Ling 5801: Lecture Notes 12
From Probabilistic CFGs to Probability Models

Contents

12.1 Full joint models 1
12.2 Inference / prediction 2
12.3 Factored models 3
12.4 Graphical representation ('Bayes net') 5
12.5 Conditional probability model class 5
12.6 Estimating θ from fully-specified ('annotated') data 7
12.7 Inducing θ from not-fully-specified ('unannotated') data 8
12.8 Simplification using independence assumptions 8
12.9 Another example: speech components (Phone, Voice, Back, Formant frequencies (binned)) 8
12.10Generative vs discriminative models 10
12.11Efficient inference by 'message passing' 10
12.12Example 10
12.13Example program 12
12.14Limits of message passing 12

12.1 Full joint models

A ***probability model*** M is a tuple consisting of:

- an ordered set $\mathcal{X}_M$ of ***random variables*** $X_1, ..., X_{|\mathcal{X}_M|}$ with uncertain values,
- a ***domain mapping*** $\mathcal{D}_M$ from random variables X_i in $\mathcal{X}_M$ to their domains $\mathcal{D}_{X_i}$, and
- a ***full joint probability distribution*** θ_M over each combination of values of $\mathcal{X}_M$

$$M = \langle \mathcal{X}_M, \mathcal{D}_M, \theta_M \rangle \text{ where } \mathcal{X}_M = \langle X_1, ..., X_{|\mathcal{X}_M|} \rangle$$
$$\text{and } \mathcal{D}_M = \{ \langle X_1, \mathcal{D}_{X_1} \rangle, ..., \langle X_{|\mathcal{X}_M|}, \mathcal{D}_{X_{|\mathcal{X}_M|}} \rangle \}$$
$$\text{and } \theta_M \in \mathcal{D}_{X_1} \times ... \times \mathcal{D}_{X_{|\mathcal{X}_M|}} \rightarrow \mathbb{R}_0^1$$

For example, a model of pronunciation variation:

$$\begin{aligned} M_{Pron} = \langle\, & \langle Reg, Wrd, Acou \rangle, \\ & \{ \langle Reg, \mathcal{D}_{Reg} \rangle, \langle Wrd, \mathcal{D}_{Wrd} \rangle, \langle Acou, \mathcal{D}_{Acou} \rangle \}, \\ & \theta_{Pron} \,\rangle \end{aligned}$$

1. Domains of random variables (upper case) are sets of mutually-exclusive ***values*** (lower case):

   $\mathcal{D}_{X_i} = \{ x_i, x_i', x_i'', ... \}$

   For example:

$\mathcal{D}_{Reg} = \{\text{ohio}, \text{phil}, ...\}$, (speaker region)
$\mathcal{D}_{Wrd} = \{\text{/nɛk/}, \text{/næk/}, ...\}$, (speaker's intended word)
$\mathcal{D}_{Acou} = \{\text{[nɛk]}, \text{[næk]}, ...\}$ (listener's observed phone)

2. P_θ assigns a probability to each combination of random variable values ('atomic event'):
   - $\forall_{x_1,...,x_{|\mathcal{X}_M|}\in\mathcal{D}_{X_1}\times...\times\mathcal{D}_{X_{|\mathcal{X}_M|}}}\ \mathsf{P}_{\theta_M}(x_1,...,x_{|\mathcal{X}_M|}) \in \mathbb{R}_0^1$ (each prob is 0 to 1),
   - $\sum_{x_1,...,x_{|\mathcal{X}_M|}} \mathsf{P}_{\theta_M}(x_1,...,x_{|\mathcal{X}_M|}) = 1$ (total prob of all atomic events is 1)

   For example, Ohians never pronounce /ɛ/ as [æ], but Philadelphians often do:

   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{ohio}, Wrd=\text{/næk/}, Acou=\text{[næk]}) = .2$
   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{ohio}, Wrd=\text{/næk/}, Acou=\text{[nɛk]}) = 0$
   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{ohio}, Wrd=\text{/nɛk/}, Acou=\text{[næk]}) = 0$ (never)
   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{ohio}, Wrd=\text{/nɛk/}, Acou=\text{[nɛk]}) = .3$
   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{phil}, Wrd=\text{/næk/}, Acou=\text{[næk]}) = .2$
   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{phil}, Wrd=\text{/næk/}, Acou=\text{[nɛk]}) = 0$
   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{phil}, Wrd=\text{/nɛk/}, Acou=\text{[næk]}) = .2$ (often)
   $\mathsf{P}_{\theta_{Pron}}(Reg=\text{phil}, Wrd=\text{/nɛk/}, Acou=\text{[nɛk]}) = .1$

   Write as probability table, rows sum to one:

   $\mathsf{P}_{\theta_{Pron}}(Reg, Wrd, Acou) =$

ohio	ohio	ohio	ohio	phil	phil	phil	phil
/næk/	/næk/	/nɛk/	/nɛk/	/næk/	/næk/	/nɛk/	/nɛk/
[næk]	[nɛk]	[næk]	[nɛk]	[næk]	[nɛk]	[næk]	[nɛk]
.2	0	0	.3	.2	0	.2	.1

   Each probability in the model is called a ***parameter***.

