DMML 5 Feb 2019

Naïve Bayesian Classifiers

$$P(A_1 = a_1, A_2 = a_2, \ldots, A_k = a_k \mid C = c)$$

$$= P(A_1 = a_1 \mid C = c) \times \ldots \times P(A_k = a_k \mid C = c)$$

Text classification

Documents $d_1, \ldots, d_N$ over vocabulary V

Topics $t_1, \ldots, t_M$

Each d_i has topic t_j

Bodean model of documents — set of words

$P(w_i | t_j)$ – fraction of docs of topic t_j that contain w_i

$P(t_j)$ – fraction of documents of type t_j

$$P(t_j | d) = \prod_{w \in V} P(t_j | d(w))$$

$d(w)$: 0 or 1

$$P(w | t_j) \cdot P(t_j)$$

Richer model - count multiplicities of words

Set of words → Set with multiplicities
= multiset / bag

$f: V \to \{0,1\}$ $\quad$ $g: V \to \mathbb{N}_0$

Assume we have $P(w_i | t_j)$

- Pick a topic $P(t_j)$
- Pick a length l - $P(l | t_j)$?
  - Assume $P(l)$ is independent of t_j

- For $i = 1$ to l,

  generate a word from V

  roll a $|V|$-sided die – each face represents a word w, and it is chosen with $P(w|t_j)$

d of length l: $w_1 w_2 \ldots w_i \ldots w_l$

topic t_j ↓ $P(w_i|t_j)$

$$P(d|t_j) = \prod_{i=1}^{l} P(w_i|t_j)$$. Permutations etc

Rewrite this expression

For each $w \in V$, n_w is number of times w occurs in d

$$|d|! \cdot \prod_{w \in V} \frac{P(w|t_j)^{n_w}}{n_w!}$$

— $P(t)$ not required!

$P(t|d)$ — invert and write in terms of $P(w|t_j)$

Main hurdle with supervised learning is labelling training data

Semi Supervised Learning – iteratively improve labelling starting from "simple" state

Problem Mixture of models

Each topic t_j has a family of parameters $P(w_i|t_j)$

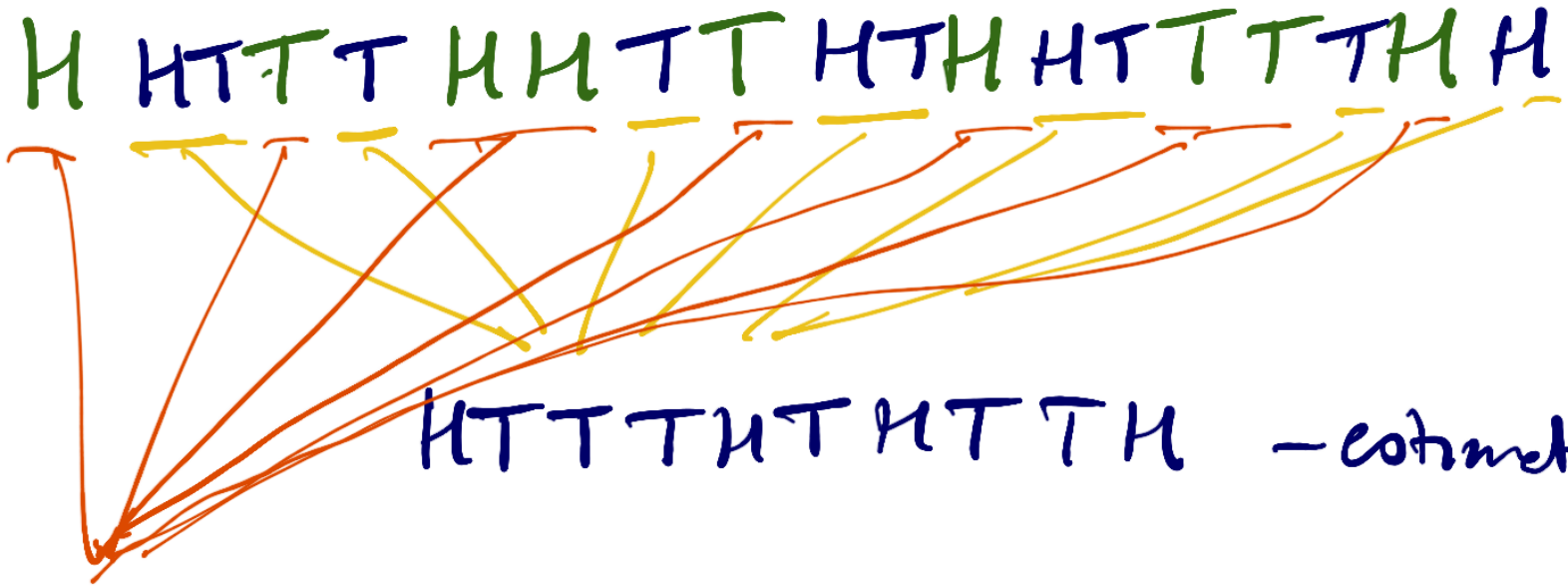
Suppose we have two coins with possibly different characteristics

