

Last Time:

- Learning parameters in the discrete & continuous cases
- Examples with binomial samples using Beta distribution
- Bucket examples → 101 coins (discrete)
∞ coins (continuous)
 - Pick 1 coin, assign probability for heads
 - If not all probabilities (relative frequencies) are equiprobable, model this belief with the Beta distribution.

- Beta examples
 - coin: $\text{Beta}(f; 50, 50)$
 - Thumbtack $\text{Beta}(f; 3, 3)$
 - Teeth Brushed $\text{Beta}(f; 13, 2)$

- Learning a relative frequency (updating belief)

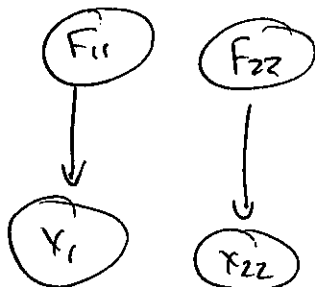
$p(f)$: prior $p(f|d)$ = posterior

- update example: thumbtack

- prior: $\text{beta}(f; 3, 3)$, observe 8h 2t
- posterior: $\text{beta}(f; 3+8, 3+2) = \text{beta}(f; 11, 5)$



- More than one param to learn - 2 buckets



update each posterior, re-compute probabilities for toss combinations HH, HT, TH, TT,

This time:

1 more example, but this time our 2nd toss depends on the outcome of the 1st.

2. Transition to multinomial variables
- Dirichlet

23/ Structure Learning Intro... Brief.

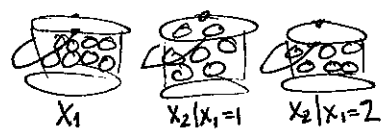
But first, from last time, ~~we~~ we had some trouble ~~going~~ ~~from~~ getting to

$$P(X^{(n+1)} = i | d) = E(F | d) = \frac{\alpha + s}{N + M} \text{ (Corollary 6.4)}$$

for beta distr'd F.

- Lemma 6.4 (proved using lemma 6.2) relates a beta distr.'s ~~exp~~ expectation in terms of 2 extra variables, s, b.
- Lemma 6.5 is just 1 more step ^{proved using 6.4 and} definition of Beta distr.
- Theorem 6.2 (is next, which gives us the relative frequency with which we'll obtain data d in the experiment, relating it to previous lemmas
- Corollary 6.2 relates thm 6.2 to betas
- Theorem 6.3 models posterior for data in form of prior
- Corollary 6.3 connects it to betas
- Theorem 6.4 shows our posterior estimate of the relative frequency (E(F|d)) is the same as the probability of the next experiment
- Corollary 6.4 relates it to betas.

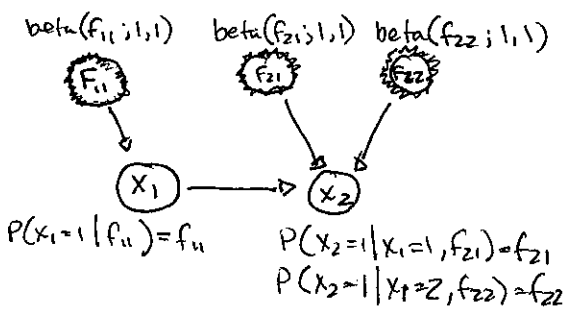
Three ~~BUCKETS~~ Example:



3 buckets in terms of variables. Each bucket has 10 coins with propensity for heads uniformly distributed between [0, 1].

1. Toss a coin from bucket-labelled X_1
2. If heads, (1) Toss a coin from $X_2|X_1=1$, else (2) toss from $X_2|X_1=2$

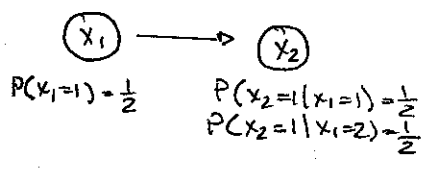
X_1 : r.v. whose value is the result of the 1st toss, X_2 for the 2nd toss.