   There's one for each combination of values: in this case $2 \cdot 2 \cdot 2 = 8$.

12.2 Inference / prediction

We can use a probability model to infer answers to questions.

First, we will adopt a notational shorthand $\mathsf{P}_{\theta_M}(\mathcal{X})$, which means:

$\mathsf{P}_{\theta_M}(X_1=x_1,...,X_{|\mathcal{X}|}=x_{|\mathcal{X}|})$ for all $X_1,...,X_{|\mathcal{X}|}\in\mathcal{X}$ and all $x_1,...,x_{|\mathcal{X}|}\in X_1,...,X_{|\mathcal{X}|}$

Now we can define...

1. ***Marginals***: for any query $\mathcal{X}'$, sum probability over values of all variables not in $\mathcal{X}'$:

$$\forall_{\mathcal{X}'\subseteq\mathcal{X}_M}\ \mathsf{P}_{\theta_M}(\mathcal{X}') = \sum_{\mathcal{X}_M-\mathcal{X}'} \mathsf{P}_{\theta_M}(\mathcal{X}_M)$$

   For example:

$$\mathsf{P}_{\theta_{Pron}}(Wrd\!=\!w, Acou\!=\!a) = \sum_{r \in Reg} \mathsf{P}_{\theta_{Pron}}(Reg\!=\!r, Wrd\!=\!w, Acou\!=\!a)$$

$\mathsf{P}_{\theta_{Pron}}(Wrd, Acou) =$

/næk/	/næk/	/nɛk/	/nɛk/
[næk]	[nɛk]	[næk]	[nɛk]
.4	0	.2	.4

Another example:

$$\mathsf{P}_{\theta_{Pron}}(Acou\!=\!a) = \sum_{r \in Reg, w \in Wrd} \mathsf{P}_{\theta_{Pron}}(Reg\!=\!r, Wrd\!=\!w, Acou\!=\!a)$$

$\mathsf{P}_{\theta_{Pron}}(Acou) =$

[næk]	[nɛk]
.6	.4

2. ***Conditionals***: divide modeled & conditioned variables over conditioned variables

$$\forall_{\mathcal{X}', \mathcal{X}'' \subseteq \mathcal{X}_M} \; \mathsf{P}_{\theta_M}(\mathcal{X}' \mid \mathcal{X}'') = \frac{\mathsf{P}_{\theta_M}(\mathcal{X}' \cup \mathcal{X}'')}{\mathsf{P}_{\theta_M}(\mathcal{X}'')}$$

For example:

$$\mathsf{P}_{\theta_{Pron}}(Wrd\!=\!w \mid Acou\!=\!a) = \frac{\mathsf{P}_{\theta_{Pron}}(Wrd\!=\!w, Acou\!=\!a)}{\mathsf{P}_{\theta_{Pron}}(Acou\!=\!a)}$$

$$= \frac{\sum_{r \in Reg} \mathsf{P}_{\theta_{Pron}}(Reg\!=\!r, Wrd\!=\!w, Acou\!=\!a)}{\sum_{r \in Reg, w \in Wrd} \mathsf{P}_{\theta_{Pron}}(Reg\!=\!r, Wrd\!=\!w, Acou\!=\!a)}$$

Table has one row for each combination of conditioned-on variable values:

$\mathsf{P}_{\theta_{Pron}}(Wrd \mid Acou) =$

Acou	/næk/	/nɛk/
[næk]	.667	.333
[nɛk]	0	1

So, by our model, we can calculate that [næk] is probably from /næk/ in the general case:

$$\mathsf{P}_{\theta_{Pron}}(Wrd\!=\!\text{/næk/} \mid Acou\!=\![\text{næk}]) = \frac{.2 + .2}{.2 + 0 + .2 + .2} = .667$$

but if the speaker is from Philadelphia, the odds are even:

$$\mathsf{P}_{\theta_{Pron}}(Wrd\!=\!\text{/næk/} \mid Acou\!=\![\text{næk}], Reg\!=\!\text{phil}) = \frac{.2}{.2 + .2} = .5$$

Practice

Write a Python program to calculate $\mathsf{P}(Wrd \mid Acou)$ in a dictionary `WgivA[w,a]` given a full joint distribution $\mathsf{P}(Reg, Wrd, Acou)$ in a dictionary `RWA[r,w,a]`.