Green : $p(H) = a$

Blue : $p(H) = b$

Single coin

Given a random sequence of coin tosses, estimate $p(H) = \frac{\#\text{heads}}{\text{length}}$



Suppose we don't have blue/green labels?

HTHTTTHHTHTHTHTTTHT

Need a generative explanation

- Pick Blue/Green with probability $1/2$
- Toss it
- Repeat

Fix a specific point in the sequence

H T H T H T H T H T H T H ...

$P(\text{Green}_{\text{coin}}) = \frac{a}{a+b}$ $\quad P(\text{Blue}) = \frac{b}{a+b}$

Attribute fractional weight for each toss to the two coins according to given assumption about a & b

$a = \frac{2}{3}$

$b = \frac{1}{3}$

$\frac{2}{3}\ \frac{1}{3}\ \frac{2}{3}\ \frac{2}{3}\ \frac{1}{3}$ $= \frac{8}{3}$ Green tosses

H T H H T

$\frac{1}{3}\ \frac{2}{3}\ \frac{1}{3}\ \frac{1}{3}\ \frac{2}{3}$

$= \frac{7}{3}$ blue tosses

of which $\frac{6}{3} = H$

↓

Revise a to $\frac{6}{8} = \frac{3}{4}$

$\frac{3}{3}$ are H

$\Rightarrow$ Revise b to $\frac{3}{7}$

Hopefully this converges ?

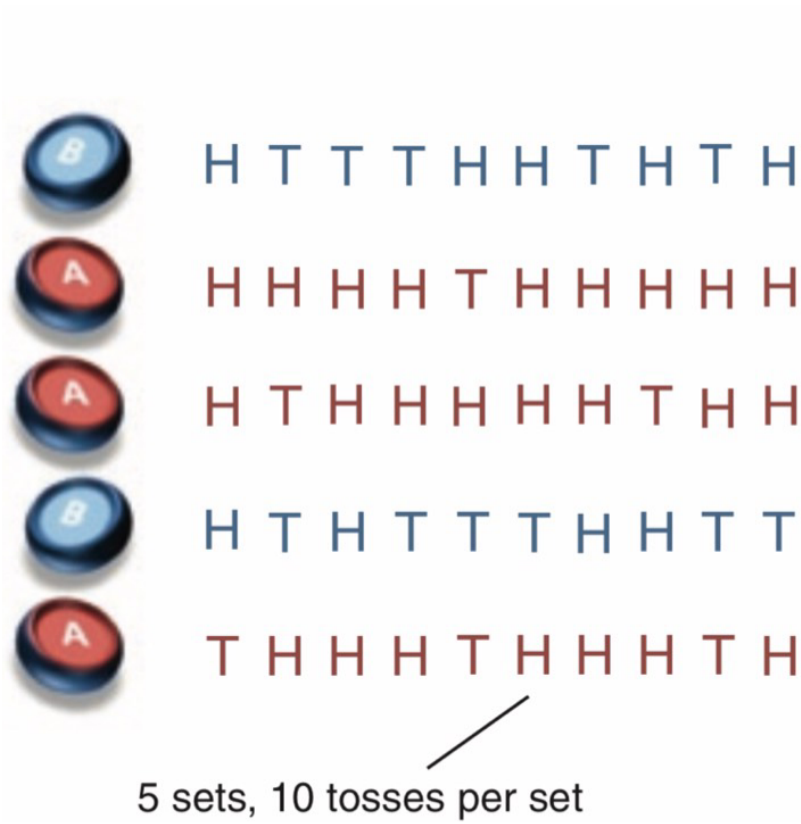
Expectation - Maximization (EM)

