Generative vs. Maximum Entropy Models

Mausam

(Slides by Dan Jurafsky, Chris Manning, Pushpak Bhattacharya)

Joint vs. Conditional Models

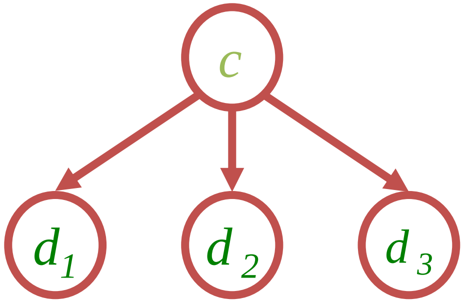
- We have some data {(*d*, *c*)} of paired observations *d* and hidden classes *c*.
- Joint (generative) models place probabilities over both observed data and the hidden stuff (generate the observed data from hidden stuff):
  - All the classic Stat-NLP models:
    - *n*-gram models, Naive Bayes classifiers, hidden Markov models, probabilistic context-free grammars, IBM machine translation alignment models

Joint vs. Conditional Models

- Discriminative (conditional) models take the data as given, and put a probability over hidden structure given the data:
  - Logistic regression, conditional loglinear or maximum entropy models, conditional random fields
  - Also, SVMs, (averaged) perceptron, etc. are discriminative classifiers (but not directly probabilistic)

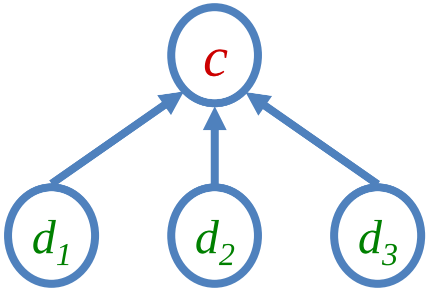
Bayes Net/Graphical Models

- Bayes net diagrams draw circles for random variables, and lines for direct dependencies
- Some variables are observed; some are hidden
- Each node is a little classifier (conditional probability table) based on incoming arcs