12.3 Factored models

The full joint distribution θ_M can sometimes be very large.

To simplify it, we'll first ***factor*** it into an ordered set of distributions for each variable:

1. A ***factored probability model*** M is a tuple:

   $$\begin{aligned} M = \langle \mathcal{X}_M, \mathcal{D}_M, \mathcal{C}_M, \theta_M \rangle \text{ where } \mathcal{X}_M &= \langle X_1, ..., X_{|\mathcal{X}_M|} \rangle \\ \text{and } \mathcal{D}_M &= \{ \langle X_1, \mathcal{D}_{X_1} \rangle, ..., \langle X_{|\mathcal{X}_M|}, \mathcal{D}_{X_{|\mathcal{X}_M|}} \rangle \} \\ \text{and } \mathcal{C}_M &= \{ \langle X_1, \mathcal{C}_{X_1} \rangle, ..., \langle X_{|\mathcal{X}_M|}, \mathcal{C}_{X_{|\mathcal{X}_M|}} \rangle \} \\ \text{and } \theta_M &= \{ \langle X_1, \theta_{X_1} \rangle, ..., \langle X_{|\mathcal{X}_M|}, \theta_{X_{|\mathcal{X}_M|}} \rangle \} \end{aligned}$$

   and, for now, $\forall_{X_i \in \mathcal{X}_M} \; \mathcal{C}_{X_i} = \{ X_1, ..., X_{i-1} \}$

   Now our model of pronunciation variation becomes:

   $$\begin{aligned} M_{Pron} = \langle \, & \langle Reg, Wrd, Acou \rangle, \\ & \{ \langle Reg, \{\text{ohio, phil}\} \rangle, \langle Wrd, \{\text{/nɛk/, /næk/}\} \rangle, \langle Acou, \{\text{[nɛk], [næk]}\} \rangle \}, \\ & \{ \langle Reg, \emptyset \rangle, \langle Wrd, \{Reg\} \rangle, \langle Acou, \{Reg, Wrd\} \rangle \}, \\ & \{ \langle Reg, \theta_{Reg} \rangle, \langle Wrd, \theta_{Wrd} \rangle, \langle Acou, \theta_{Acou} \rangle \} \, \rangle \end{aligned}$$

2. We can now define full joint probabilities as a product of these factors.

   This is a ***chain rule decomposition***, using the definition of conditional probability:

   $$\begin{aligned} \mathsf{P}_{\theta_M}(X_1, ..., X_{|\mathcal{X}_M|}) &= \frac{\mathsf{P}_{\theta_M}(X_1)}{1} \cdot \frac{\mathsf{P}_{\theta_M}(X_1, X_2)}{\mathsf{P}_{\theta_M}(X_1)} \cdot \frac{\mathsf{P}_{\theta_M}(X_1, X_2, X_3)}{\mathsf{P}_{\theta_M}(X_1, X_2)} \cdot \ldots \cdot \frac{\mathsf{P}_{\theta_M}(X_1, ..., X_{|\mathcal{X}_M|})}{\mathsf{P}_{\theta_M}(X_1, ..., X_{|\mathcal{X}_M|-1})} \\ &= \mathsf{P}_{\theta_M}(X_1) \cdot \mathsf{P}_{\theta_M}(X_2 \mid X_1) \cdot \mathsf{P}_{\theta_M}(X_3 \mid X_1, X_2) \cdot \ldots \cdot \mathsf{P}_{\theta_M}(X_{|\mathcal{X}_M|} \mid X_1, ..., X_{|\mathcal{X}_M|-1}) \\ &= \textstyle\prod_{i=1}^{|\mathcal{X}_M|} \mathsf{P}_{\theta_M}(X_i \mid X_1, ..., X_{i-1}) \\ &= \textstyle\prod_{i=1}^{|\mathcal{X}_M|} \mathsf{P}_{\theta_M}(X_i \mid \mathcal{C}_{X_i}) \end{aligned}$$

   We now have a bunch of ***conditional probability distributions*** instead of a full joint.

