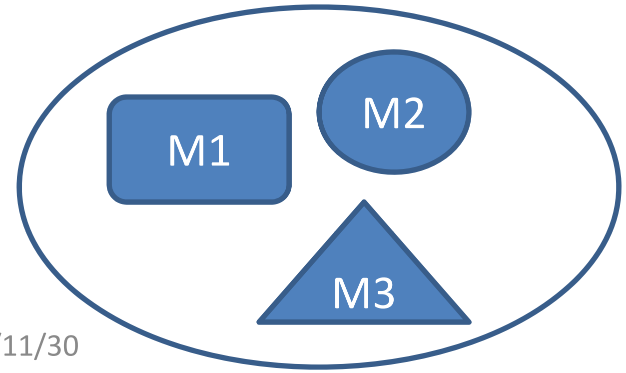
Accessing the Models

Prof. Shou-de Lin
CSIE/GINM, NTU
Sdlin@csie.ntu.edu.tw

Questions

- For a theory (or hypothesis of model) T, how do we know if one set of parameter estimates is better than another?
- Which is better? Theory T with parameter estimates X or theory S with parameter estimates P?
- Knowledge Discovery is about finding a best model with optimal parameter that fits the given data, then use the model to find something useful.



Model Assessment: A Bayesian Approach

- d= data (observation), m=model (how the data are generated)

$\underset{m}{aug \max}\, p(m \mid d)$ ➔ most likely model given data

Baye's Rule

$$\underset{m}{aug \max}\, p(m \mid d) = \underset{m}{aug \max} \frac{p(m) * p(d \mid m)}{p(d)} =$$

$$\underset{m}{aug \max}\, p(m) * p(d \mid m)$$

Does this model look reasonable?

Given the fixed model m, does the observed data stream look reasonable?

Maximum Likelihood Estimation (MLE)

- If p(m) in unknown, then we can only evaluate

$$\arg\max_{m} p(d \mid m)$$

,which is usually quantitative !! ➔ thank god ☺

- E.g. d=H H T H
  - M1: coin is unbiased p(d|m)= $0.5^4 = 0.066$
  - ☺ – M2: coin is biased s.t. p(H)=3/4, p(d|m)= $\frac{3}{4} * \frac{3}{4} * \frac{1}{4} * \frac{3}{4} = 0.1$
  - M3: coin is biased so that P(H)=0.9, p(d|m) $= 0.073$

$$\underset{m}{aug\max}\; p(m) * p(d \mid m)$$

- What if p(m) is not uniform (e.g. we examine the coin and find nothing wrong with it)
- E.g. P(M1): 0.9, P(M2):0.05, P(M3):0.05
- Then in the previous example,
  - P(M1)*P(d|M1)= 0.9*0.066=0.059 ☺
  - P(M2)*P(d|M2)= 0.1*0.1=0.01
  - P(M3)*P(d|M3)= 0.1*0.073=0.0073

Unsupervised Learning

Prof. Shou-de Lin
GSIE/GINM, NTU

To Bring you Back to the Earth

In the "whatever I want to do lecture", I'll teach

- **Supervised learning.** (2 hours)
  - Generative learning algorithms. Gaussian discriminant analysis.
- **Unsupervised learning.** (3 hours)
  - *EM (why? Because it is as magical as you should know).*

    Note: Last year I used 3 full lectures teaching EM
  - *Clustering: K-means (why? Because it is as simple as you should know)*
- **Reinforcement learning** (0.5 hour)
  - Value iteration and policy iteration.
  - Q-learning & SARSA

What is Unsupervised Learning

- Supervised learning: we are given a set of training data X given a class, and we want to learn a function f(x)= y that maps x to y
- Unsupervised Learning:
  - Clustering: given x, grouping x into different clusters.
  - EM: given x and partial information about y, trying to learn f(x).
    - EM is the key solution to many knowledge discovery tasks.

