Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 21, 2015

Today:
- Bayes Rule
- Estimating parameters
  - MLE
  - MAP

Readings:

Machine Learning (ML), Ch. 2

Probability review:

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

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

$P(A,B)$

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

…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

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

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

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

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

$\cancel{P(U|X,B)}$

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) = 0.20

what is P(flu | cough) = P(A|B)? $= \frac{\overset{0.80}{P(B|A)}\,\overset{0.05}{P(A)}}{P(B)}$

$P(B) = \overset{0.80}{P(B|A)}\,\overset{0.05}{P(A)} + \overset{0.20}{P(B|\sim A)}\,\overset{0.95}{P(\sim A)}$

$f : X \to Y$. $\qquad f(x) = x^T\beta$. $\qquad Y = X^T\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$

or $\quad P(Y|X) \qquad \hat{G}(x) = \arg\max_g P_r(G=g|X=x) \qquad$ LR : $P(Y|X) = N(X^T\beta, \Sigma)$, MLE $\to \beta$

what does all this have to do with function approximation?

instead of F: X → Y,
learn P(Y | X)

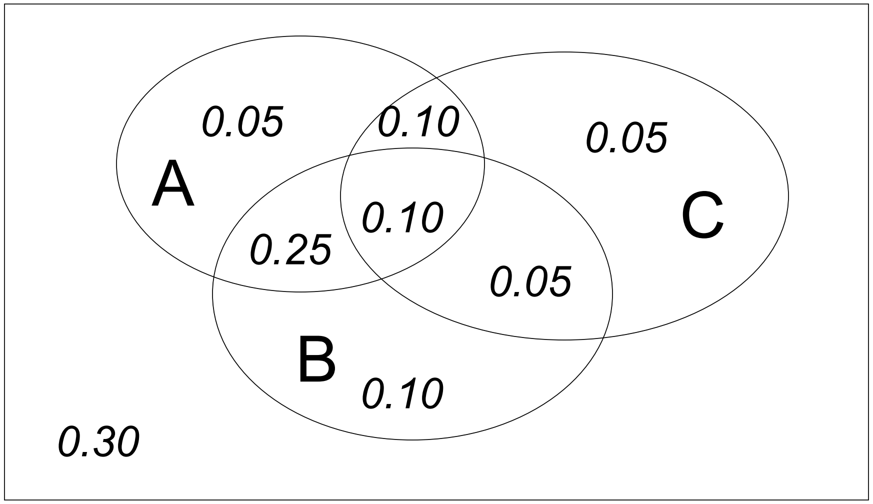
The Joint Distribution

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

Pr(A,B,C)



[A. Moore]

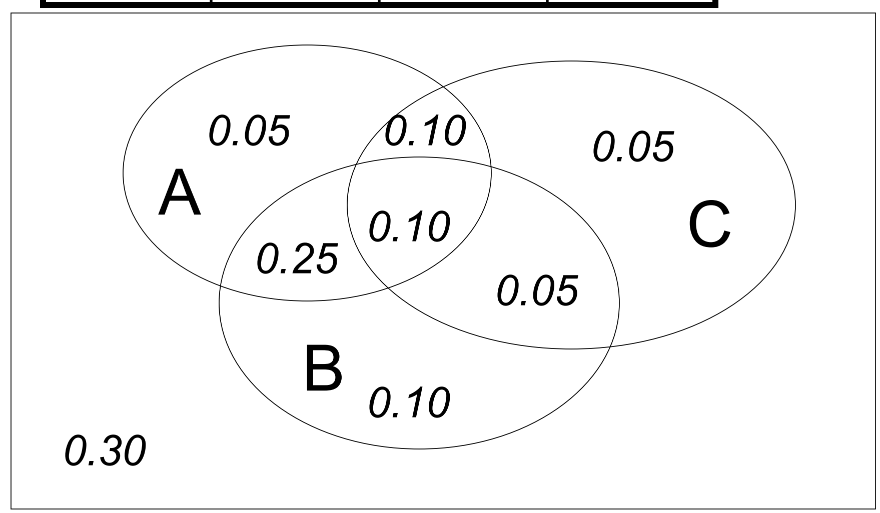
The Joint Distribution

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 → 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

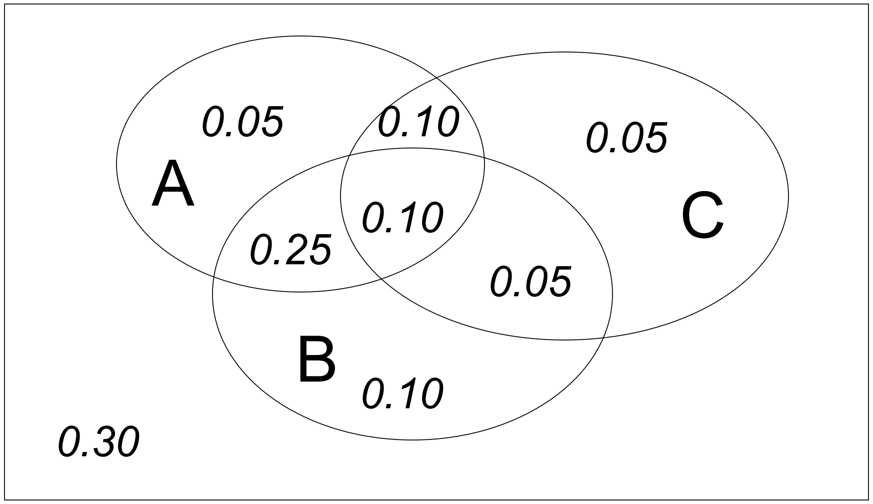
Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values (M Boolean variables → 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