Naive Bayes

Generative



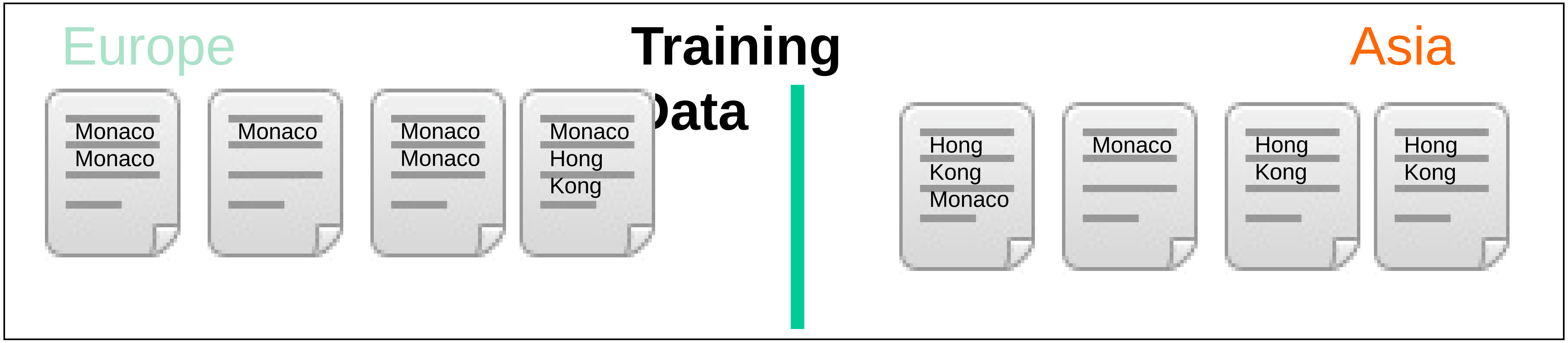
Logistic Regression

Discriminative

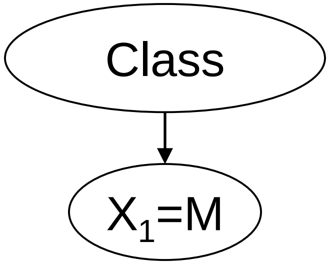
Conditional vs. Joint Likelihood

- A *joint* model gives probabilities P(*d*,*c*) and tries to maximize this joint likelihood.
  - It turns out to be trivial to choose weights: just relative frequencies.
- A *conditional* model gives probabilities P(*c*|*d*). It takes the data as given and models only the conditional probability of the class.
  - We seek to maximize conditional likelihood.
  - Harder to do (as we'll see…)
  - More closely related to classification error.

Text classification: Asia or Europe



NB Model



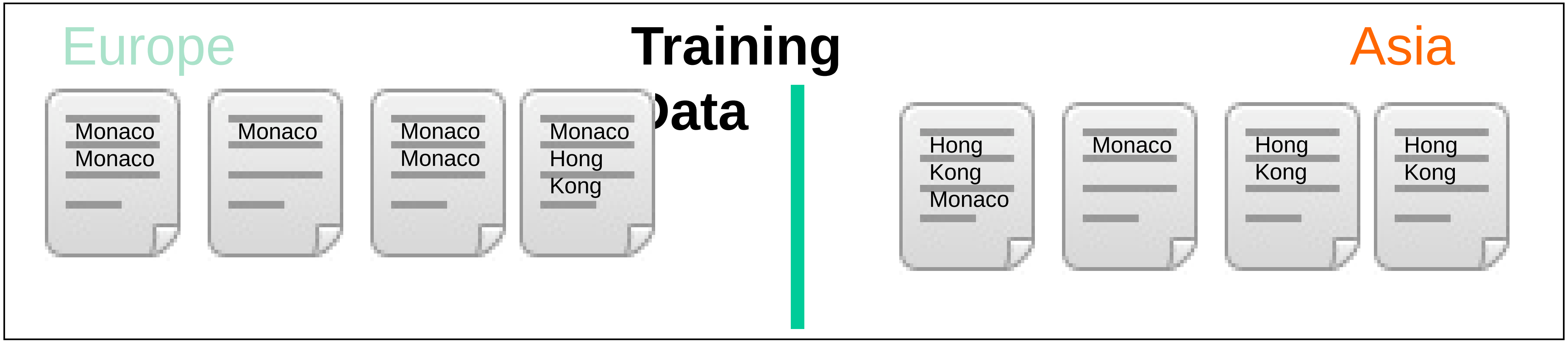
NB FACTORS:

- P(A) = P(E) =
- P(M|A) =
- P(M|E) =

PREDICTIONS:

- P(A,M) =
- P(E,M) =
- P(A|M) =
- P(E|M) =

Text classification: Asia or Europe



Europe

Training Data

Asia

Monaco Monaco

Monaco

Monaco Monaco

Monaco Hong Kong

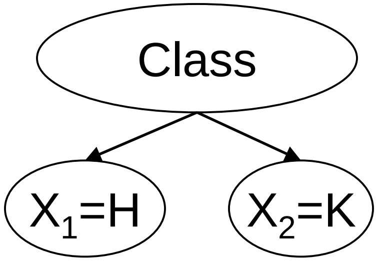
Hong Kong Monaco

Monaco

Hong Kong

Hong Kong

NB Model



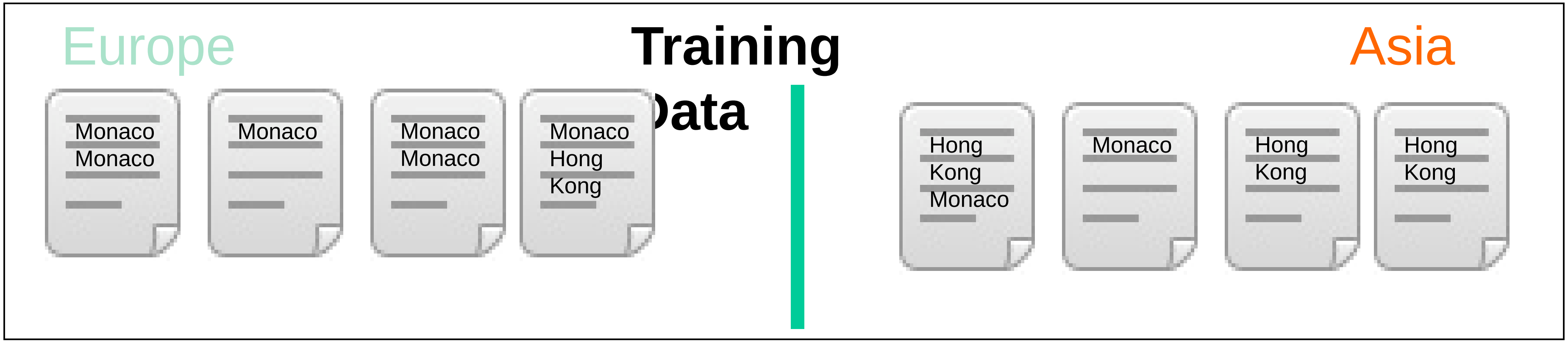
NB FACTORS:

- P(A) = P(E) =
- P(H|A) = P(K|A) =
- P(H|E) = PK|E) =

PREDICTIONS:

- P(A,H,K) =
- P(E,H,K) =
- P(A|H,K) =
- P(E|H,K) =

Text classification: Asia or Europe



Europe

Training Data

Asia

Monaco Monaco

Monaco

Monaco Monaco

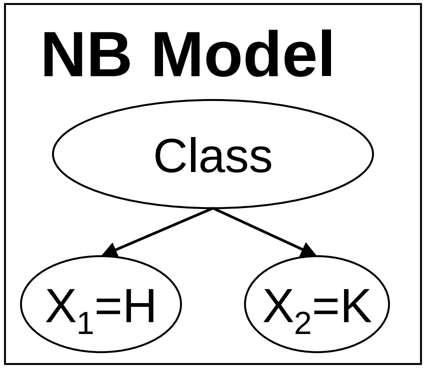
Monaco Hong Kong

Hong Kong Monaco

Monaco

Hong Kong

Hong Kong



NB Model

Class

X_1=H

X_2=K

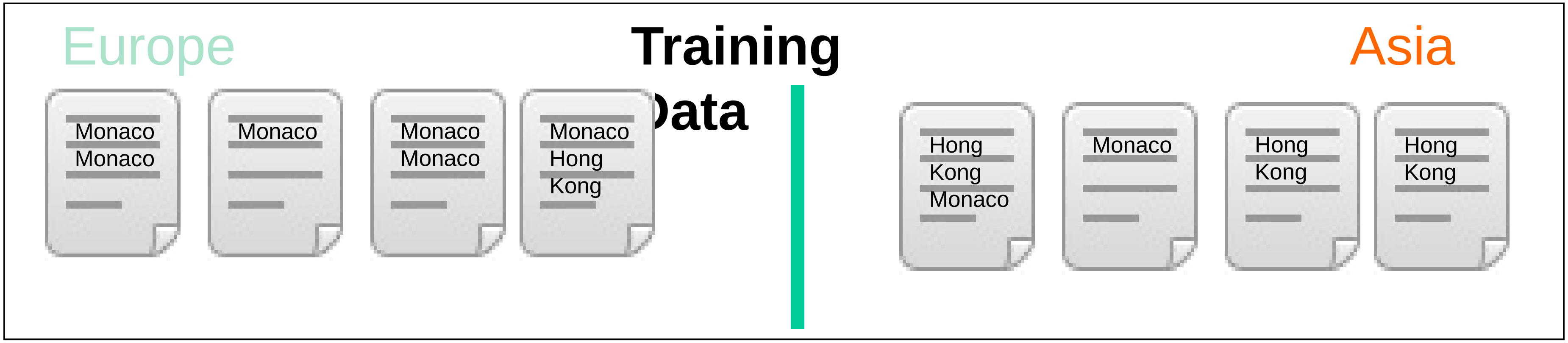
NB FACTORS:

- P(A) = P(E) =
- P(H|A) = P(K|A) =
- P(H|E) = PK|E) =

PREDICTIONS:

- P(A,H,K) =
- P(E,H,K) =
- P(A|H,K) =
- P(E|H,K) =

Text classification: Asia or Europe



Europe

Training Data

Asia

Monaco Monaco

Monaco

Monaco Monaco

Monaco Hong Kong

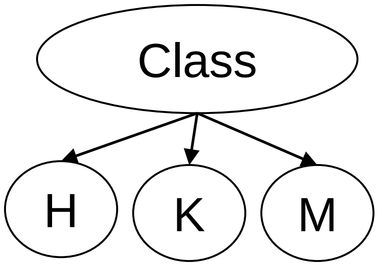
Hong Kong Monaco

Monaco

Hong Kong

Hong Kong

NB Model



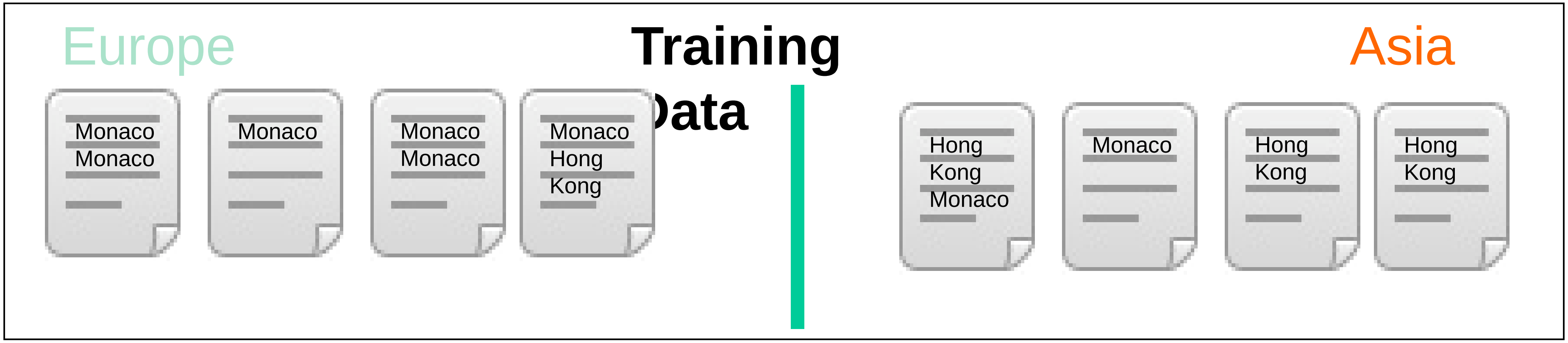
NB FACTORS:

- P(A) = P(E) =
- P(M|A) =
- P(M|E) =
- P(H|A) = P(K|A) =
- P(H|E) = PK|E) =

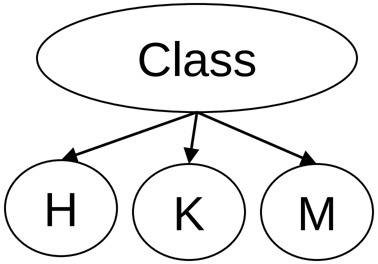
PREDICTIONS:

- P(A,H,K,M) =
- P(E,H,K,M) =
- P(A|H,K,M) =
- P(E|H,K,M) =

Text classification: Asia or Europe



NB Model



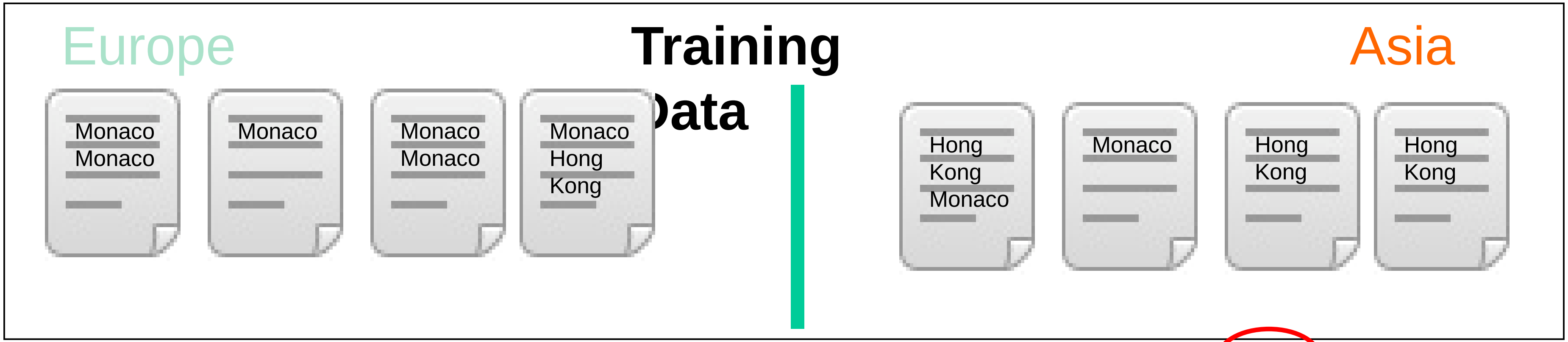
NB FACTORS:

- P(A) = P(E) =
- P(M|A) =
- P(M|E) =
- P(H|A) = P(K|A) =
- P(H|E) = PK|E) =

PREDICTIONS:

- P(A,H,K,M) =
- P(E,H,K,M) =
- P(A|H,K,M) =
- P(E|H,K,M) =

Text classification: Asia or Europe



Europe

Training Data

Asia

Monaco Monaco

Monaco

Monaco Monaco

Monaco Hong Kong

Hong Kong Monaco

Monaco

Hong Kong

Hong Kong

$$P(Y\mid X)=\begin{aligned}&\frac{e^{2me}}{e^{2me}+e^{2ma}}\frac{e^{me}}{e^{me}+e^{ma}}\frac{e^{2me}}{e^{2me}+e^{2ma}}\frac{e^{me+he+ke}}{e^{me+he+ke}+e^{ma+ha+ka}}\cdot\\&\frac{e^{ha+ka+ma}}{e^{ha+ka+ma}+e^{he+ke+me}}\frac{e^{ma}}{e^{ma}+e^{me}}\frac{e^{ha+ka}}{e^{ha+ka}+e^{he+ke}}\frac{e^{ha+ka}}{e^{ha+ka}+e^{he+ke}}\end{aligned}$$

$$P(Y\mid X)=\begin{aligned}&\frac{e^{2me}}{e^{2me}+e^{2ma}}\frac{e^{me}}{e^{me}+e^{ma}}\frac{e^{2me}}{e^{2me}+e^{2ma}}\frac{e^{me+hke}}{e^{me+hke}+e^{ma+hka}}\cdot\\&\frac{e^{hka+ma}}{e^{hka+ma}+e^{hke+me}}\frac{e^{ma}}{e^{ma}+e^{me}}\frac{e^{hka}}{e^{hka}+e^{hke}}\frac{e^{hka}}{e^{hka}+e^{hke}}\end{aligned}$$

Both equations are the same: ha+ka=hka; he+ke=hke… will have same optima

Naive Bayes vs. Maxent Models

- Naive Bayes models multi-count correlated evidence
  - Each feature is multiplied in, even when you have multiple features telling you the same thing

- Maximum Entropy models (pretty much) solve this problem
  - weight features so that model expectations match the observed (empirical) expectations
  - by dividing the weights into all features

