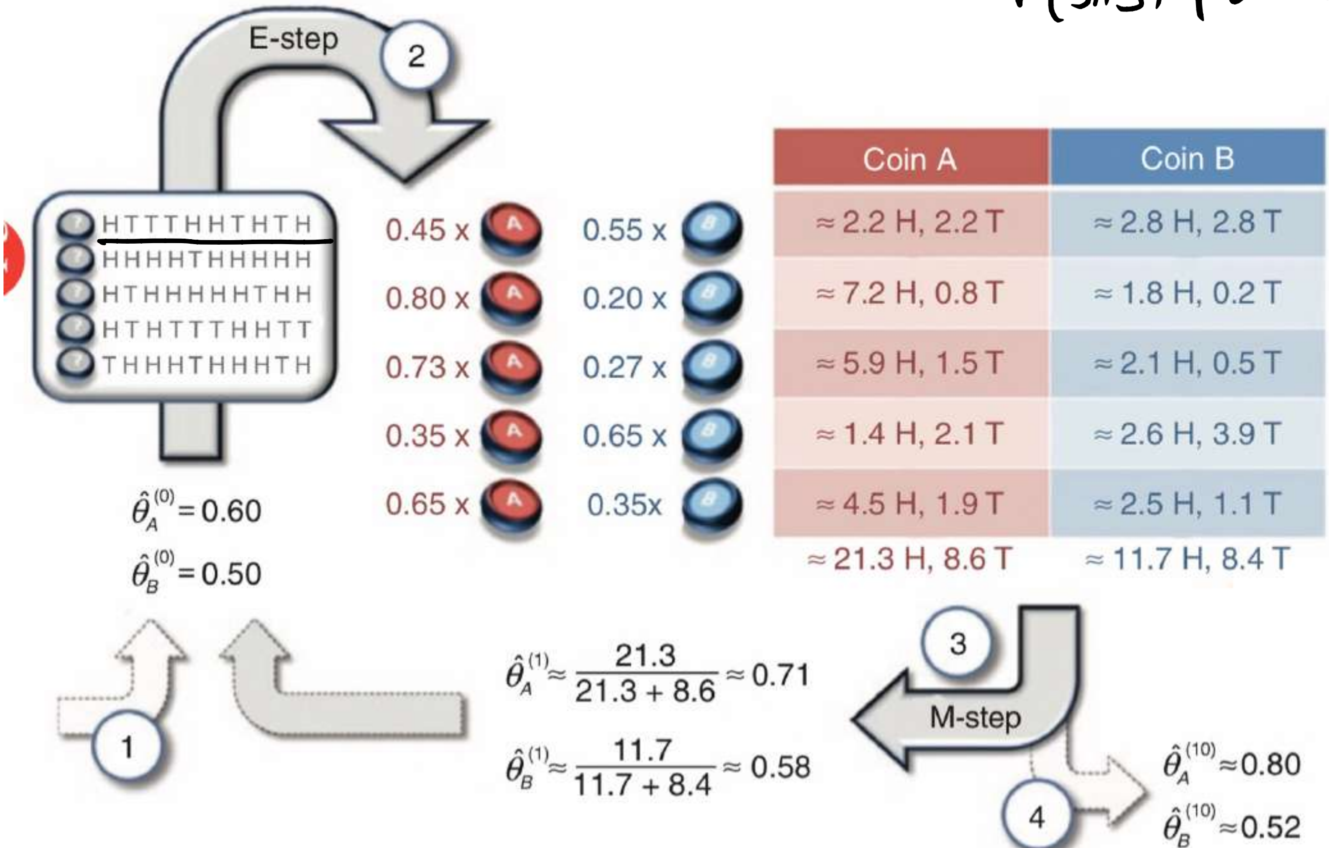
DMML, 20 Feb 2020

5H5T: $P(5H5T \mid \theta = 0.60) = P_1$

$P(5H5T \mid \theta = 0.50) = P_2$

$\frac{P_1}{P_1+P_2}$ $\frac{P_2}{P_1+P_2}$

0.45 0.55



Bottleneck of supervised learning - need for labelled training data

Volume + Timeliness

Bootstrapping — Semi Supervised Learning

1. EM - Expectation Maximization

Text Classification - Naive Bayes

$$P(w_i | c_j)$$

$$P(d_i | c_j)$$

$$P(w_i|c_j) = \frac{\text{Occ. of } w_i \text{ in } c_j}{\text{All words in } c_j}$$

$$= \frac{\sum_{d_k \in c_j} n_{ik}}{\sum_{w_l \in V} \sum_{d_k \in c_j} n_{lk}}$$

n_{ik} = # of occurrences of w_i in d_k

0 if $d_k \notin c_j$

1 if $d_k \in c_j$

Instead

$$\frac{\sum_{d_k} n_{ik} P(d_k|c_j)}{\sum_{w_l \in V} \sum_{d_k} n_{lk} P(d_k|c_j)}$$

Semi Supervised Learning

Supervised case — compute $P(w_i|c_j)$, $P(d_k|c_j)$

Given a new d

Compute $\underline{P(c_i|d)}$ for all c_i

$$\arg\max_i P(c_i|d)$$

Fractional assignments of categories that we discard after arg max

Semi Supervised

Label, say, 10% of documents "randomly selected"

↓

Naive Bayes model

↓

Fractional topic allocation per document
class

↓

Now $P(d_k|c_i)$ makes sense!

New estimate of $P(w_i|c_j)$, $P(d_k|c_j)$

Another example

MNIST digit images

training data (1000) → test data

label only 50

Cluster remaining 950 using these 50 as centroids

- labels are inherited from centroid

Use distance from centroid for logistic regression

50 labelled point $\xrightarrow[\text{1000}]{\text{extend to}}$ 1000 labelled point

50 labelled point → Model → Test data accuracy A_1

1000 labelled point → Model → Accuracy A_2

50 labelled point → Extend labels to "nearby" point

Nearest 20% pts in each cluster → Model A_3

$A_1 < A_2 < A_3$