   We'll name them after the variables they model:

   $$\mathsf{P}_{\theta_{X_i}}(X_i \mid \mathcal{C}_{X_i}) \overset{\text{def}}{=} \mathsf{P}_{\theta_M}(X_i \mid \mathcal{C}_{X_i})$$

   Now, in our example:

   $$\begin{aligned} \mathsf{P}_{\theta_{Pron}}(Reg,\, Wrd,\, Acou) &= \frac{\mathsf{P}_{\theta_{Pron}}(Reg)}{1} \cdot \frac{\mathsf{P}_{\theta_{Pron}}(Reg, Wrd)}{\mathsf{P}_{\theta_{Pron}}(Reg)} \cdot \frac{\mathsf{P}_{\theta_{Pron}}(Reg, Wrd, Acou)}{\mathsf{P}_{\theta_{Pron}}(Reg, Wrd)} \\ &= \mathsf{P}_{\theta_{Reg}}(Reg) \cdot \mathsf{P}_{\theta_{Wrd}}(Wrd \mid Reg) \cdot \mathsf{P}_{\theta_{Acou}}(Acou \mid Reg,\, Wrd) \end{aligned}$$

3. Now the number of parameters in any variable's conditional probability distribution:

   $$\mathsf{P}_{\theta_{X_i}}(X_i \mid X_1, ..., X_{i-1})$$

   is the product of the cardinalities of its modeled and conditioned-on variables:

   $$|\theta_{X_i}| = |X_i| \cdot |X_1| \cdot \ldots \cdot |X_{i-1}|$$

   In our example:

   $$|\theta_{Acou}| = |Acou| \cdot |Reg| \cdot |Wrd| = 2 \cdot 2 \cdot 2 = 8 \text{ parameters}$$

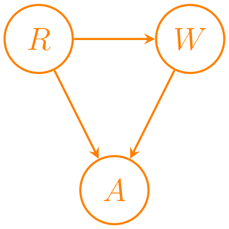
   So far this is no better than the full joint distribution... but that's ok for now.

12.4 Graphical representation (‘Bayes net’)

Terms $\mathsf{P}_{\theta_{X_i}}(X_i \,|\, \mathcal{C}_{X_i})$ from chain can be represented graphically —

each variable X_i is a circle, with an edge from each variable in the condition $\mathcal{C}_{X_i}$:

For example:



Graphical models are like sudoku: some values given, some unknown, values interdepend

In fact, they can be considered a *generalization* of sudoku puzzles!

12.5 Conditional probability model class

Extend ‘`model.py`’ to implement conditional model: dictionary of dictionary

(You may find this useful for certain problem set questions!)

```
import re
# define distribution to map value tuples to probs (or frequencies or scores)
class Model(dict):
    # init with model id
    def __init__(self,i=’’):
        self.id = i
    # read model
    def read(self,s):
        m = re.search(’^ *’+self.id+’ +: +(.*?) += +(.*) *’,s)
        if m is not None:
            v = tuple(re.split(’ +’,m.group(1)))
            if len(v)==1: v = v[0]
            self[v] = float(m.group(2))
    # write model
    def write(self):
        for v in sorted(self):
            if self[v]>0.0:
                print self.id,
                print ’:’,
                if type(v) is tuple:
                    for f in v:
                        print f,
                else: print v,
                print ’=’,self[v]
```

```
# define model to map condition tuples to distributions
class CondModel(dict):
    # populate with default values when queried on missing keys
    def __missing__(self,k):
        self[k]=Model()
        return self[k]
    # define get without promiscuity
    def get(self,k):
        return dict.get(self,k,Model())
    # init with model id
    def __init__(self,i):
        self.id = i
    # read model
    def read(self,s):
        m = re.search(’^ *’+self.id+’ +(.*?) +: +(.*?) += +(.*) *’,s)
        if m is not None:
            c = tuple(re.split(’ +’,m.group(1)))
            if len(c)==1: c = c[0]
            v = tuple(re.split(’ +’,m.group(2)))
            if len(v)==1: v = v[0]
            self[c][v] = float(m.group(3))
    # write model
    def write(self):
        for c in sorted(self):
            for v in sorted(self[c]):
                if self[c][v]>0.0:
                    print self.id,
                    if type(c) is tuple:
                        for f in c:
                            print f,
                    else: print c,
                    print ’:’,
                    if type(v) is tuple:
                        for f in v:
                            print f,
                    else: print v,
                    print ’=’,self[c][v]
```