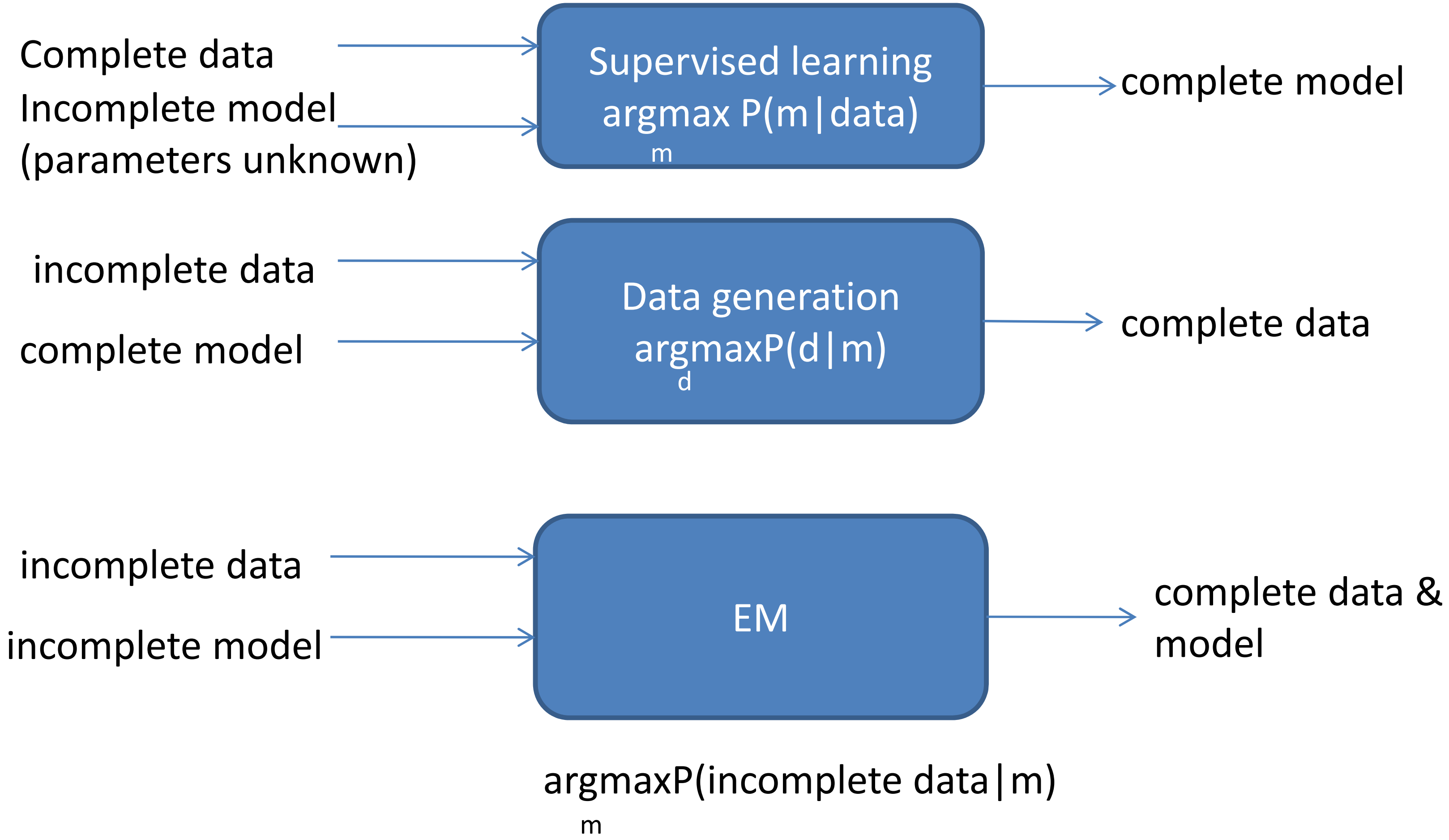
Analogy: Decipherment

- SL: given a bunch of words X and its cipher Y, trying to figure out f(X)=Y. For example, (X,Y)=

