Machine Learning 10-601 10-301

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

March 31, 2021

Today:
- Probabilistic learning
- Joint probabilities
- Estimating parameters
  - MLE
  - MAP

Required Reading:
- Estimating Probabilities [Mitchell]

Optional Probability Review:
- Goodfellow, Ch 3-3.9

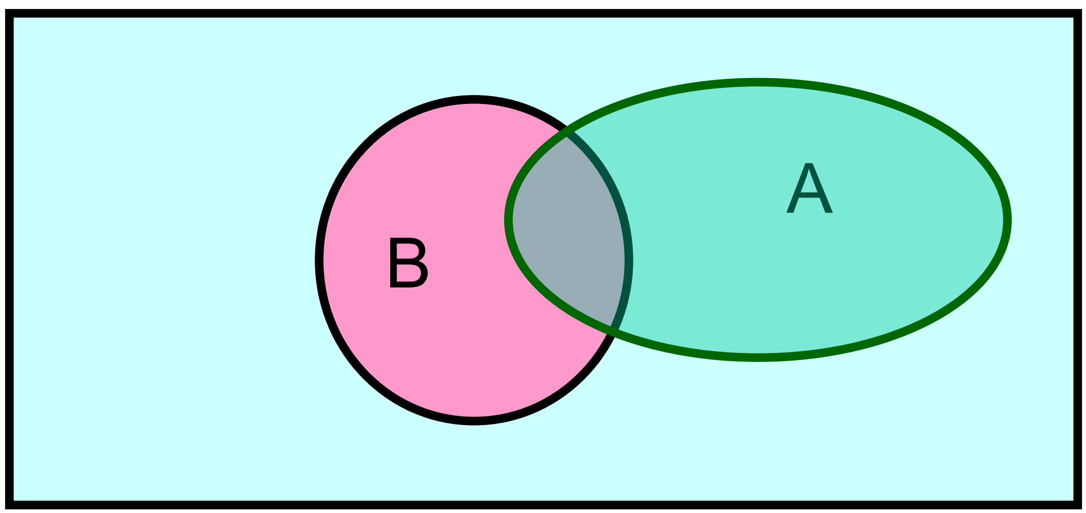
some of these slides are derived from
William Cohen, Andrew Moore, Aarti Singh, Eric Xing, Carlos Guestrin.
- Thanks!

probabilistic function approximation:

instead of $F: X \rightarrow Y$,
learn $P(Y \mid X)$

Definition of Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$



Definition of Conditional Probability

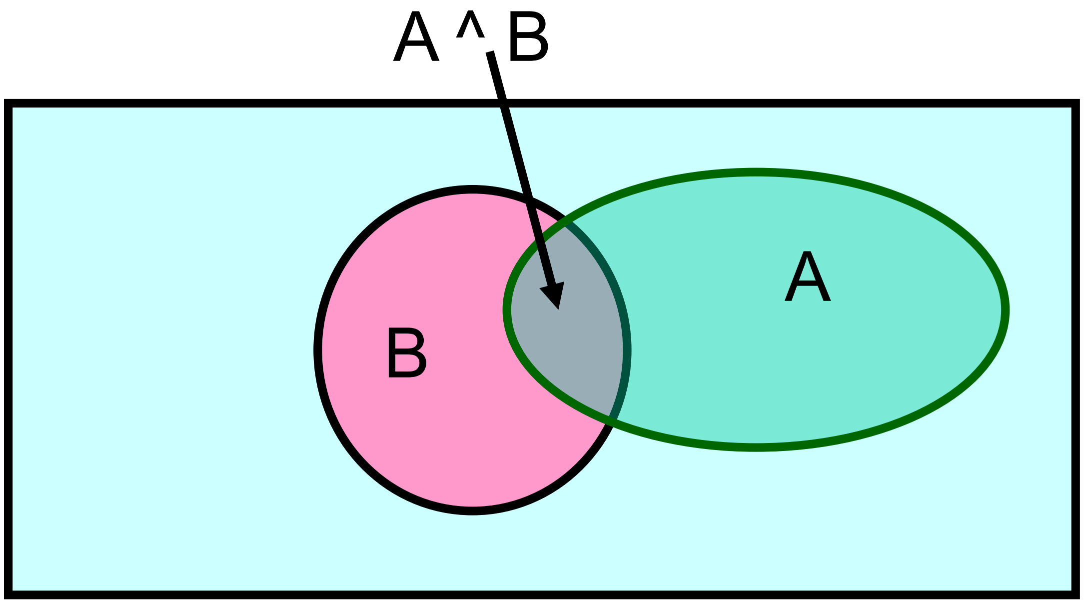
$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\, P(B)$$