Example, e.g. type into a program file ‘`myprog.py`’:

```
import re
import sys
import model

# read in params beginning with ’R’ (prior model: no value before colon)
# read in params beginning with ’W’ (prior model: no value before colon)
# read in params beginning with ’A’ (cond model: value before colon)
```

```
R = model.Model(’R’)
W = model.Model(’W’)
A = model.CondModel(’A’)
for line in sys.stdin:
    R.read(line)
    W.read(line)
    A.read(line)

# use a prior model
for r in R:
  print ( ’prob of ’+r+’ is ’+str(R[r]) )
# use a conditional model
for r,w in A:
    for a in A[r,w]:
        print ( ’prob of ’+a+’ given ’+r+’ and ’+w+’ is ’+str(A[r,w][a]) )
# calc prob of W=’/naek/’ given R=’phil’ and A=’[naek]’
probAnyW = 0.0
for w in W:
  probAnyW = probAnyW + R[’phil’] * W[w] * A[’phil’,w][’[naek]’]
print ( R[’phil’] * W[’/naek/’] * A[’phil’,’/naek/’][’[naek]’] / probAnyW )
```

Reads files in the following format:

```
R : ohio = .5
R : phil = .5
W : /naek/ = .4
W : /nek/  = .6
A ohio /naek/ : [naek] = 1.0
A ohio /nek/  : [nek]  = 1.0
A phil /naek/ : [naek] = 1.0
A phil /nek/  : [nek]  = .333333
A phil /nek/  : [naek] = .666667
⋮
```

then describes the information in the file and then prints the probability of /næk/:

```
0.499999875
```

12.6 Estimating θ from fully-specified (‘annotated’) data

Every θ_{X_i} is a conditional probability table $\mathsf{P}_{\theta_{X_i}}(X_i \,|\, \mathcal{C}_{X_i})$.

Simply count instances of $x_1, ..., x_i \in \mathcal{C}_{X_i} \times X_i$ and divide by count of $x_1, ..., x_{i-1} \in \mathcal{C}_{X_i}$.

This is called ***relative frequency estimation***.

12.7 Inducing θ from not-fully-specified (‘unannotated’) data

You’ll study this in comp ling 2!

12.8 Simplification using independence assumptions

We can now make ***independence assumptions*** about which dependencies don’t matter:

Define conditioned-on variables as a subset of preceding vars:

$\forall_{X_i \in \mathcal{X}_M}\, \mathcal{C}_{X_i} \subseteq \{X_1, ..., X_{i-1}\}$

For example, our model of pronunciation variation becomes:

$$\begin{aligned} M_{Pron} = \langle\, & \langle \mathit{Reg}, \mathit{Wrd}, \mathit{Acou} \rangle, \\ & \{ \langle \mathit{Reg}, \{\text{ohio}, \text{phil}\} \rangle, \langle \mathit{Wrd}, \{\text{/nɛk/}, \text{/næk/}\} \rangle, \langle \mathit{Acou}, \{\text{[nɛk]}, \text{[næk]}\} \rangle \}, \\ & \{ \langle \mathit{Reg}, \emptyset \rangle, \langle \mathit{Wrd}, \emptyset \rangle, \langle \mathit{Acou}, \{\mathit{Reg}, \mathit{Wrd}\} \rangle \}, \\ & \{ \langle \mathit{Reg}, \theta_{\mathit{Reg}} \rangle, \langle \mathit{Wrd}, \theta_{\mathit{Wrd}} \rangle, \langle \mathit{Acou}, \theta_{\mathit{Acou}} \rangle \} \,\rangle \end{aligned}$$

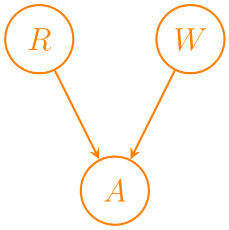
The joint probability is:

$$\mathsf{P}_{\theta_{Pron}}(Reg, Wrd, Acou) = \mathsf{P}_{\theta_{Reg}}(Reg) \cdot \mathsf{P}_{\theta_{Wrd}}(Wrd \mid Reg) \cdot \mathsf{P}_{\theta_{Acou}}(Acou \mid Reg, Wrd)$$

$$\mathsf{P}_{\theta_{Pron}}(Reg, Wrd, Acou) \overset{\text{def}}{=} \mathsf{P}_{\theta_{Reg}}(Reg) \cdot \mathsf{P}_{\theta_{Wrd}}(Wrd) \cdot \mathsf{P}_{\theta_{Acou}}(Acou \mid Reg, Wrd)$$

because of independence assumption: $\mathsf{P}_{\theta_{Pron}}(Wrd \mid Reg) \overset{\text{def}}{=} \mathsf{P}_{\theta_{Pron}}(Wrd)$