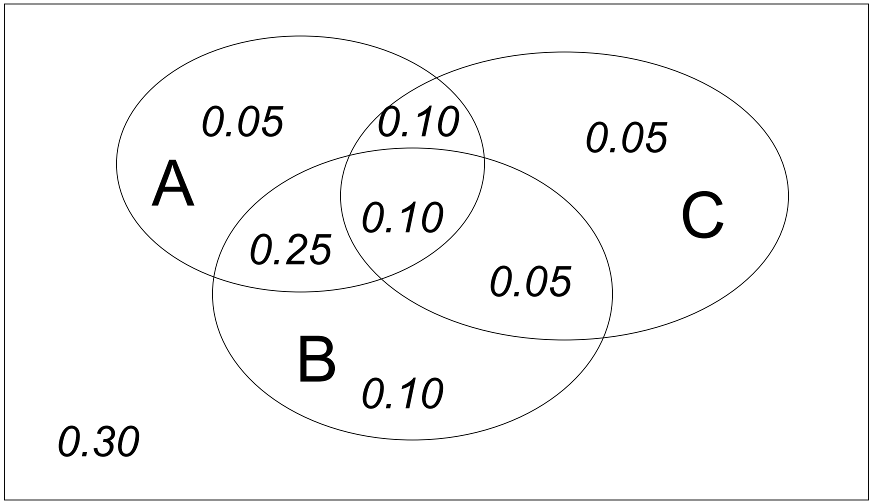
Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values (M Boolean variables → 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

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

[A. Moore]

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

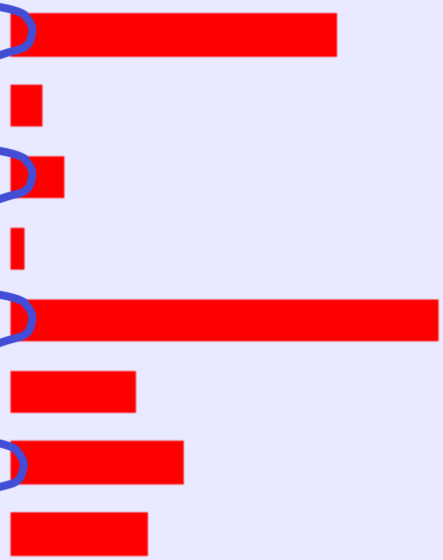
P(Poor Male) = 0.4654

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

[A. Moore]

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

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

[A. Moore]

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

$$= \frac{P(\text{Male}, \text{Poor})}{P(\text{Poor})}$$

[A. Moore]

Learning and the Joint Distribution

G gender	H hours_worked	W 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-) = $\frac{P(W=rich, G=f, H=40.5-)}{P(G=f, H=40.5-)}$

[A. Moore]

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 → Y, or P(Y | X).

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

JDT

consider learning Joint Dist. with 100 attributes

\# of rows in this table? $2^{100} \approx (10^3)^{10} = 10^{30}$

\# of people on earth? $10^9 << 10^{30}$

fraction of rows with 0 training examples?

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

1. Be smart about how we estimate probabilities

Estimating Probability of Heads

X=1 X=0

- 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....?

MLE

$$P(X=1) = \frac{\alpha_1}{n} = \frac{\alpha_1}{\alpha_1+\alpha_0}$$

Estimating θ = P(X=1)

X=1 X=0

Test A:

$\alpha_1 = 51$ $\alpha_0 = 49$

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

$$P(X=1) = \frac{51}{51+49} = 0.51$$

Test B:

$\alpha_1 = 2$ $\alpha_0 = 1$

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

$$P(X=1) = \frac{2}{3} \approx 66.7\%$$

Estimating θ = P(X=1)

X=1 X=0

Case C: (online learning)

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

Prior : $\hat{P}(X=1) = \frac{1}{2} = 50\%$

α_1 : # heads

α_0 : # tails.

$P(X=1) = \frac{\alpha_1}{\alpha_1+\alpha_0} = \frac{\alpha_1}{n}$ ← MLE

① $P(X=1) = \frac{1}{n}\cdot\frac{1}{2} + (1-\frac{1}{n})\frac{\alpha_1}{\alpha_1+\alpha_0}$

larger → confidence.

⇒ ② $P(X=1) = \frac{\alpha_1+10}{(\alpha_1+10)+(\alpha_0+10)}$. ← MAP

$(10+10=20)$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood): MLE

- choose parameters θ that maximize **P(data | θ)** D
- e.g., $$\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Principle 2 (maximum a posteriori prob.): MAP

- choose parameters θ that maximize **P(θ | data)** $\propto P(\text{data}|\theta)P(\theta)$
  $P(\text{data})$
- e.g.