E - Compute prob. of observation with current estimates

M - Re-estimate - MLE calculation

Converges to a local optimum

Start with some estimates for A & B



	Coin A	Coin B
B		5 H, 5 T
A	9 H, 1 T	
A	8 H, 2 T	
B		4 H, 6 T
A	7 H, 3 T	
	24 H, 6 T	9 H, 11 T

If we had used

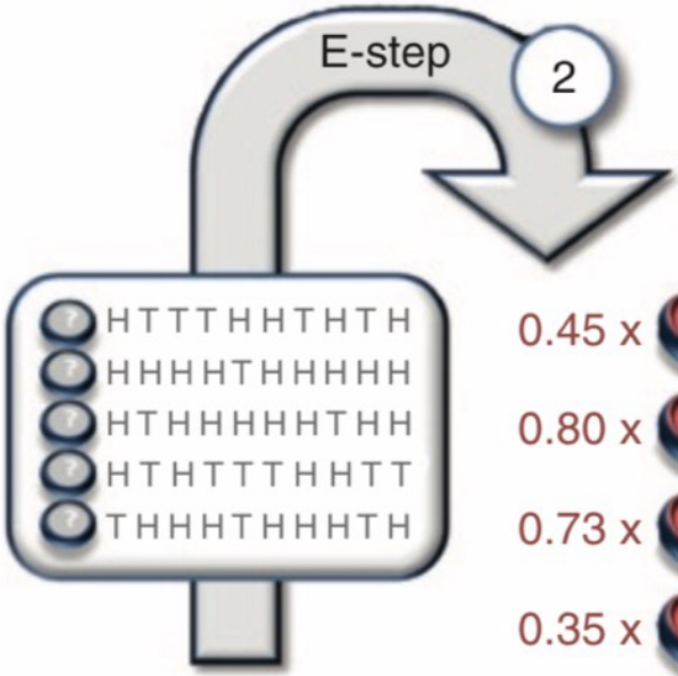
		Coin A	Coin B
0.45 x	0.55 x	≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
0.80 x	0.20 x	≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
0.73 x	0.27 x	≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
0.35 x	0.65 x	≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
0.65 x	0.35x	≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T

Initial estimate $P(A) = 0.6$

$P(B) = 0.5$

$P(5H, 5T)$ for $A = P_A$

$P(5H, 5T)$ for $B = P_B$



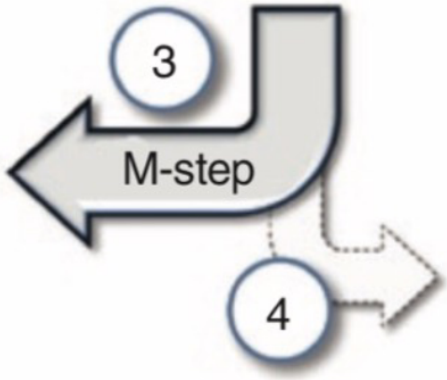
$\hat{\theta}_A^{(0)} = 0.60$

$\hat{\theta}_B^{(0)} = 0.50$

	Coin A	Coin B
0.45 x A, 0.55 x B	≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
0.80 x A, 0.20 x B	≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
0.73 x A, 0.27 x B	≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
0.35 x A, 0.65 x B	≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
0.65 x A, 0.35x B	≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
	≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$$