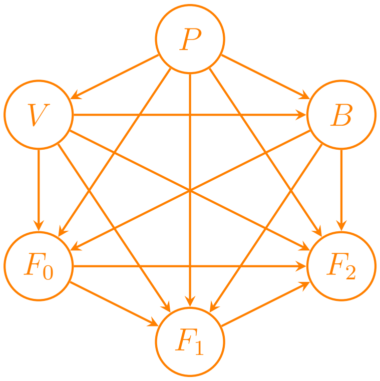
Graphically:



12.9 Another example: speech components (Phone, Voice, Back, Formant frequencies (binned))

1. Chain rule decomposition (no independence assumptions):

$$\begin{aligned} M_{Sp} = \langle\, & \langle P, V, B, F_0, F_1, F_2 \rangle, \\ & \{ \langle P, \{\text{/i/}, \text{/u/}\} \rangle, \langle V, \{+, -\} \rangle, \langle B, \{+, -\} \rangle, \langle F_0, \mathbb{I}_0^{99} \rangle, \langle F_1, \mathbb{I}_0^{99} \rangle, \langle F_2, \mathbb{I}_0^{99} \rangle \}, \\ & \{ \langle P, \emptyset \rangle, \langle V, \{P\} \rangle, \langle B, \{P,V\} \rangle, \langle F_0, \{P,V,B\} \rangle, \langle F_1, \{P,V,B,F_0\} \rangle, \langle F_2, \{P,V,B,F_0,F_1\} \rangle \}, \\ & \{ \langle P, \theta_P \rangle, \langle V, \theta_V \rangle, \langle B, \theta_B \rangle, \langle F_0, \theta_{F_0} \rangle, \langle F_1, \theta_{F_1} \rangle, \langle F_2, \theta_{F_2} \rangle \} \,\rangle \end{aligned}$$

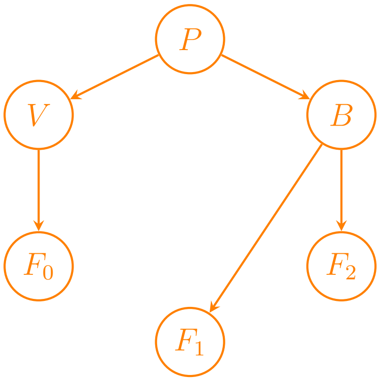


$$\mathsf{P}_{\theta_{Sp}}(P,V,B,F_0,F_1,F_2) = \mathsf{P}_{\theta_P}(P)\cdot\mathsf{P}_{\theta_V}(V \mid P)\cdot\mathsf{P}_{\theta_B}(B \mid P,V)\cdot \\ \mathsf{P}_{\theta_{F_0}}(F_0 \mid P,V,B)\cdot\mathsf{P}_{\theta_{F_1}}(F_1 \mid P,V,B,F_0)\cdot\mathsf{P}_{\theta_{F_2}}(F_2 \mid P,V,B,F_0,F_1)$$

Here $|\theta_{F_2}| = 100\cdot 2\cdot 2\cdot 2\cdot 100\cdot 100 = 8,000,000$ parameters!

2. With independence assumptions:

$$\begin{aligned} M_{Sp} = \langle\, & \langle P,V,B,F_0,F_1,F_2\rangle, \\ & \{\langle P,\{\text{/i/},\text{/u/}\}\rangle,\langle V,\{+,-\}\rangle,\langle B,\{+,-\}\rangle,\langle F_0,\mathbb{I}_0^{99}\rangle,\langle F_1,\mathbb{I}_0^{99}\rangle,\langle F_2,\mathbb{I}_0^{99}\rangle\}, \\ & \{\langle P,\emptyset\rangle,\langle V,\{P\}\rangle,\langle B,\{P\}\rangle,\langle F_0,\{V\}\rangle,\langle F_1,\{B\}\rangle,\langle F_2,\{B\}\rangle\}, \\ & \{\langle P,\theta_P\rangle,\langle V,\theta_V\rangle,\langle B,\theta_B\rangle,\langle F_0,\theta_{F_0}\rangle,\langle F_1,\theta_{F_1}\rangle,\langle F_2,\theta_{F_2}\rangle\}\,\rangle \end{aligned}$$



$$\mathsf{P}_{\theta_{Sp}}(P,V,B,F_0,F_1,F_2) \overset{\text{def}}{=} \mathsf{P}_{\theta_P}(P)\cdot\mathsf{P}_{\theta_V}(V \mid P)\cdot\mathsf{P}_{\theta_B}(B \mid P)\cdot \\ \mathsf{P}_{\theta_{F_0}}(F_0 \mid V)\cdot\mathsf{P}_{\theta_{F_1}}(F_1 \mid B)\cdot\mathsf{P}_{\theta_{F_2}}(F_2 \mid B)$$

Now $|\theta_{F_2}| = 100\cdot 2 = 200$ parameters!