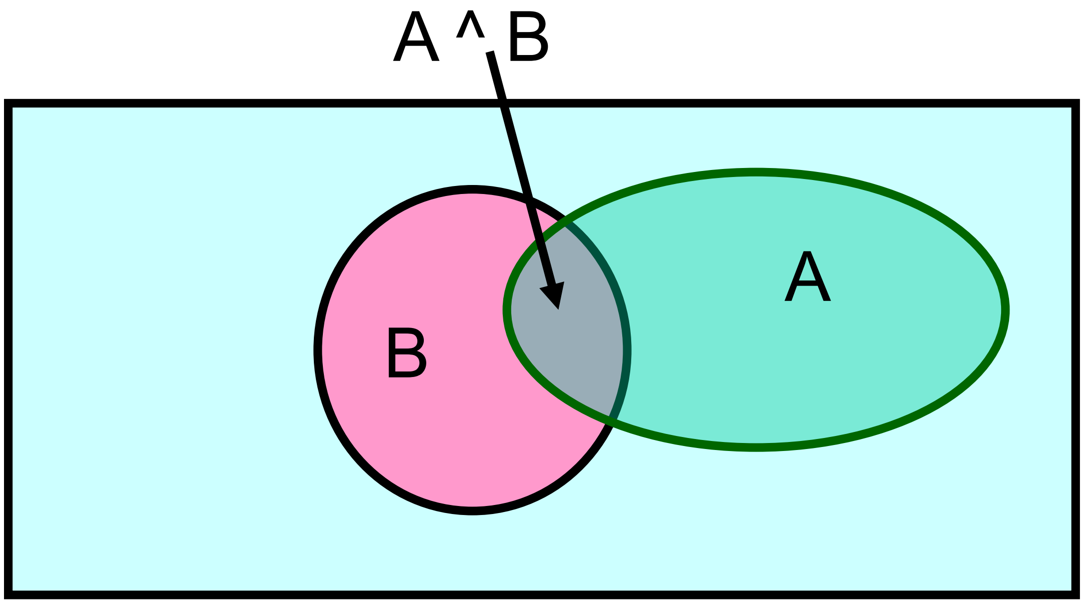
Bayes Rule

- let's write 2 expressions for P(A ^ B)



Bayes Rule

- let's write 2 expressions for P(A ^ B)



$$P(A \wedge B) = P(A|B)P(B) = P(B|A)\ P(B)$$

implies:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes’ rule

we call P(A) the “prior”

and P(A|B) the “posterior”

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Other Forms of Bayes Rule

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

$$P(A|B \wedge X) = \frac{P(B \mid A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

Applying Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

A = you have the flu,   B = you just coughed

Assume:
P(A) = 0.05
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

The Awesome Joint Probability Distribution $P(X_1, X_2, \ldots X_N)$

from which we can calculate
$P(X_1|X_2 \ldots X_N)$,
and <u>every</u> other probability we desire
over subsets of $X_1 \ldots X_N$

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	**B**	**C**
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

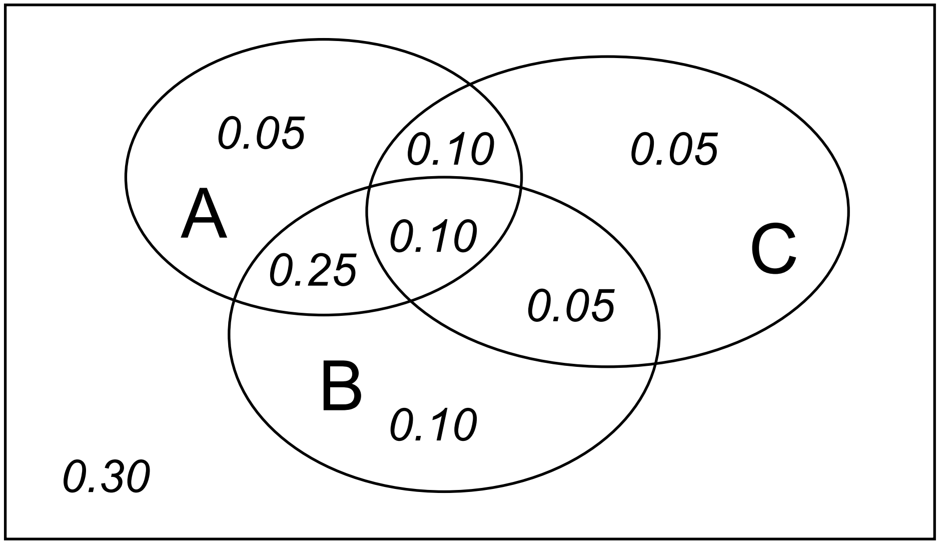
A	**B**	**C**	**Prob**
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Once you have the JD you can ask for the probability of **any** logical expression involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Once you have the JD you can ask for the probability of **any** logical expression involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Poll question 1:

What is P(rich, female)?