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

$P(x|\theta) = \theta^{x}(1-\theta)^{1-x}$. $p(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$

Maximum Likelihood Estimation

X=1 X=0

P(X=1) = θ P(X=0) = (1-θ)

likelihood

Data D: $\{1, 0, 0, 1, 0\}$, $P(D|\theta) = P(\{1,0,0,1,0\}|\theta) = \theta^{2}(1-\theta)^{3}$

(i.i.d) $\theta, 1-\theta, 1-\theta, \theta, 1-\theta$

#heads: α_1,

#tail: α_0, $\max_{\theta} P(D|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$

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

Prob. → log-concave

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

← concave

$$= \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

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

[C. Guestrin]

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

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

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

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

$$\frac{\partial \ln(1-\theta)}{\partial (1-\theta)} \cdot \frac{\partial(1-\theta)}{\partial \theta}$$

$$\frac{1}{1-\theta} \cdot (-1)$$

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

$$\frac{\partial Q}{\partial \theta} = \alpha_1 \cdot \frac{1}{\theta} + \alpha_0 (-1) \cdot \frac{1}{1-\theta} = 0$$

$$\Rightarrow \boxed{\theta = \frac{\alpha_1}{\alpha_1 + \alpha_0}}$$

Summary:
Maximum Likelihood Estimate

X=1 X=0

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)

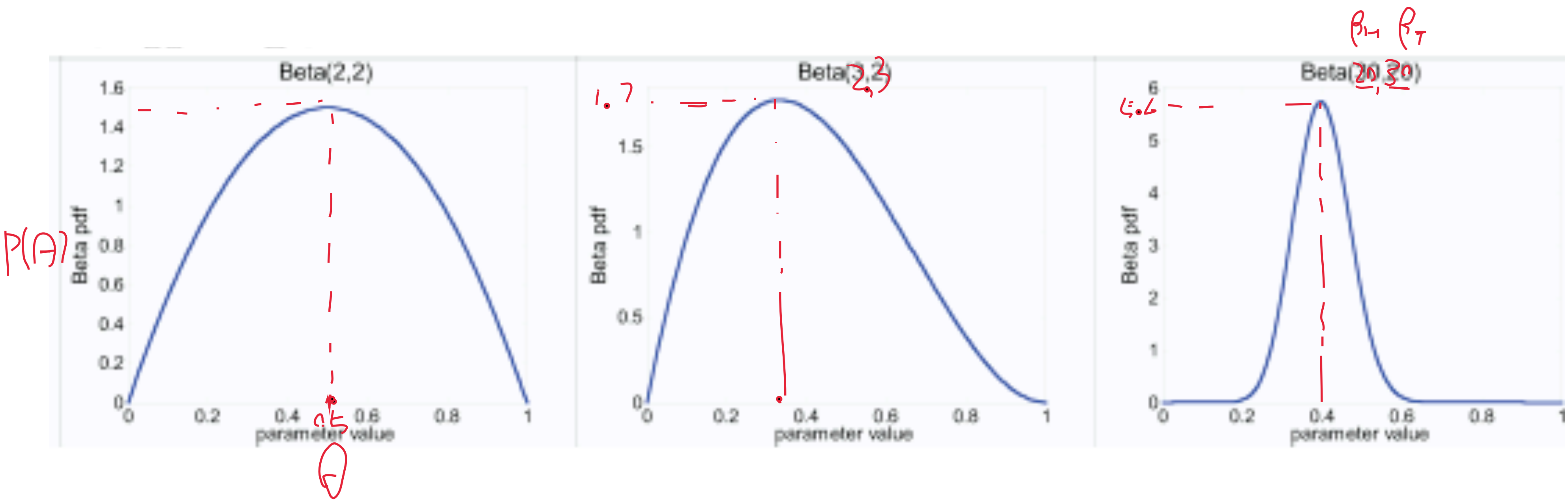
- $$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \;\; \propto \;\; P(\mathcal{D} \mid \theta) P(\theta)$

$$\boxed{\ln P(\theta|D)} \propto \ln\left(\theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}\right)$$

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

Beta prior distribution – P(θ)

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



[C. Guestrin]

Eg. 1 Coin flip problem

$D = \{x_1, x_2, \ldots, x_m\}$

$p(x) = \theta^{x} \cdot (1-\theta)^{1-x}$

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

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

and MAP estimate is therefore

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

$$P(x|\theta) = \theta_1^{\mathbb{1}_{x=1}} \theta_2^{\mathbb{1}_{x=2}} \cdots \theta_K^{\mathbb{1}_{x=K}}$$

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

Categorical distribution

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$) $\left(\sum_{k=1}^{K} \theta_k = 1\right)$

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

If prior is Dirichlet distribution,

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

Then posterior is Dirichlet distribution

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

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

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

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

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

e.g., X=gender, Y=playsFootball

or X=gender, Y=leftHanded

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