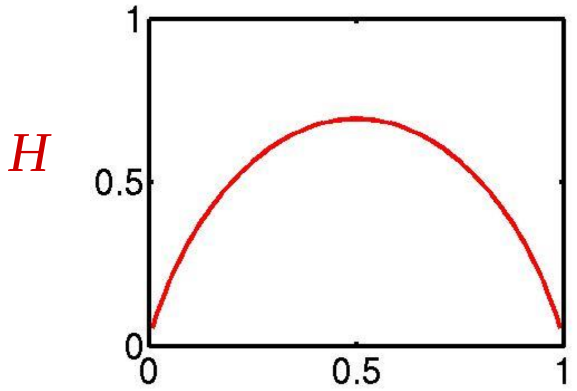
Principle of Maximum Entropy

- Lots of distributions out there, most of them very spiked, specific, overfit.
- We want a distribution which is uniform except in specific ways we require.
- Uniformity means high entropy – we can search for distributions which have properties we desire, but also have high entropy.

Ignorance is preferable to error and he is less remote from the truth who believes nothing than he who believes what is wrong – Thomas Jefferson (1781)

(Maximum) Entropy

- Entropy: the uncertainty of a distribution.
- Quantifying uncertainty ("surprise"):
  - Event x
  - Probability p_x
  - "Surprise" $\log(1/p_x)$
- Entropy: expected surprise (over p):



A coin-flip is most uncertain for a fair coin.

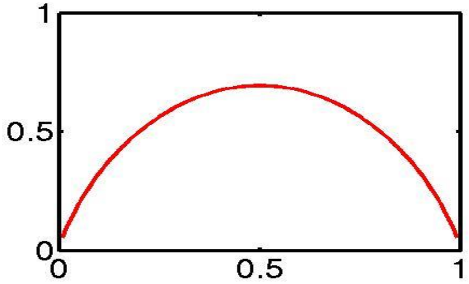
$$\mathrm{H}(p) = E_p\left[\log_2 \frac{1}{p_x}\right] = -\sum_x p_x \log_2 p_x$$

Maxent Examples I

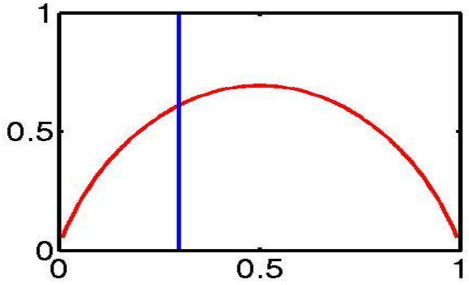
- What do we want from a distribution?
  - Minimize commitment = maximize entropy.
  - Resemble some reference distribution (data).
- Solution: maximize entropy *H*, subject to feature-based constraints:

$$E_p\left[f_i\right] = E_{\hat{p}}\left[f_i\right] \Leftrightarrow \sum_{x \in f_i} p_x = C_i$$

- Adding constraints (features):
  - Lowers maximum entropy
  - Raises maximum likelihood of data
  - Brings the distribution further from uniform
  - Brings the distribution closer to data