F_{11} 's probability distribution is our belief concerning the relative frequency with which the 1st coin lands heads (1).
 F_{21} 's PD is for 2nd coin = heads given 1st = heads
 F_{22} 's PD is for 2nd coin = ~~heads~~ given 1st = tails.

In previous example, the same coin is used for the second toss regardless of 1st coin's outcome, so in that case F_{21} and F_{22} are completely dependent and are 1 node.

Above is ABN, here is regular BN:



Above has probability distribution for experiment. Here we have the marginal distribution of X_1, X_2 .

From this network, any heads-tails combination has probability $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$.

Experiment:

Data:

case	1	2	3	4	5	6	7
X_1	1	1	1	1	2	2	2
X_2	1	1	1	2	1	1	2

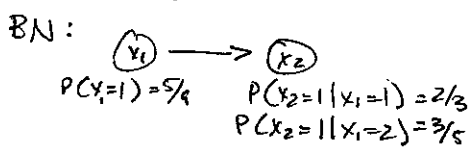
s_{11} : # times $X_1=1$	s_{22} : # times $X_2=1 X_1=2$
t_{11} : 2	t_{22} : $X_2=2 X_1=2$
s_{21} : $X_2=1 X_1=1$	
t_{21} : $X_2=2 X_1=1$	

Let's update the distros.

$P(f_{11}|d) = \text{beta}(f_{11}; a_{11}+s_{11}, b_{11}+t_{11}) = \text{beta}(f_{11}; 1+4, 1+3) = \text{beta}(f_{11}; 5, 4)$
 $P(f_{21}|d) = (f_{21}; 1+3, 1+1) = \text{beta}(f_{21}; 4, 2)$
 $P(f_{22}|d) = (f_{22}; 1+2, 1+1) = \text{beta}(f_{22}; 3, 2)$

Now the networks:

ABN: change betas.



Now the probabilities:

$P(X_1=1, X_2=1) = P(X_2=1|X_1=1)P(X_1=1) = (\frac{2}{3})(\frac{5}{9}) = 10/27$

1	2	$(\frac{1}{3})(\frac{5}{9}) = 5/27$
2	1	$(\frac{3}{5})(\frac{4}{9}) = 4/15$
2	2	$(\frac{2}{5})(\frac{4}{9}) = 8/45$

We've been doing only the binary case - what about multinomial variables?

- 1) Quantize
- 2) use more notation

Dirichlet density function with parameters $a_1, \dots, a_r \geq 1$, $N = \sum_{k=1}^r a_k$

$$p(f_1, f_2, \dots, f_{r-1}) = \frac{\Gamma(N)}{\prod_{k=1}^r \Gamma(a_k)} f_1^{a_1-1} f_2^{a_2-1} \dots f_r^{a_r-1} \quad 0 \leq f_k \leq 1, \sum_{k=1}^r f_k = 1$$

Note: $\text{Dir}(f_1, f_2, \dots, f_{r-1}; a_1, a_2, \dots, a_r)$ (Since f_r is uniquely determined by the first $r-1$ variables (i.e. $f_r = 1 - \sum_{h=1}^{r-1} f_h$), p is only a function of $r-1$ variables)

Dirichlet generalizes Beta. We've seen the k^{th} value occur a_k times in N trials. Like Beta:

$$E(F_k) = \frac{a_k}{N} \quad P(X = k | F_k) = f_k \quad P(X = k) = E(F_k) = \frac{a_k}{N}$$

Theorems, Lemmata and Corollaries on handout are applicable to Dirichlet too.

What about continuous variables?

- Normally distributed: ~~assume unknown mean, known variance~~ unknown mean, known variance
- Gamma Distribution (aka chi-square w/ k degrees of freedom), known mean, unknown variance
- t distribution: unknown mean, unknown variance
- More & more...

Structure Learning Introduction.

- Assumptions. Given a repeatable experiment whose outcome determines n random vars, assume
1. Relative frequency distributions of the variables admits a faithful DAG representation.
 2. Our belief concerning probability of the outcome of M execut

Assume: Data is generated by a BN.
All variables visible in every iteration

Score the models