



Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 21, 2015

Today:

- Bayes Rule
- Estimating parameters
 - MLE
 - MAP

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

Readings:

Probability review

- Bishop Ch. 1 thru 1.2.3
- Bishop, Ch. 2 thru 2.2
- Andrew Moore's online tutorial

Announcements

- Class is using Piazza for questions/discussions about homeworks, etc.
 - see class website for Piazza address
 - <http://www.cs.cmu.edu/~ninamf/courses/601sp15/>
- Recitations thursdays 7-8pm, Wean 5409
 - videos for future recitations (class website)
- HW1 was accepted to Sunday 5pm for full credit
- HW2 out today on class website, due in 1 week
- HW3 will involve programming (in Octave)

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



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**

we call $P(A)$ the “prior”
and $P(A|B)$ the “posterior”

...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

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

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

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

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:

$$P(A) = 0.05$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.20$$

what is $P(\text{flu} | \text{cough}) = P(A|B)$?

$$P(A|B) = \frac{0.8 \cdot 0.05}{0.05 \cdot 0.8 + 0.95 \cdot 0.20} =$$

what does all this have to do with
function approximation?

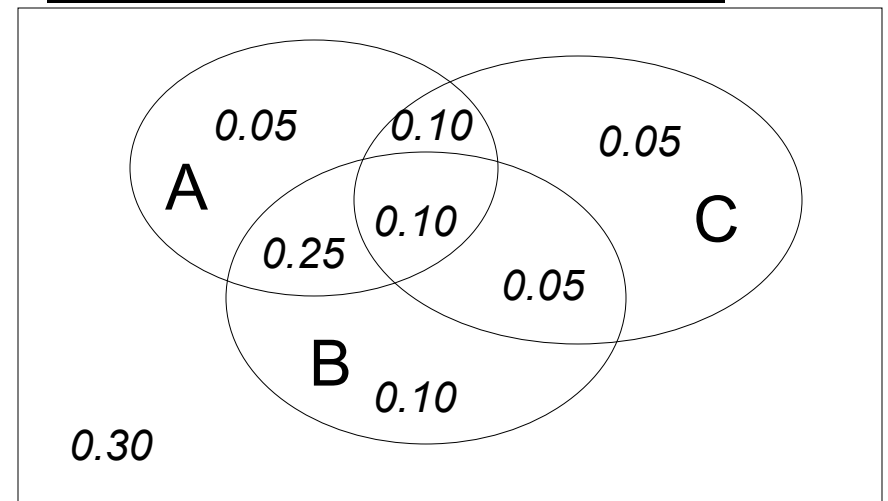
instead of $F: X \rightarrow Y$,
learn $P(Y | X)$

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

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



[A. Moore]

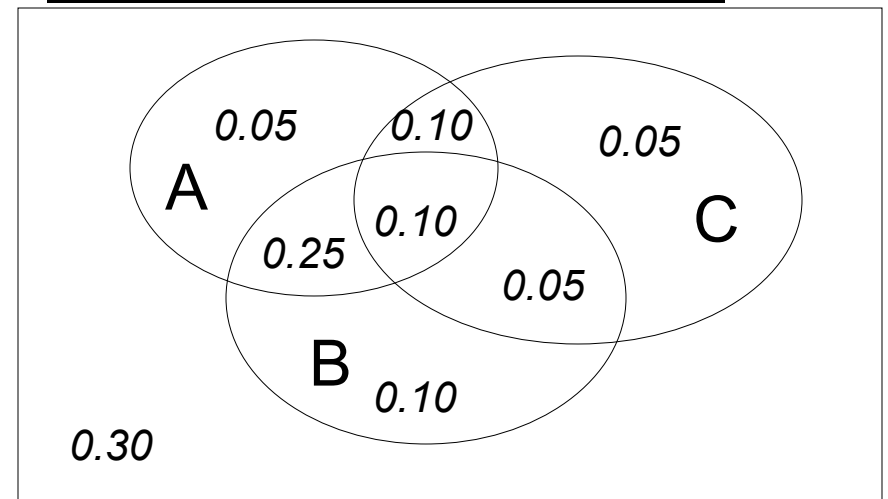
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values (M Boolean variables $\rightarrow 2^M$ rows).

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



[A. Moore]

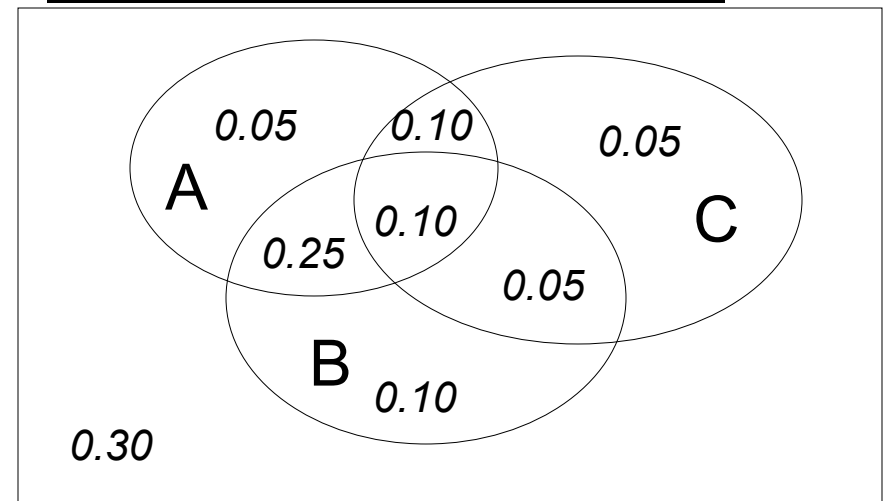
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values (M Boolean variables $\rightarrow 2^M$ 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



