Lecture 3: Probabilistic models

Helen Yannakoudakis and Oana Cocarascu

Department of Informatics
King's College London

January 31, 2022

Lecture 3 Q&A

Can you elaborate (with a small example, if possible) how overfitting can happen when we estimate probabilities by simple counting?

Lecture 3 Q&A

- With more features can get easily underflow errors.
- Avoid this by taking logs:

$$v_{NB} = \arg\max_{v \in V} p(v) \prod_j p(a_j|v)$$

$$v_{NB} = \arg\max_{v \in V} logP(v) + \sum_j logP(a_j|v)$$

- In Naive Bayes, we compute the conditionals by counting instances.
- In sentiment classification:

$$P(v) = \frac{N_c}{N_{doc}}$$

$$P(a_j|v) = P(w_j|v) = \frac{count(w_j, v)}{\sum_{w \in V} count(w, v)}$$

Lecture 3 Q&A

$$v_{NB} = \arg\max_{v \in V} logP(v) + \sum_j logP(a_j|v)$$

- In sentiment classification:

$$P(v) = \frac{N_c}{N_{doc}}$$

$$P(a_j|v) = P(w_j|v) = \frac{count(w_j,v)}{\sum_{w \in V} count(w,v)}$$

$$P(w_j|v) = P(\text{“}fantastic\text{''}\,|\,positive) = \frac{count(\text{“}fantastic\text{''}, positive)}{\sum_{w \in V} count(w, positive)} = 0$$

Lecture 3 Q&A

- If the number of training examples is very small, then our calculations are not very accurate.

- Add-one smoothing (aka *Laplace* smoothing):

$$P(a_j|v) = P(w_j|v) = \frac{count(w_j,v)+1}{\sum_{w \in V}(count(w,v)+1)} = \frac{count(w_j,v)+1}{(\sum_{w \in V}(count(w,v))+|V|}$$

(see Murphy's book and Jurafsky & Martin, Chapters 3 & 4 on Naive Bayes and smoothing zero counts, including worked examples).

Expectation Maximisation coin-flipping example.

Coin flipping experiment

- We have a pair of coins A and B.
- Coin A will land on heads with probability θ_A and tails with $1 - \theta_A$.
- Coin B will land on heads with probability θ_B and tails with $1 - \theta_B$.
- How can we estimate $\theta = (\theta_A, \theta_B)$?

Coin flipping experiment

- We have a pair of coins A and B.
- Coin A will land on heads with probability θ_A and tails with $1 - \theta_A$.
- Coin B will land on heads with probability θ_B and tails with $1 - \theta_B$.
- How can we estimate $\theta = (\theta_A, \theta_B)$?
- Randomly pick a coin (with equal probability) and perform 10 independent coin tosses; Repeat this 5 times, so total 50 coin tosses.

Coin flipping experiment

- We have a pair of coins A and B.
- Coin A will land on heads with probability θ_A and tails with $1 - \theta_A$.
- Coin B will land on heads with probability θ_B and tails with $1 - \theta_B$.
- How can we estimate $\theta = (\theta_A, \theta_B)$?
- Randomly pick a coin (with equal probability) and perform 10 independent coin tosses; Repeat this 5 times, so total 50 coin tosses.
- So assuming we know which coin was used in each set of tosses:

Coin flipping experiment

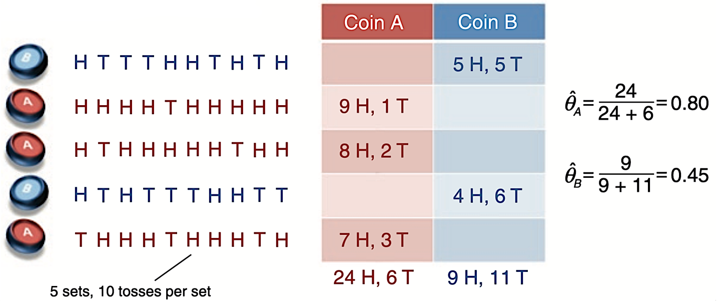
- We have a pair of coins A and B.
- Coin A will land on heads with probability θ_A and tails with $1 - \theta_A$.
- Coin B will land on heads with probability θ_B and tails with $1 - \theta_B$.
- How can we estimate $\theta = (\theta_A, \theta_B)$?
- Randomly pick a coin (with equal probability) and perform 10 independent coin tosses; Repeat this 5 times, so total 50 coin tosses.
- So assuming we know which coin was used in each set of tosses:

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

5 sets, 10 tosses per set

Coin flipping experiment

- We have a pair of coins A and B.
- Coin A will land on heads with probability θ_A and tails with $1 - \theta_A$.
- Coin B will land on heads with probability θ_B and tails with $1 - \theta_B$.
- How can we estimate $\theta = (\theta_A, \theta_B)$?
- Randomly pick a coin (with equal probability) and perform 10 independent coin tosses; Repeat this 5 times, so total 50 coin tosses.
- So assuming we know which coin was used in each set of tosses:

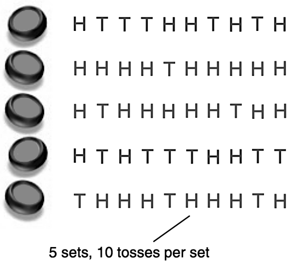


Coin flipping experiment

- We have a pair of coins A and B.
- Coin A will land on heads with probability θ_A and tails with $1 - \theta_A$.
- Coin B will land on heads with probability θ_B and tails with $1 - \theta_B$.
- How can we estimate $\theta = (\theta_A, \theta_B)$?
- Randomly pick a coin (with equal probability) and perform 10 independent coin tosses; Repeat this 5 times, so total 50 coin tosses.
- But what if we don't know which coin was used in each set of tosses?

Coin flipping experiment

- We have a pair of coins A and B.
- Coin A will land on heads with probability θ_A and tails with $1 - \theta_A$.
- Coin B will land on heads with probability θ_B and tails with $1 - \theta_B$.
- How can we estimate $\theta = (\theta_A, \theta_B)$?
- Randomly pick a coin (with equal probability) and perform 10 independent coin tosses; Repeat this 5 times, so total 50 coin tosses.
- But what if we don't know which coin was used in each set of tosses?

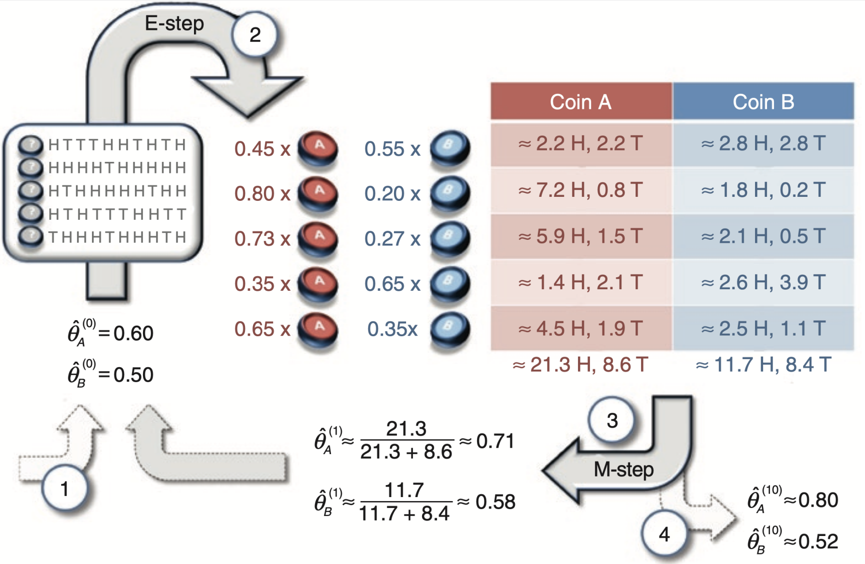


Coin flipping experiment – EM steps

- Start with random guesses for the parameters theta.
- E step: Using these, estimate a probability distribution over the coins for each set of tosses (your hidden variables).
- M step: Based on the above, now use maximum likelihood to estimate new parameters theta.
- Alternate between the two steps until convergence.

Coin flipping experiment – EM

2. $P(A|data, theta) = P(data|A)P(A)/(P(data)) = 0.000796/(0.0009766 + 0.000796) = 0.45$



Does the prior, p(h), depend on the geometry (or some properties) of the model's hypothesis space?

Could you please explain the calculation process of K-means clustering with an example?