  (byf, axe) (hppe, good) (bqqmf, apple), f=?
- However, this is not how decipherment works in the real world. People didn't decipher Egyptian or Maya this way. They did it through an unsupervised manner (only X is given, and they need to translate it into Y):

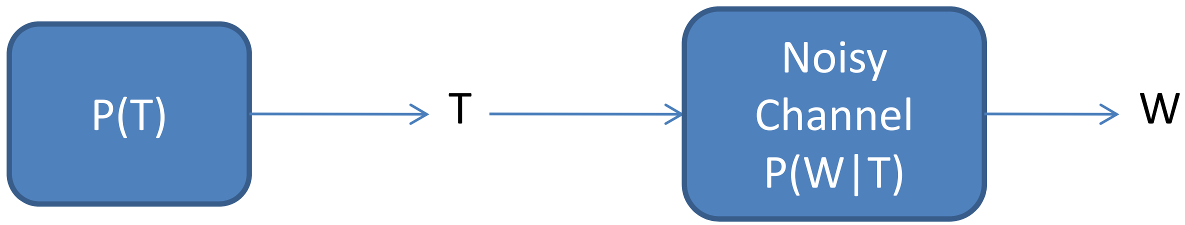
  X=(byf, hppe, bqqmf …), f=?

Data and Model



Ideal vs. Available Data – Sequential Labeling (POS tagging)

- Part of speech tagging:



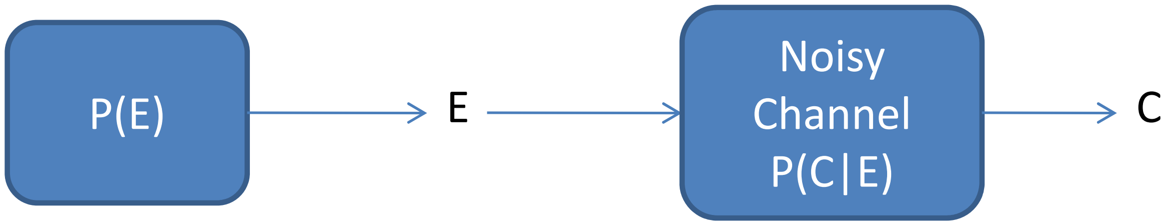
- Ideal: t_1 t_2 t_3 .....

  w_1 w_2 w_3 ....

- Available: w_1 w_2 w_3 ....

Ideal vs. Available Data - Cryptography

- Cryptography:



- Ideal: $e_1\ e_2\ e_3$ ..... (solvable by SL)

  $\downarrow\ \downarrow\ \downarrow$

  $c_1\ c_2\ c_3$ .....

