential distribution:

$$S: \{ s(y) : s(y) = k\, e^{\sum_{f=1}^{|F|} \mu_f F_f} \}$$

e.g. s(“promotion”, “NN”)

Pythagorean theorem

- D (r || s) = D (r || p*) + D (p* || s)

p * is an intersection of r and s; r ε R and s ε S, constrained and exponential distributions respectively

Intuition

- The intuition is:
  - Constrained distribution brings in likelihood
  - Exponential distribution is due to entropy

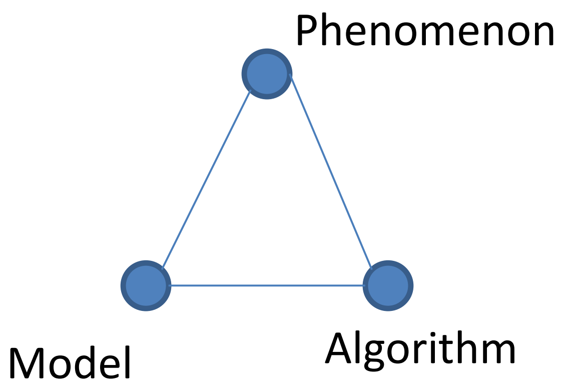
  - Hence, the duality

Recap: POS Tagging (1/2)

- Vocabulary, V : V_1, V_2, V_3 … $V_{|V|}$
- Tag set, T : T_1, T_2, T_3 ….. $T_{|T|}$
- Features, F: F_1, F_2, F_3 … $F_{|F|}$

“Promotion”,	“NN”,	has ‘tion’
V_i	T_j	F_k

Goal:



Recap: POS Tagging (2/2)

- P: P_1 P_2 P_3 ..... $P_{|V| \text{ X } |T|}$

- P(T|W) = P(T,W) / P(W)
- Hence, P("Promotion", "NN") is the kind of probabilities we wish to estimate

- Observations: O: O_1, O_2, O_3... O_m

{o} ε {V} X {T} .... In a tagged corpus

Constrained Distribution

- R: $\{r \mid E_r (F_k) = E_{\tilde{p}} (F_k) \}$ $K = 1 \ldots. |F|$

Where,

$\tilde{P}$ = empirical distribution

$$E_r(F_k) = \sum_{\text{All values } F_k} V(F_k). \Pr (F_k)$$

F_k are such that $V(F_k) = 0$ or 1

$$E_r(F_k) = \Pr(F_k) = \sum_x \Pr (F_k, x)$$

$$= \sum_x \Pr(x) \Pr (F_k|x)$$

$X \varepsilon \{V\} X \{T\}$

Exponential distribution

$$S: \{\ s(y) : s(y) = k\, e^{\sum_{f=1}^{|F|} \mu_f F_f}\ \}$$

Weights of features

- Uniform distribution, U is a member of the exponential family where $\mu_{f = 0}$

Pythagorian theorem

- Let there be a p* that is a member of R ∩ S

- The theorem states that for such a p*

$D(r||s) = D(r||p^*) + D(p^*||s)$

where *r* comes from the family of constrained distributions and *s* comes from the family of exponential distributions

Proof (1/4)

- LHS = D(r||s)

$$= \sum_{x} r(x).\ \log\ (r(x)) - \sum_{x} r(x).\ \log\ (s(x))$$

$$= -\ H(r) - E_r\ \log\ (s(x))$$

Proof (2/4)

- RHS = D(r||p*) + D(p*||s)

$$= \sum_x r(x) \log r(x) - \sum_x r(x) \log (p^*(x))$$

$$+\sum_x p^*(x) \log (p^*(x)) - \sum_x p^*(x) \log (s(x))$$

Proof (3/4)

- Terms 2 and 3 cancel and the last term in LHS will be equal to remaining in RHS. Why?
- For any two distributions r_1 and r_2 from constrained distributions family and any distribution *s* from exponential distribution family

$$\sum_{X} r_1(x) \log s(x) = \sum_{X} r_2(x) \log s(x)$$

Proof (3/4)

- The following important result is used:

$$\sum_{X} r_1(x)\ F_i(x) = \sum_{X} p\sim(x)\ F_i(x) = \sum_{X} r_2(x)\ F_i(x),\ \forall i$$

All these quantities are expectations of feature values

Taking the ME path (1/1)

- With Pythagorian theorem in view, we wish to show that: $p^* = \underset{r}{\text{argmax}}\ (H(r))$

- Consider $D(r||u)$ where u is uniform distribution

- $D(r||u) = D(r||p^*) + D(p^*||u)$

Since $D(.) >= 0$,

$D(r||u) >= D(p^*||u)$

Taking the ME path (2/2)