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Poll question 2:

What is P(female | poor, v0:40.5)

Learning and the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

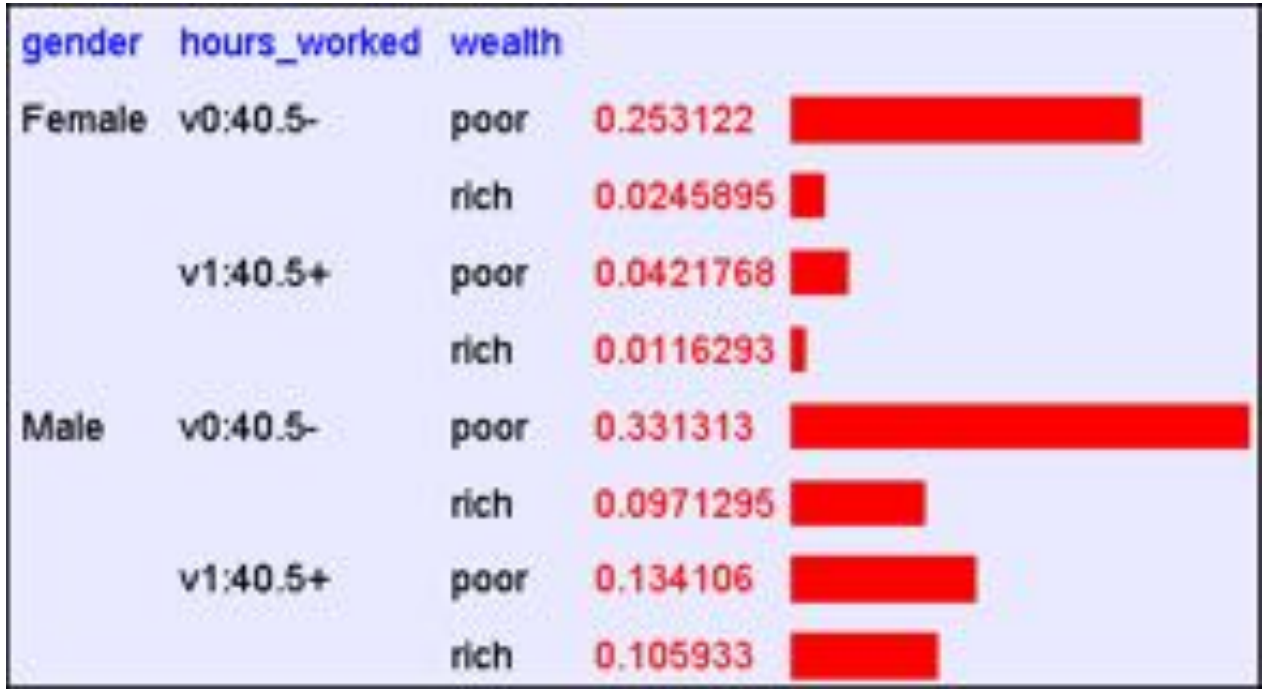
Suppose we want to learn the function f: <G, H> → W

Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) =

Learning and the Joint Distribution



gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Suppose we want to learn the function f: <G, H> → W

Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) = 0.024 / (0.024 + 0.253)
= 0.087

sounds like the solution to learning $F: X \rightarrow Y$, or $P(Y \mid X)$.

Are we done?

sounds like the solution to learning $F: X \rightarrow Y$, or $P(Y \mid X)$.

Main problem: learning $P(Y|X)$ can require more data than we have

consider learning Joint Dist. with 100 attributes

# of rows in this table?

# of people on earth?

sounds like the solution to
learning F: X →Y,
or P(Y | X).

Main problem: learning P(Y|X)
can require more data than we have

consider learning Joint Dist. with 100 attributes
# of rows in this table? $2^{100} > 10^{30}$
# of people on earth? 10^{10}
fraction of rows with 0 training examples? 99.99

What to do?