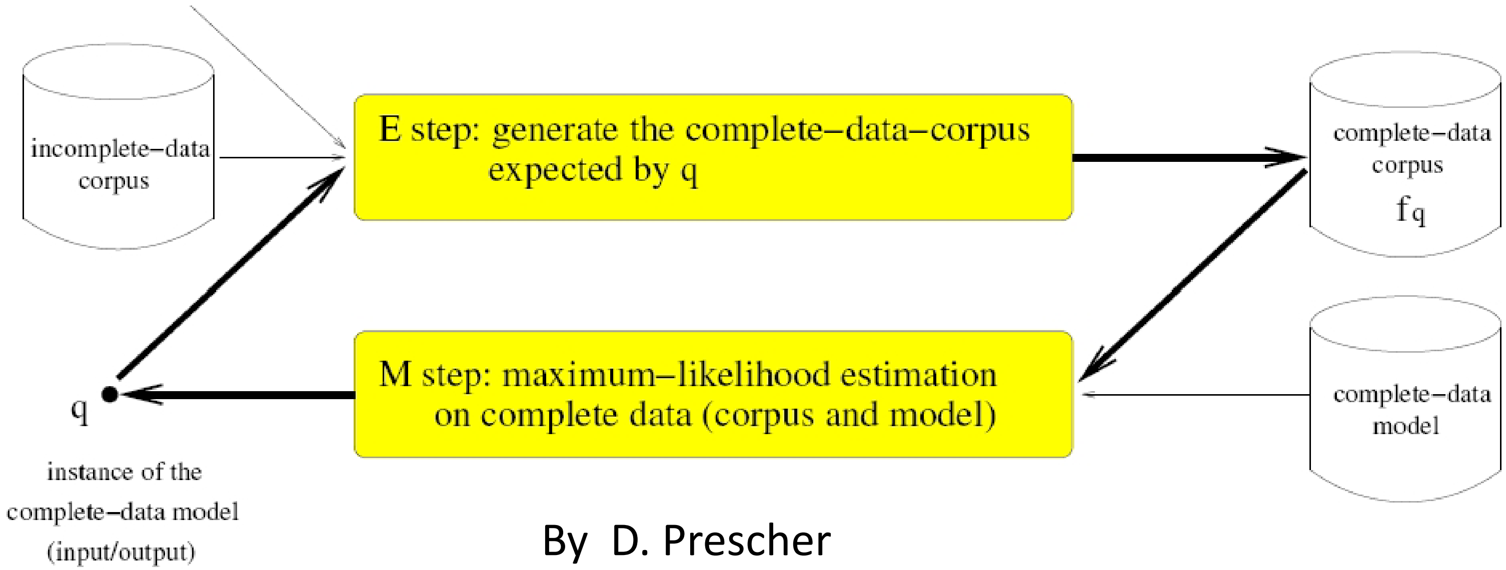
- Available: $c_1\ c_2\ c_3$ .... (need EM)

Introducing EM

- Expectation Maximization (EM) is perhaps most often used and mostly half understood algorithm for unsupervised learning.
  – It is very intuitive.
  – Many people rely on their intuition to apply the algorithm in different problem domains.
  – It is not an algorithm instead a framework. Different algorithms can be designed based on EM framework.
- Note: The following slides integrate some people's materials and viewpoints about EM, including Kevin Knight, Dekang Lin, D. Prescher, and Dan Klein.

EM framework

- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to re-estimate parameters



By D. Prescher

Handling incomplete Data

- Our goal is to build a probabilistic model of data (e.g. LM), defined by a set of parameters θ
- The model parameters can be estimated from a set of IID training examples: $x_1, x_2, \ldots, x_n$
- Unfortunately, we only get to observe partial information about x's, for example:
  - $x_i=(t_i, y_i)$ and we can only observe y_i. The t_i's are the so-called "hidden" data that will be modeled by the "hidden" variables in EM.
- How can we still construct the model?

Example MLE

- A coin with P(H)=p, P(T)=q. We observed m H's and n T's.
- Q: What are p and q according to MLE?
- Solution:
- Maximize $\Sigma_i \log P_\theta(y_i)= \log p^m q^n = m \log p + n \log q$, under the constraint: p+q=1
- Lagrange Method:
  - Define $g(p,q)=m \log p + n \log q+\lambda(p+q-1)$
  - Solve the equations: $\frac{\partial g(p,q)}{\partial p}=0, \quad \frac{\partial g(p,q)}{\partial q}=0, \quad p+q=1$