[A. Moore]

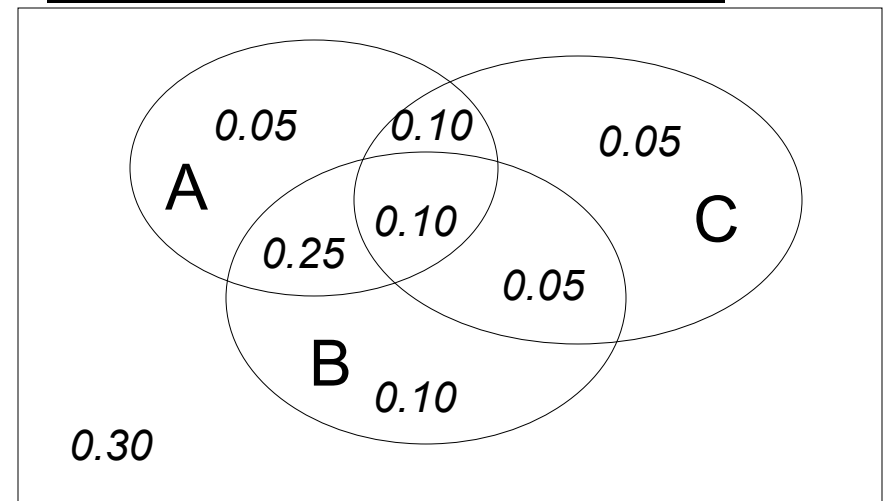
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:









1. Make a truth table listing all combinations of values (M Boolean variables $\rightarrow 2^M$ rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those probabilities 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



[A. Moore]









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 

$$P(\text{Poor Male}) = 0.4654$$

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

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

$$P(\text{Poor}) = 0.7604$$

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

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

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






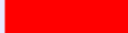


$$P(\text{Male} | \text{Poor}) = 0.4654 / 0.7604 = 0.612$$

$$P(F|R) = 1 - P(M|R)$$

$$P(\wedge | R)$$

[A. Moore]

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 $f: \langle G, H \rangle \rightarrow W$

Equivalently, $P(W | G, H)$

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

e.g., $P(W=\text{rich} | G = \text{female}, H = 40.5-) = \frac{P(F, H, R) \cdot P(R)}{P(F, H)}$

$$P(F, H)$$

[A. Moore]

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

Are we done?

sounds like the solution to
learning $F: X \rightarrow Y$,
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?

of people on earth?

fraction of rows with 0 training examples?

$$2^{100} = 2^{10} \cdot 2^{90} \approx 10^3 \cdot 10^{27} = 10^{30}$$
$$10^4 < 10^{30}$$

What to do?

1. Be smart about how we estimate probabilities from sparse data
 - maximum likelihood estimates $\propto L$
 - maximum a posteriori estimates $\propto AP$
2. Be smart about how to represent joint distributions
 - Bayes networks, graphical models

1. Be smart about how we estimate probabilities

Estimating Probability of Heads



- I show you the above coin X , and hire 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 $P(X = 1)$ is....?

$$\frac{\alpha_1}{\alpha_1 + \alpha_0}$$

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



X=1 X=0

Test A:

100 flips: 51 Heads ($X=1$), 49 Tails ($X=0$)

$$\hat{p}(X=1) = \frac{51}{100} = 0.51$$

Test B:

3 flips: 2 Heads ($X=1$), 1 Tails ($X=0$)

$$\hat{p}(X=1) = \frac{2}{3} \approx 0.667$$

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



Case C: (online learning)

- keep flipping, want single learning algorithm that gives reasonable estimate after each flip

$$\frac{\alpha_1}{\alpha_1 + \alpha_0}$$

α_1 : #heads
 α_0 : #Tails

prior 1/5

1. $\frac{1}{n} \cdot \frac{1}{2} + \frac{n-1}{n} \cdot \frac{\alpha_1}{\alpha_1 + \alpha_0}$

2. $\frac{\alpha_1 + 1}{(\alpha_1 + 1) + (\alpha_0 + 1)}$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

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

- e.g.,
$$\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Principle 2 (maximum a posteriori prob.):

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

- e.g.