1. Be smart about how we estimate probabilities from sparse data
   – maximum likelihood estimates
   – maximum a posteriori estimates

2. Be smart about how to represent joint distributions
   – Bayes networks, graphical models, conditional independencies

1. Be smart about how we estimate probabilities

Estimating Probability of Heads

- I show you the above coin X, and ask you to estimate the probability that it will turn up heads (X=1) or tails (X=0)

- You flip it repeatedly, observing
  - it turns up heads α_1 times
  - it turns up tails α_0 times

- Your estimate for $\hat{\theta} = \hat{P}(X = 1)$ is …?

Estimating Probability of Heads

X=1 X=0

- I show you the above coin X, and ask you to estimate the probability that it will turn up heads (X=1) or tails (X=0)

- You flip it repeatedly, observing
  - it turns up heads α_1 times
  - it turns up tails α_0 times

Algorithm 1 (MLE): $\hat{\theta} = \hat{P}(X=1) = \dfrac{\alpha_1}{\alpha_1 + \alpha_0}$

Estimating $\theta = P(X=1)$

X=1 X=0

Test A:

100 flips: 51 Heads, 49 Tails

Test B:

3 flips: 2 Heads, 1 Tails

Estimating Probability of Heads


X=1 X=0

When data sparse, might bring in prior assumptions to bias our estimate

- e.g., represent priors by "hallucinating" γ_1 heads, and γ_0 tails, to complement sparse observed α_1, α_0

Alg 2 (MAP): $\hat{\theta} = \hat{P}(X=1) = \frac{(\alpha_1 + \gamma_1)}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$

Estimating Probability of Heads

X=1 X=0

When data sparse, might bring in prior assumptions to bias our estimate

- e.g., represent priors by "hallucinating" γ_1 heads, and γ_0 tails, to complement sparse observed α_1, α_0

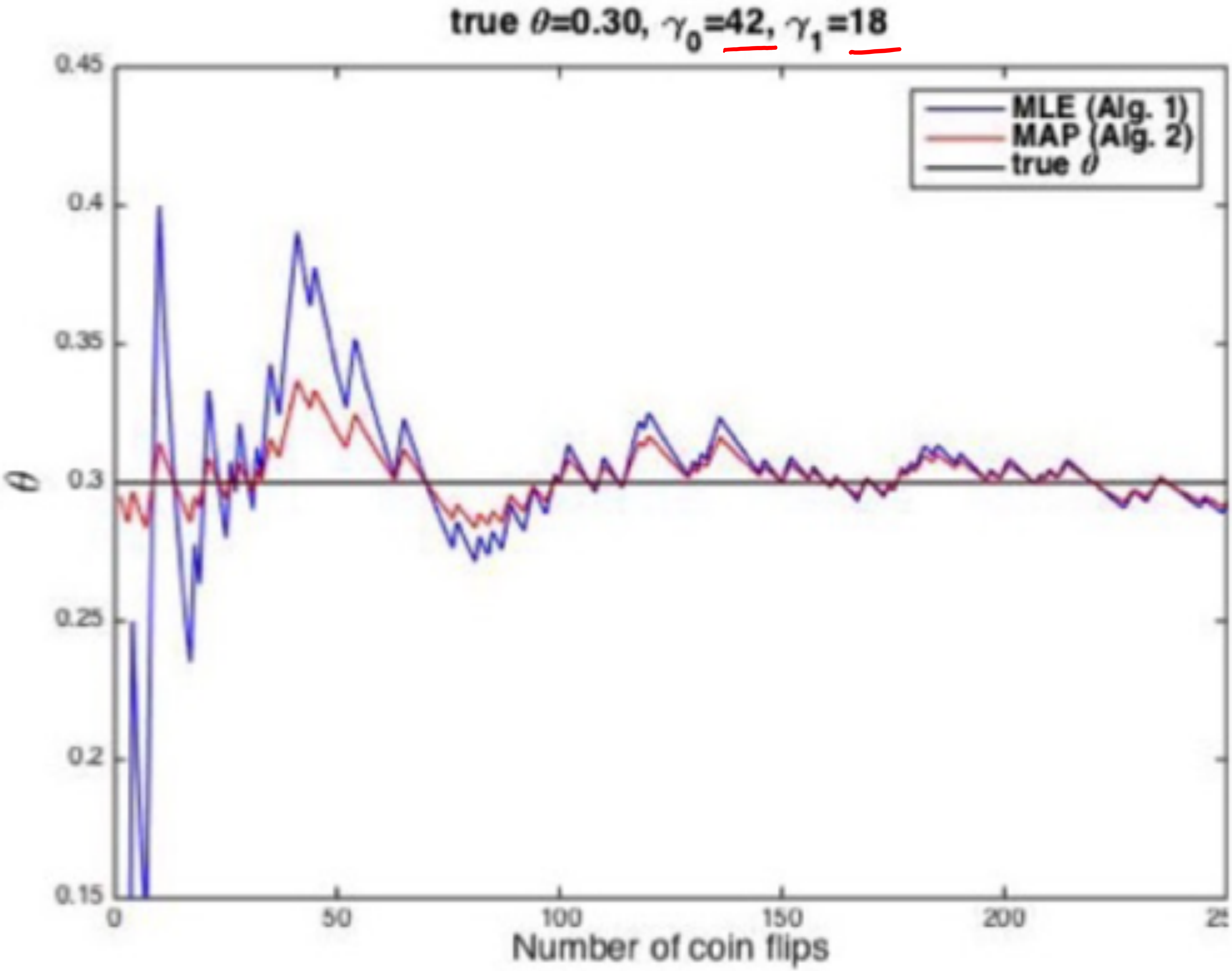
Alg 2 (MAP): $\hat{\theta} = \hat{P}(X=1) = \dfrac{(\alpha_1 + \gamma_1)}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$