But if the data is incomplete

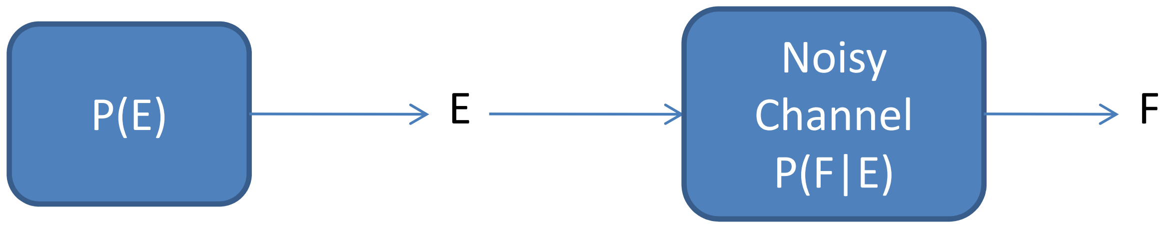
- Suppose we have two coins. Coin 1 is fair. Coin 2 has probability p generating H.
- They each have x probability to be chosen.
- We only know the result of the toss, but don't know when coin was chosen.
  - The complete data is (1, H), (1, T), (2, T), (1, H), (2, T)
  - The observed data is H, T, T, H, T.
- What are p, q and x?

EM Properties

- EM is a general technique for learning anytime we have incomplete data (x,y)
- Each step of EM is guaranteed to increase data likelihood - a hill climbing procedure
- Not guaranteed to find global maximum of data likelihood
  - Data likelihood typically has many local maxima for a general model class and rich feature set
  - Many “patterns” in the data that we can fit our model to…

Ideal vs. Available Data – Alignment Problem for Machine Translation

- MT:



- Ideal: e_1 e_2 e_3 ….. (solvable by SL)

  f_1 f_2 f_3 ….
- Available: e_1 e_2 e_3 ….. (need EM)

  f_1 f_2 f_3 ….

Ex: English-French Alignment

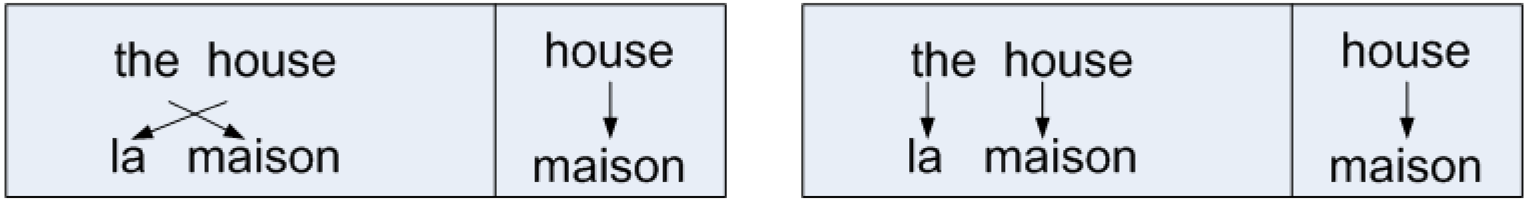
- Data: the house → la maison,

  house → maison
- Alignments are missing!!
- Theory: English words are translated first, then permuted.
- Parameters: P(la|the), p(maison|the), p(la|house), p(maison|house)

Ex: EMTraining on MT

Model to learn:
P(la|the)=?
P(maison|the)=?
P(la|house)=?
P(maison|house)=?

- Possible assignments:



(a) (b)

initialize uniformly:

P(la|the)=1/2
P(maison|the)=1/2
P(la|house)=1/2
P(maison|house)=1/2

Score

p(a)= 1/8
p(b)= 1/8

Fractional counts

C(la|the)=0*1/8+1*1/8=1/8
C(maison|the)=1 *1/8+0*1/8=1/8
C(la|house)=1*1/8+0*1/8=1/8
C(maison|house)=1*1/8+2*1/8=3/8

normalize

p(la|the)=1/2
p(maison|the)=1/2
p(la|house)=1/4
p(maison|house)=3/4

score

P(a)=3/32
P(b)=9/32

Fractional counts

C(la|the)=9/32
C(maison|the)=3/32
C(la|house)=3/32
C(maison|house)=21/32

normalize

p(la|the)=3/4
p(maison|the)=1/4
p(la|house)=1/8
p(maison|house)=7/8

P(a)=7/256
P(b)=147/256