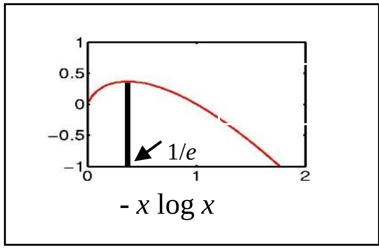


Unconstrained, max at 0.5



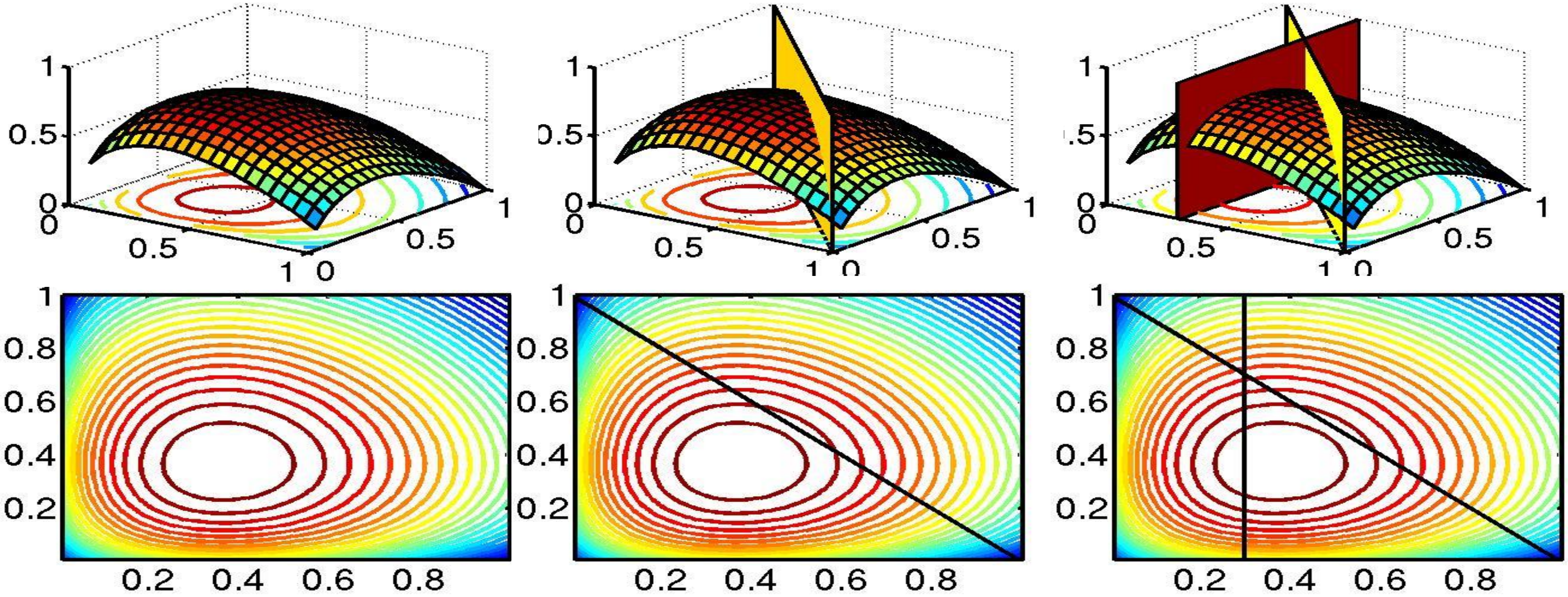
Constraint that $p_{\mathrm{HEADS}} = 0.3$

Maxent Examples II



- $x \log x$

$H(p_H\, p_T,)$ $p_H + p_T = 1$ $p_H = 0.3$



Maxent Examples III

- Let’s say we have the following event space:

Plan	Agt	ML	NLP	Alg	CoTh

- … and the following empirical data:

3	5	11	13	3	1

- Maximize H:

$1/e$	$1/e$	$1/e$	$1/e$	$1/e$	$1/e$

- … want probabilities: E[Plan, Agt, ML, NLP, Alg, CoTh] = 1

1/6	1/6	1/6	1/6	1/6	1/6

Maxent Examples IV

- Too uniform!
- AI is more common than Theory, so we add the feature f_{AI} = {Plan, Agt, ML, NLP}, with E[f_{AI}] =32/36

Plan	Agt	ML	NLP	Alg	CoTh
8/36	8/36	8/36	8/36	2/36	2/36

- … and empirical AI is more frequent than theoretical AI, so we add f_E = {ML, NLP}, with E[f_E] =24/36

4/36	4/36	12/36	12/36	2/36	2/36

- … we could keep refining the models, e.g., by adding a feature to distinguish single vs. multi-agent AI or theory types.

Feature Overlap

- Maxent models handle overlapping features well.
- Unlike a NB model, there is no double counting!

Empirical

	A	a
B	2	1
b	2	1

	A	a
B	⬛	⬛
b	⬛	⬛

All = 1

	A	a
B	1/4	1/4
b	1/4	1/4

	A	a
B		
b		

	A	a
B	🟥	
b	🟥	

A = 2/3

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B	w_A	
b	w_A	

	A	a
B	⬛	
b	⬛	

A = 2/3

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B	$w'_A+w''_A$	
b	$w'_A+w''_A$	

Example: Named Entity Feature Overlap

Grace is correlated with PERSON, but does not add much evidence on top of already knowing prefix features.