Consider $\gamma_1 = 1 \;\; \gamma_0 = 1$

versus $\gamma_1 = 1000 \;\; \gamma_0 = 1000$

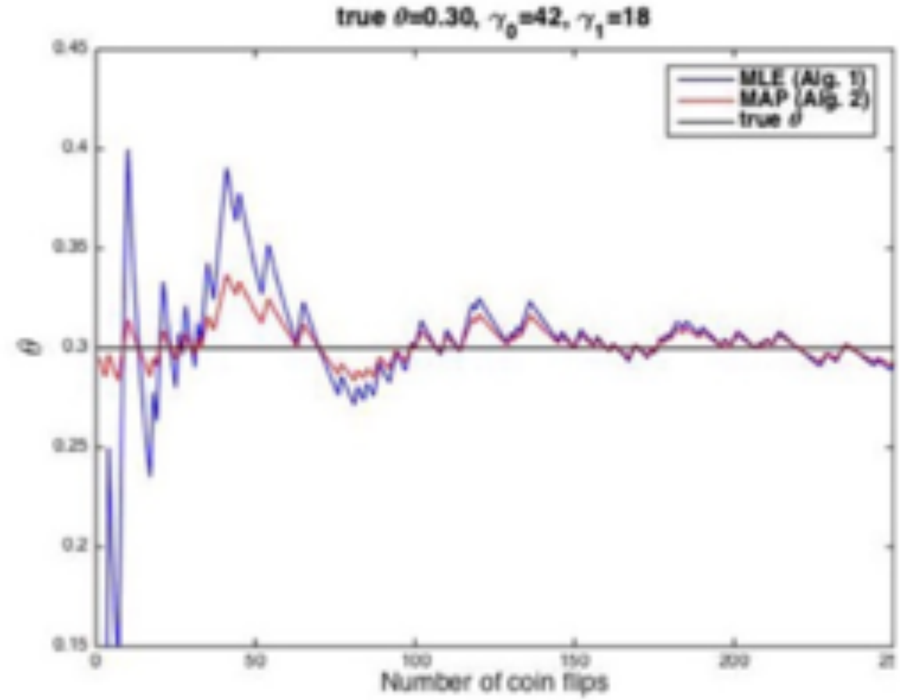
versus $\gamma_1 = 500 \;\; \gamma_0 = 1500$