- $D(r||u) = -H(r) - \sum_{x} r(x). \log (u(x))$

  $= - H(r) + \log (|V| X |T|)$

- $D(p^*||u) = -H(p^*) - \sum_{x} p^*(x). \log (u(x))$

  $= - H(p^*) + \log (|V| X |T|)$

Therefore,

$-H(r) + \log (|V|X|T|) >= -H(p^*) + \log (|V|X|T|)$

$-- H(r) >= -H(p^*)$

$H(r) <= H(p^*)$ ---------------- (A)

Taking the MLE Path (1/3)

- Consider $D(\tilde{p} \,||\, s) = D(\tilde{p} \,||\, p^*) + D(p^* \,||\, s)$

Where $\tilde{p}$ is a constrained distribution

$$D(\tilde{p} \,||\, s) >= D(\tilde{p} \,||\, p^*)$$

$$-H(\tilde{p}) - \sum \tilde{p}(x).\ \log (s(x)) >= -\ H(\tilde{p}) - \sum \tilde{p}(x).\ \log (p^*(x))$$

$$\sum \tilde{p}(x).\ \log (s(x)) <= \sum \tilde{p}(x).\ \log (p^*(x))$$

Taking the MLE path (2/3)

- Suppose $\tilde{p}(x) = fx / |o|$

$$\sum f(x).\ \log(s(x)) <= \sum f(x).\ \log(p^*(x))$$

$$\sum \log(s(x)^{f(x)}) <= \sum \log(p^*(x)^{f(x)})$$

$$LL(s) <= LL(p^*) \text{ ------ } (B)$$

The Duality!

- By (A) and (B),

p* maximizes entropy (From (A))

maximizes log-likelihood (From B)

A significant application

Word alignment as the crux of Statistical Machine Translation

English

(1) three rabbits

a b

(2) rabbits of Grenoble

b c d

French

(1) trois lapins

w x

(2) lapins de Grenoble

x y z

Initial Probabilities: each cell denotes *t(a←→w), t(a←→x) etc.*

	a	b	c	d
w	1/4	1/4	1/4	1/4
x	1/4	1/4	1/4	1/4
y	1/4	1/4	1/4	1/4
z	1/4	1/4	1/4	1/4

“counts”

a b ←→ ***w x***	a	b	c	d
w	1/2	1/2	0	0
x	1/2	1/2	0	0
y	0	0	0	0
z	0	0	0	0

b c d ←→ ***x y z***	a	b	c	d
w	0	0	0	0
x	0	1/3	1/3	1/3
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

Revised probabilities table

	a	b	c	d
w	1/2	1/4	0	0
x	1/2	5/12	1/3	1/3
y	0	1/6	1/3	1/3
z	0	1/6	1/3	1/3

“revised counts”

a b ←→ ***w x***	a	b	c	d
w	1/2	3/8	0	0
x	1/2	5/8	0	0
y	0	0	0	0
z	0	0	0	0

b c d ←→ ***x y z***	a	b	c	d
w	0	0	0	0
x	0	5/9	1/3	1/3
y	0	2/9	1/3	1/3
z	0	2/9	1/3	1/3

Re-Revised probabilities table

	a	b	c	d
w	1/2	3/16	0	0
x	1/2	**85/144**	1/3	1/3
y	0	1/9	1/3	1/3
z	0	1/9	1/3	1/3

Continue until convergence; notice that (b,x) binding gets progressively stronger; b=rabbits, x=lapins

Derivation: Key Notations

English vocabulary : V_E
French vocabulary : V_F
No. of observations / sentence pairs : S
Data D which consists of S observations looks like,

$$e^1{}_1, e^1{}_2, \dots, e^1{}_{l^1} \Leftrightarrow f^1{}_1, f^1{}_2, \dots, f^1{}_{m^1}$$

$$e^2{}_1, e^2{}_2, \dots, e^2{}_{l^2} \Leftrightarrow f^2{}_1, f^2{}_2, \dots, f^2{}_{m^2}$$

.....

$$e^s{}_1, e^s{}_2, \dots, e^s{}_{l^s} \Leftrightarrow f^s{}_1, f^s{}_2, \dots, f^s{}_{m^s}$$

.....

$$e^S{}_1, e^S{}_2, \dots, e^S{}_{l^S} \Leftrightarrow f^S{}_1, f^S{}_2, \dots, f^S{}_{m^S}$$