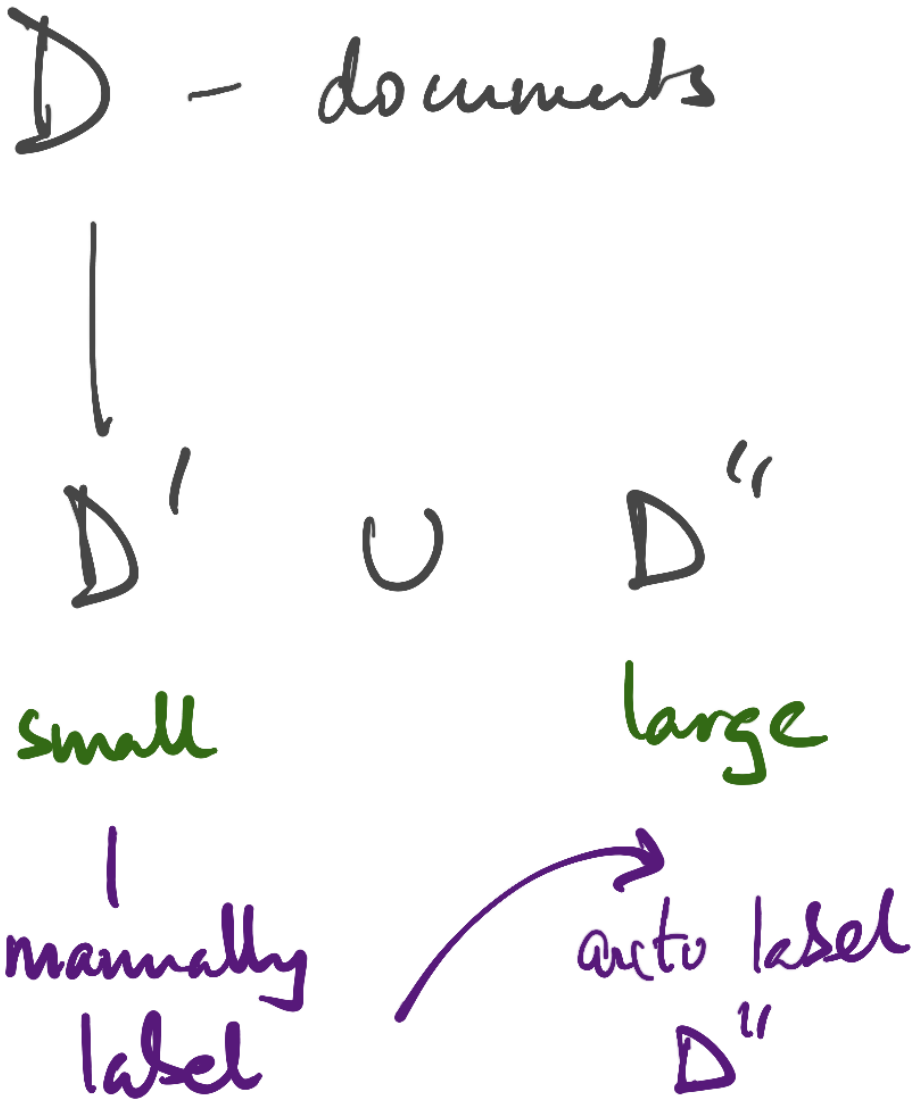
$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$$



$\hat{\theta}_A^{(10)} \approx 0.80$

$\hat{\theta}_B^{(10)} \approx 0.52$

Topic classification using EM



Automatic topic discovery

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

Assume we fix the number of topics

- **Sentences 1 and 2**: 100% Topic A
- **Sentences 3 and 4**: 100% Topic B
- **Sentence 5**: 60% Topic A, 40% Topic B
- **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
- **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

Latent Dirichlet Analysis (LDA)

- Initially randomly assign each word to a topic
- Assign (fractional) topics to each doc. based on assignment of topics to words
- Count words using fractional topics & get $P(w|t)$
- Reassign random topics to each w with $P(t|w)$