Correct MAP Priors

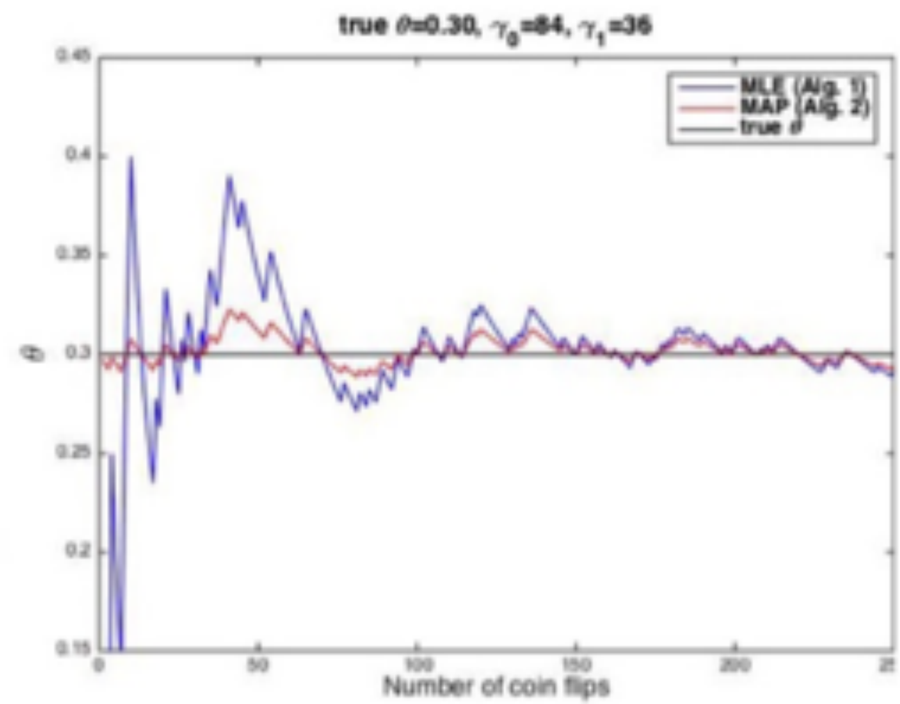


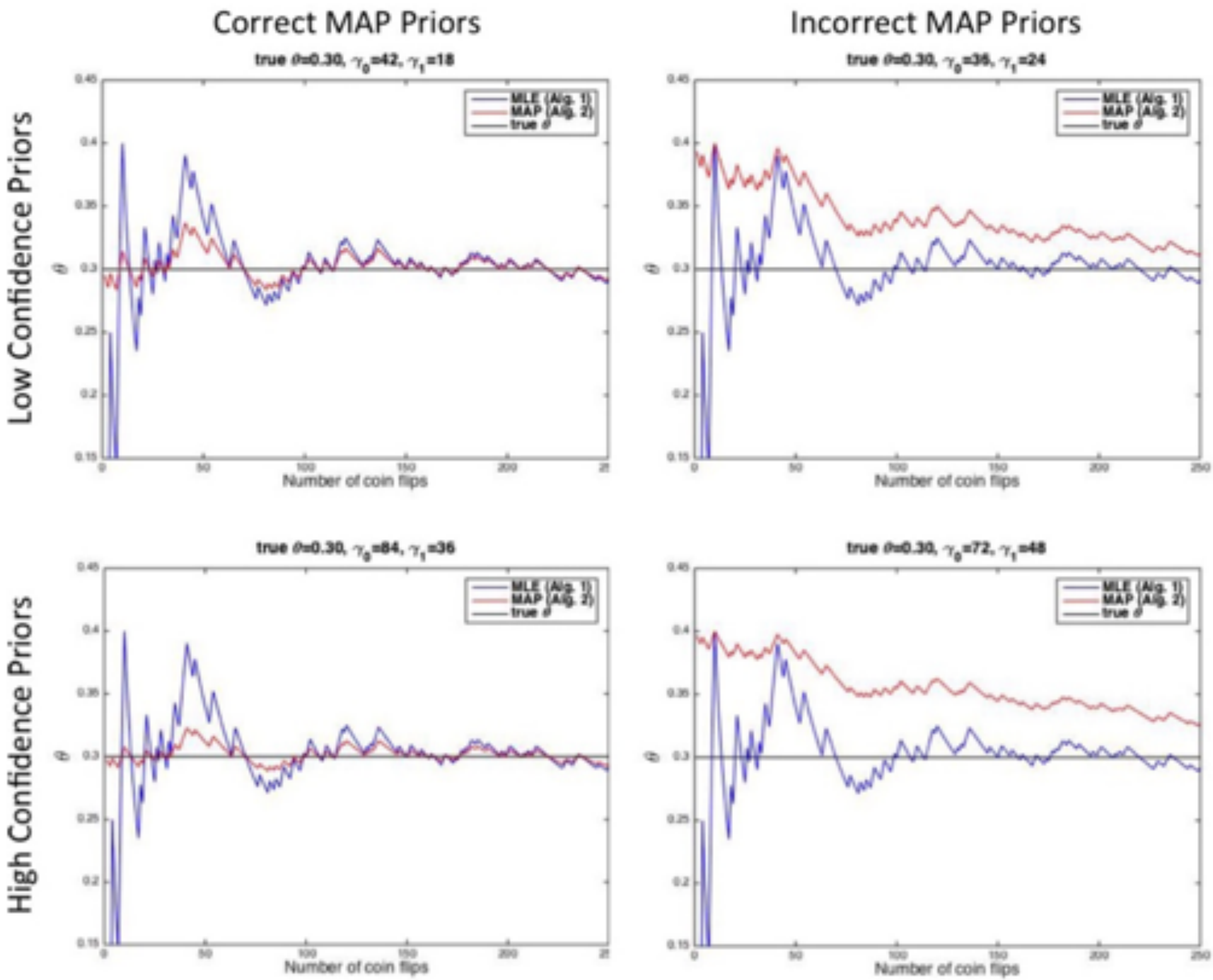
Correct MAP Priors

Low Confidence Priors



High Confidence Priors




Correct MAP Priors
Incorrect MAP Priors
Low Confidence Priors
High Confidence Priors
true θ=0.30, γ0=42, γ1=18
true θ=0.30, γ0=36, γ1=24
true θ=0.30, γ0=84, γ1=36
true θ=0.30, γ0=72, γ1=48
MLE (Alg. 1)
MAP (Alg. 2)
true θ
θ
Number of coin flips