No. words on English side in s^{th} sentence : l^s
No. words on French side in s^{th} sentence : m^s
$index_E(e^s{}_p)$ =Index of English word $e^s{}_p$in English vocabulary/dictionary
$index_F(f^s{}_q)$ =Index of French word $f^s{}_q$in French vocabulary/dictionary

(Thanks to Sachin Pawar for helping with the maths formulae processing)

Modeling: Hidden variables and parameters

Hidden Variables (Z) :
Total no. of hidden variables $= \sum_{s=1}^{S} l^s\, m^s$ where each hidden variable is as follows:
$z_{pq}^{s} = 1$, if in s^{th} sentence, p^{th} English word is mapped to q^{th} French word.
$z_{pq}^{s} = 0$, otherwise

Parameters (*Θ*) :
Total no. of parameters $= |\, V_E| \times |V_F|$, where each parameter is as follows:
$P_{i,j} =$ Probability that i^{th} word in English vocabulary is mapped to j^{th} word in French vocabulary

Likelihoods

Data Likelihood *L(D; Θ)* :

$$L(D;\Theta) = \prod_{s=1}^{S}\prod_{p=1}^{l^S}\prod_{q=1}^{m^S}\left(P_{index_E(e_p^S),index_F(f_q^S)}\right)^{z_{pq}^S}$$

Data Log-Likelihood LL(D; Θ) :

$$LL(D;\Theta) = \sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} z_{pq}^s \, log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)$$

Expected value of Data Log-Likelihood E(LL(D; Θ)) :

$$E(LL(D;\Theta)) = \sum_{s=1}^{S}\sum_{p=1}^{l^S}\sum_{q=1}^{m^S} E(z_{pq}^s)\; log\left(P_{index_E(e_p^s),index_F(f_q^S)}\right)$$

Constraint and Lagrangian

$$\sum_{j=1}^{|V_F|} P_{i,j} = 1 \ , \forall i$$

$$\sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} E(z_{pq}^s)\, log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right) - \sum_{i=1}^{|V_E|} \lambda_i \left(\sum_{j=1}^{|V_F|} P_{i,j} - 1\right)$$

Differentiating wrt P_{ij}

$$\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\left(\frac{E(z_{pq}^s)}{P_{i,j}}\right)-\lambda_i=0$$

$$P_{i,j}=\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\,E(z_{pq}^s)$$

$$\sum_{j=1}^{|V_F|}P_{i,j}=1=\sum_{j=1}^{|V_F|}\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\,E(z_{pq}^s)$$

Final E and M steps

M-step

$$P_{i,j} = \frac{\sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} \delta_{index_E(e_p^s),i}\, \delta_{index_F(f_q^s),j} E(z_{pq}^s)}{\sum_{j=1}^{|V_F|} \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} \delta_{index_E(e_p^s),i}\, \delta_{index_F(f_q^s),j} E(z_{pq}^s)}, \forall i,j$$

E-step

$$E(z_{pq}^s) = \frac{P_{index_E(e_p^s),index_F(f_q^s)}}{\sum_{q'=1}^{m^s} P_{index_E(e_p^s),index_F(f_{q'}^s)}}, \forall s,p,q$$

Summary (1/2)

- Motivated the problem: necessity of probability
- Estimation of parameters in 2 ways: MLE and ME
- Defined principles of MLE and ME
- Showed reduction in entropy when data arrives
- Introduced constrained distribution and exponential distribution

Summary (2/2)

- Used Pythagorean theorem for establishing duality
- The intersection of exponential distribution family and the constrained distribution family is the all important zone.
- Exponential family addresses the entropy part.
- Constrained family addresses the likelihood part

Conclusions (1/2)

- Only two paths to parameter estimation
- When no data, only thing one can do is use MLE
- When data is present MLE is good; but forming the likelihood expression may be cumbersome
- ME can then give a way to incorporate data through constraints
- Usual constraint is expected values of features (idea and empirical) should match

Conclusions (2/2)

- The duality holds for exponential family and constrained distribution- this is important
- Algorithmically speaking, MLEs are in general easier for the training phase- typically construct a few table of ratios of counts
- For ME the training phase is cumbersome; e.g., MEMM and CRF use complicated algorithms which are forms of gradient descent and take a long time
- Both approaches, of course, need decoding- some form of Beam Search or Viterbi

Acknowledgment

- Notes of Amnon Shashua, Lectures on Machine Learning
- Notes from Adwait Ratnaparkhi’s lecture on Machine Learning
- Chapter 2 of the book “Machine Translation”, by Pushpak Bhattacharyya, CRC Press, 2015

Thank you

http://www.cse.iitb.ac.in/~pb

pb@cse.iitb.ac.in