$$\hat{\theta}^{MAP} = \frac{\alpha_1 + \# \text{hallucinated_1s}}{(\alpha_1 + \# \text{hallucinated_1s}) + (\alpha_0 + \# \text{hallucinated_0s})}$$

Maximum Likelihood Estimation

$$P(X=1) = \theta \quad P(X=0) = (1-\theta)$$



Data D: 1 1 0 0 1

$$P(D|\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta = \theta^3 (1-\theta)^2$$

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

Flips produce data D with α_1 heads, α_0 tails

- flips are independent, identically distributed 1's and 0's (Bernoulli)
- α_1 and α_0 are counts that sum these outcomes (Binomial)

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

Maximum Likelihood Estimate for Θ

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \ln P(\mathcal{D} | \theta) \\ &= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}\end{aligned}$$

$\max_{\theta} P(\text{cluster} | \text{model})$

$\{(\theta)\}$

- Set derivative to zero:

$$\frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0$$

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

■ Set derivative to zero:

$$\frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0$$

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

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

$$\ell(\theta) = \alpha_1 \ln \theta + \alpha_0 \ln (1 - \theta)$$

$$\frac{\partial \ell}{\partial \theta} = \alpha_1 \frac{1}{\theta} + \alpha_0 \left[\frac{\partial \ln (1 - \theta)}{\partial \theta} \right] = \frac{\alpha_1}{\theta} - \frac{\alpha_0}{1 - \theta} = 0$$

$$\hat{\theta} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Summary:

Maximum Likelihood Estimate



$X=1$ $X=0$

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

$$P(X=0) = 1-\theta$$

(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)

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

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

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

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

$$P(\text{model} \mid \text{data})$$

Beta prior distribution – $P(\theta)$

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

- Likelihood function:** $P(\mathcal{D} | \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

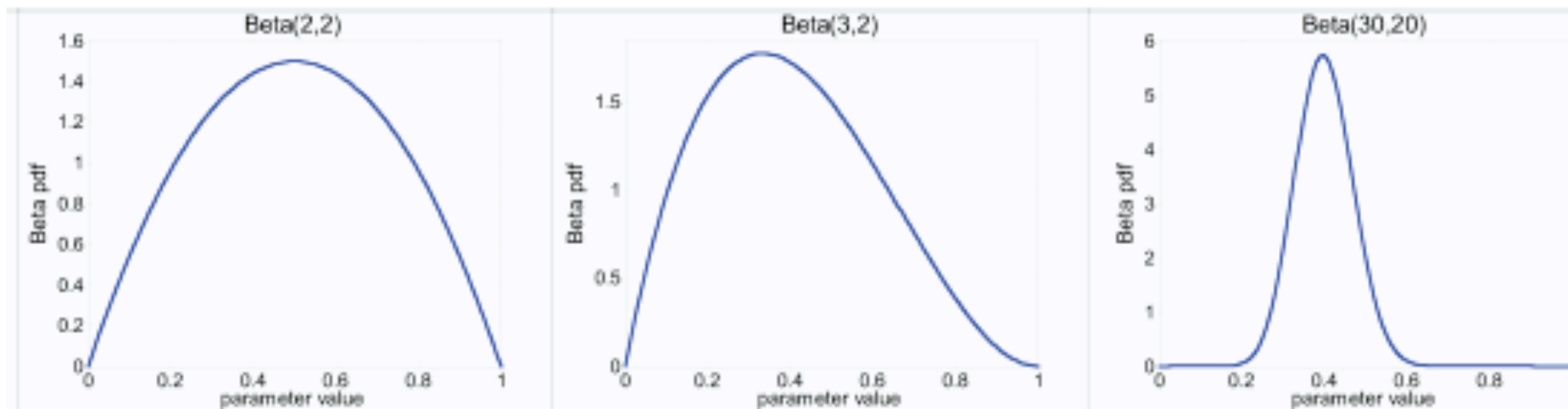
- Posterior:** $P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$

$$\text{numerator} = \theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_T - 1}$$

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

Beta prior distribution – $P(\theta)$

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



Eg. 1 Coin flip problem

Likelihood is \sim Binomial

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

If prior is Beta distribution,

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

Then posterior is Beta distribution

$$P(\theta \mid D) \sim \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

and MAP estimate is therefore

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



Eg. 2 Dice roll problem (6 outcomes instead of 2)



Likelihood is $\sim \text{Multinomial}(\theta = \{\theta_1, \theta_2, \dots, \theta_k\})$

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

If prior is Dirichlet distribution,

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

Then posterior is Dirichlet distribution

$$P(\theta | D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \dots, \beta_k + \alpha_k)$$

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 forms of $P(\theta)$ and $P(\theta \mid \text{data})$ are the same.

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 – binomial, Beta, Dirichlet, ...
 - conjugate priors

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)

Picture “A independent of B”

Expected values

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

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

Example:

x	$P(X)$
0	0.3
1	0.2
2	0.5

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

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

e.g., $X=\text{gender}$, $Y=\text{playsFootball}$

or $X=\text{gender}$, $Y=\text{leftHanded}$

Remember: $E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$