Principles for Estimating Probabilities

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and observed data

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

Principles for Estimating Probabilities

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and observed data

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} P(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \\ &= \arg\max_{\theta} P(\mathcal{D} \mid \theta)P(\theta)\end{aligned}$$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize **P(data | θ)**
- result in our case: $$\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Principle 2 (maximum a posteriori probability):

- choose parameters θ that maximize **P(θ | data)**
- result in our case:

$$\hat{\theta}^{MAP} = \frac{\alpha_1 + \#\text{hallucinated\_1s}}{(\alpha_1 + \#\text{hallucinated\_1s}) + (\alpha_0 + \#\text{hallucinated\_0s})}$$

Maximum Likelihood Estimation

given data D, choose θ that maximizes P(D | θ)

X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

Data D:

$$P(D|\theta) =$$

$$\hat{\theta} = \arg\max_{\theta} \ \ln P(D|\theta)$$

$$= \arg\max_{\theta} \ \ln \left[\theta^{\alpha_1}(1-\theta)^{\alpha_0}\right]$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ \ln P(\mathcal{D} \mid \theta) = 0}$

hint: $\frac{\partial \ln \theta}{\partial \theta} = \frac{1}{\theta}$

Summary:
Maximum Likelihood Estimate
for Bernoulli random variable

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros (Binomial)

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \text{argmax}_\theta \, P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize $P(\text{data} \mid \theta)$

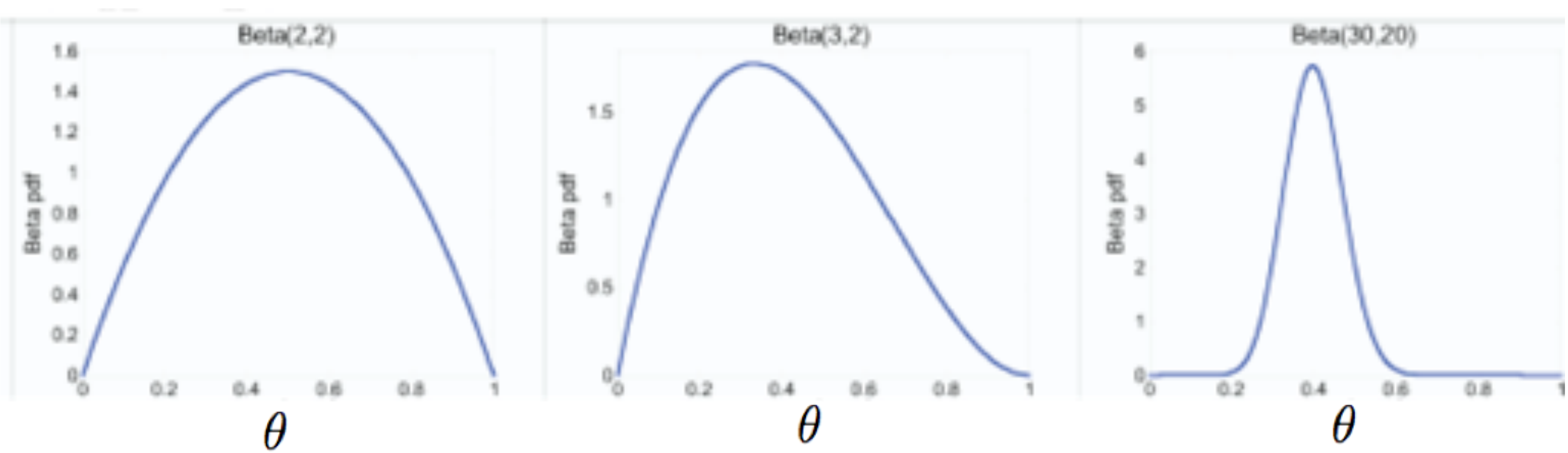
Principle 2 (maximum a posteriori prob.):

- choose parameters θ that maximize

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)\, P(\theta)}{P(\text{data})}$$