Feature Weights

Feature Type	Feature	PERS	LOC
Previous word	*at*	-0.73	0.94
Current word	*Grace*	0.03	0.00
Beginning bigram	*<G*	0.45	-0.04
Current POS tag	NNP	0.47	0.45
Prev and cur tags	IN NNP	-0.10	0.14
Previous state	Other	-0.70	-0.92
Current signature	Xx	0.80	0.46
Prev state, cur sig	O-Xx	0.68	0.37
Prev-cur-next sig	x-Xx-Xx	-0.69	0.37
P. state - p-cur sig	O-x-Xx	-0.20	0.82
…			
Total:		**-0.58**	**2.68**

Local Context

	Prev	Cur	Next
State	Other	???	???
Word	at	Grace	Road
Tag	IN	NNP	NNP
Sig	x	Xx	Xx

Feature Interaction

- Maxent models handle overlapping features well, but do not automatically model feature interactions.

Empirical

	A	a
B	1	1
b	1	0

	A	a
B	■	■
b	■	■

All = 1

	A	a
B	🟥	
b	🟥	

A = 2/3

	A	a
B	■	■
b		

B = 2/3

	A	a
B	1/4	1/4
b	1/4	1/4

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B	4/9	2/9
b	2/9	1/9

	A	a
B	0	0
b	0	0

	A	a
B	w_A	
b	w_A	

	A	a
B	w_A+w_B	w_B
b	w_A	

Feature Interaction

- If you want interaction terms, you have to add them:

Empirical

	A	a
B	1	1
b	1	0

	A	a
B	🟥	
b	🟥	

A = 2/3

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B	⬛	⬛
b		

B = 2/3

	A	a
B	4/9	2/9
b	2/9	1/9

	A	a
B	🟩	
b		

AB = 1/3

	A	a
B	1/3	1/3
b	1/3	0

- A disjunctive feature would also have done it (alone):

	A	a
B	🟦	🟦
b	🟦	

	A	a
B	1/3	1/3
b	1/3	0

Feature Interaction

- For loglinear/logistic regression models in statistics, it is standard to do a greedy stepwise search over the space of all possible interaction terms.
- This combinatorial space is exponential in size, but that’s okay as most statistics models only have 4–8 features.

- In NLP, our models commonly use hundreds of thousands of features, so that’ s not okay.
- Commonly, interaction terms are added by hand based on linguistic intuitions.

Example: NER Interaction

Previous-state and current-signature have interactions, e.g. P=PERS-C=Xx indicates C=PERS much more strongly than C=Xx and P=PERS independently.

This feature type allows the model to capture this interaction.

Local Context

	Prev	Cur	Next
State	Other	???	???
Word	at	Grace	Road
Tag	IN	NNP	NNP
Sig	x	Xx	Xx

Feature Weights

Feature Type	**Feature**	**PERS**	**LOC**
Previous word	*at*	-0.73	0.94
Current word	*Grace*	0.03	0.00
Beginning bigram	*<G*	0.45	-0.04
Current POS tag	NNP	0.47	0.45
Prev and cur tags	IN NNP	-0.10	0.14
Previous state	Other	-0.70	-0.92
Current signature	Xx	0.80	0.46
Prev state, cur sig	O-Xx	0.68	0.37
Prev-cur-next sig	x-Xx-Xx	-0.69	0.37
P. state - p-cur sig	O-x-Xx	-0.20	0.82
…			
Total:		**-0.58**	**2.68**

Max Likelihood vs. Max Entropy

- The probability distribution found by maximizing entropy is the distribution with least KL divergence from uniform distribution.
  - KL(p||q) = $\sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$

The probability distribution found by maximizing log-likelihood is the distribution with least KL divergence from frequency distribution.

Now

- bring frequencies in (1)
- bring uniformity in (2)

Duality of MaxLL and MaxEnt

- Theorem: MaxLL and MaxEnt are the same [Berger 96]
  - Distributions belong to exponential family
  - Distance measure = KL divergence