Beta prior distribution – P(θ)

- $$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



Summary:
Maximum a Posteriori (MAP) Estimate for Bernoulli random variable

X=1 X=0

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)} \sim Beta(\beta_H,\beta_T)$$

Then posterior is Beta distribution

$$P(\theta|D) \propto P(D|\theta)P(\theta) \sim Beta(\alpha_H+\beta_H, \alpha_T+\beta_T)$$

and MAP estimate is therefore

$$\hat{\theta}^{MAP} = \frac{\alpha_H+\beta_H-1}{(\alpha_H+\beta_H-1)+(\alpha_T+\beta_T-1)}$$

Maximum a Posteriori (MAP) Estimate for random variable with k possible outcomes

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2}\ldots\theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1-1}\ \theta_2^{\beta_2-1}\ \ldots\theta_k^{\beta_k-1}}{B(\beta_1,\ldots,\beta_k)} \sim \mathrm{Dirichlet}(\beta_1,\ldots,\beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \propto P(D|\theta)P(\theta) \sim \mathrm{Dirichlet}(\alpha_1+\beta_1,\ldots,\alpha_k+\beta_k)$$

and MAP estimate is therefore

$$\hat{\theta}_i^{MAP} = \frac{\alpha_i+\beta_i-1}{\sum_{j=1}^{k}(\alpha_j+\beta_j-1)}$$

Some terminology

- Likelihood function: $P(\text{data} \mid \theta)$
- Prior: $P(\theta)$
- Posterior: $P(\theta \mid \text{data})$

- Conjugate prior: $P(\theta)$ is the conjugate prior for likelihood function $P(\text{data} \mid \theta)$ if the parametric forms of $P(\theta)$ and $P(\theta \mid \text{data})$ are the same.
  - Beta is conjugate prior for Bernoulli, Binomial
  - Dirichlet is conjugate prior for Multinomial

You should know

- Probability basics
  - random variables, conditional probs, …
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution
- Estimating parameters from data
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – Bernoulli, Binomial, Beta, Dirichlet, …
  - conjugate priors
  - regularization is a form of MAP estimation

Extra slides

Independent Events

- Definition: two events A and B are *independent* if $P(A \wedge B)=P(A)\,P(B)$
- Intuition: knowing A tells us nothing about the value of B (and vice versa)

Expected values

Given a discrete random variable X, the expected value of X, written E[X] is

Probability-weighted average over all possible values of X

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

Example:

X	P(X)
0	0.3
1	0.2
2	0.5

$$E[X] =$$

Expected values

Given discrete random variable X, the expected value of X, written E[X] is

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

We also can talk about the expected value of functions of X

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) P(X = x)$$

Covariance

Given two discrete r.v.'s X and Y, we define the covariance of X and Y as

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

e.g., X=GENDER, Y=PLAYS_FOOTBALL

or X=GENDER, Y=LEFT_HANDED

Remember:

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

NAÏVE BAYES

Let’s learn classifiers by learning P(Y|X)

Consider Y=Wealth, X=<Gender, HoursWorked>

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How many parameters must we estimate?

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

Suppose $X = \langle X_1, \ldots X_n \rangle$
where X_i and Y are boolean RV's

To estimate $P(Y | X_1, X_2, \ldots X_n)$

If we have 100 boolean X_i's: $P(Y | X_1, X_2, \ldots X_{100})$

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

Equivalently:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

Can we reduce params using Bayes Rule?

Suppose $X = \langle X_1, \ldots X_n \rangle$
where X_i and Y are boolean RV's

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

How many parameters to define $P(X_1, \ldots X_n \mid Y)$?

How many parameters to define $P(Y)$?

Naïve Bayes

Naïve Bayes assumes

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

i.e., that X_i and X_j are conditionally independent given Y, for all $i \neq j$

Conditional Independence

Definition: X is <u>conditionally independent</u> of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write

$$P(X|Y,Z) = P(X|Z)$$

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$P(X_1, X_2|Y) =$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

in general: $P(X_1 ... X_n|Y) = \prod_i P(X_i|Y)$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$
\begin{aligned}
P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\
&= P(X_1|Y)P(X_2|Y)
\end{aligned}
$$

in general: $P(X_1...X_n|Y) = \prod_i P(X_i|Y)$

How many parameters to describe $P(X_1...X_n|Y)$? $P(Y)$?

- Without conditional indep assumption?
- With conditional indep assumption?

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}$$

Assuming conditional independence among X_i' s:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, to pick most probable Y for $X^{new} = < X_1, \ldots, X_n >$

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$