Practice

1. Draw a model of pet communication given random variables for vocalization, hunger, location, and tail being stepped on. The pet vocalizes when it's hungry, or when its tail is stepped on. Its tail gets stepped on only when it is in the kitchen.

2. Write the factored equation for $\mathsf{P}(\mathit{Hunger},\mathit{TailStep} \mid \mathit{Voc},\mathit{Loc})$

3. Write a program to do this calculation in `modHTgivVL[v,l][h,t]` given models `modL[l]`, `modT[l][t]`, `modH[h]`, `modV[t,h][v]`

12.10 Generative vs discriminative models

Models that condition observed variables on un-observed variables are called ***generative***.

Models that don't, and ignore the observed variable model, are called ***discriminative***.

Discrim models have annoying properties ('label-bias'), need overlapping joint variables.

12.11 Efficient inference by 'message passing'

Most queries don't need to calculate the full joint distribution (through 8,000,000 iterations):

$$\begin{aligned}\mathsf{P}_{\theta_{Sp}}(b) &= \sum_{p,v,f_0,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p,v,b,f_0,f_1,f_2) \\ &\overset{\text{def}}{=} \sum_{p,v,f_0,f_1,f_2} \mathsf{P}_{\theta_P}(p)\cdot\mathsf{P}_{\theta_V}(v\,|\,p)\cdot\mathsf{P}_{\theta_B}(b\,|\,p)\cdot\mathsf{P}_{\theta_{F_0}}(f_0\,|\,v)\cdot\mathsf{P}_{\theta_{F_1}}(f_1\,|\,b)\cdot\mathsf{P}_{\theta_{F_2}}(f_2\,|\,b)\end{aligned}$$

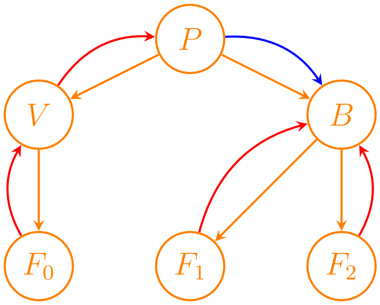
Instead, marginalize as we go, storing marginals (conditional probability tables) as 'messages':

$$\begin{aligned}\mathsf{P}_{\theta_{Sp}}(b) &= \sum_{p,v,f_0,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p,v,b,f_0,f_1,f_2) \\ &\overset{\text{def}}{=} \sum_{p} \Bigg(_{p}\, \mathsf{P}(p)\cdot\Bigg(_{p}\, \sum_{v} \mathsf{P}(v\,|\,p)\cdot\Bigg(_{v}\, \sum_{f_0} \mathsf{P}(f_0\,|\,v)\Bigg)\Bigg)\Bigg)\cdot\mathsf{P}(b\,|\,p)\cdot\Bigg(_{b}\, \sum_{f_1} \mathsf{P}(f_1\,|\,b)\Bigg)\cdot\Bigg(_{b}\, \sum_{f_2} \mathsf{P}(f_2\,|\,b)\Bigg)\end{aligned}$$

(Re-arrangement of terms just comes from distributing products over sums in the full joint.)

Blue parens show ***forward messages***: distributions over free modeled variables (subscripts).

Red parens show ***backward messages***: likelihood fns over free conditioned-on variables (subscr).



Now just need space of a conditional probability distribution per variable!

12.12 Example

For example, to solve the following query (where variable F_0 is actually observed):

$$\mathsf{P}_{\theta_{Sp}}(b,\, f_0{=}12) = \sum_{p,v,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p, v, b, f_0{=}12, f_1, f_2)$$
$$\overset{\text{def}}{=} \sum_{p} \Bigg(_{p} \mathsf{P}(p)\cdot \Bigg(_{p} \sum_{v} \mathsf{P}(v \,|\, p)\cdot \mathsf{P}(f_0{=}12 \,|\, v)\Bigg)\Bigg)\cdot \mathsf{P}(b \,|\, p)\cdot \Bigg(_{b} \sum_{f_1} \mathsf{P}(f_1 \,|\, b)\Bigg)\cdot \Bigg(_{b} \sum_{f_2} \mathsf{P}(f_2 \,|\, b)\Bigg)$$

given the following models:

$\mathsf{P}_{\theta_P}(P) =$

/i/	/u/
.4	.6

$\mathsf{P}_{\theta_V}(V \,|\, P) =$

P	$+$	$-$
/i/	.8	.2
/u/	1	0

$\mathsf{P}_{\theta_{F_0}}(F_0 \,|\, V) =$

V	…	11	12	13	…
$+$	…	.04	.02	.01	…
$-$	…	.01	.01	.01	…

$\mathsf{P}_{\theta_B}(B \,|\, P) =$

P	$+$	$-$
/i/	0	1
/u/	.5	.5

we would generate the following messages:

from F_0 to V: $\mathsf{P}(F_0{=}12 \,|\, V) =$

V	12
$+$	.02
$-$	.01

from V to P: $\mathsf{P}(F_0{=}12 \,|\, P) =$

P	$F_0 = 12$
/i/	$\mathsf{P}_{\theta_{F_0}}(12 \,\|\, +)\cdot \mathsf{P}_{\theta_V}(+ \,\|\, \text{/i/}) + \mathsf{P}_{\theta_{F_0}}(12 \,\|\, -)\cdot \mathsf{P}_{\theta_V}(- \,\|\, \text{/i/})$ $= .02\cdot .8 + .01\cdot .2 = .018$
/u/	$\mathsf{P}_{\theta_{F_0}}(12 \,\|\, +)\cdot \mathsf{P}_{\theta_V}(+ \,\|\, \text{/u/}) + \mathsf{P}_{\theta_{F_0}}(12 \,\|\, -)\cdot \mathsf{P}_{\theta_V}(- \,\|\, \text{/u/})$ $= .02\cdot 1 + .01\cdot 0 = .020$

from P to B: $\mathsf{P}(P, F_0{=}12) =$

P=/i/, F_0=12	P=/u/, F_0=12
$\mathsf{P}_{\theta_P}(\text{/i/})\cdot \mathsf{P}(F_0{=}12 \,\|\, P{=}\text{/i/})$ $= .4\cdot .018 = .0072$	$\mathsf{P}_{\theta_P}(\text{/u/})\cdot \mathsf{P}(F_0{=}12 \,\|\, P{=}\text{/u/})$ $= .6\cdot .020 = .0120$

from F_1 to B: $\mathsf{P}(\text{any } F_1 \,|\, B) =$

B	any
$+$	1
$-$	1

from F_2 to B: $\mathsf{P}(\text{any } F_2 \,|\, B) =$

B	any
$+$	1
$-$	1

Product of model and three messages at B:

$\mathsf{P}(B, F_0{=}12) =$

$B{=}+, F_0{=}12$	$B{=}-, F_0{=}12$
$\mathsf{P}(P{=}\text{/i/}, F_0{=}12) \cdot \mathsf{P}_B(+ \mid \text{/i/}) \cdot 1 \cdot 1$	$\mathsf{P}(P{=}\text{/i/}, F_0{=}12) \cdot \mathsf{P}_B(- \mid \text{/i/}) \cdot 1 \cdot 1$
$+\, \mathsf{P}(P{=}\text{/u/}, F_0{=}12) \cdot \mathsf{P}_B(+ \mid \text{/u/}) \cdot 1 \cdot 1$	$+\, \mathsf{P}(P{=}\text{/u/}, F_0{=}12) \cdot \mathsf{P}_B(- \mid \text{/u/}) \cdot 1 \cdot 1$
$= .0072 \cdot 0 \cdot 1 \cdot 1 + .0120 \cdot .5 \cdot 1 \cdot 1 = .0060$	$= .0072 \cdot 1 \cdot 1 \cdot 1 + .0120 \cdot .5 \cdot 1 \cdot 1 = .0132$

Normalized:

$\mathsf{P}(B \mid F_0{=}12) =$

$B{=}+$	$B{=}-$
$\frac{.0060}{.0060+.0132} = .3125$	$\frac{.0132}{.0060+.0132} = .6875$

12.13 Example program

Find $\mathsf{P}(B \mid F_0 = 12)$ from Model `modP`, CondModels `modV`, `modB`, `modF0`, `modF1`, `modF2`:

```
bkwdF0 = {}
for v in modF0:   # obtain likelihood of observation given V (backward message)
  bkwdF0[v] = modF0[v][’12’]
bkwdV = {}
for p in modV:    # marginalize or ’sum out’ V to get likelihood given P (bkwd msg)
  for v in modV[p]:
    bkwdV[p] = bkwdV.get(p,0.0) + (modV[p][v] * bkwdF0[v])
fwrdP = {}
for p in P:       # multiply prior over P by likelihood given P (backward message)
  fwrdP[p] = modP[p] * bkwdV[p]
...
```

Practice

Complete the above example.

12.14 Limits of message passing

Message passing degrades when network is not singly-connected.

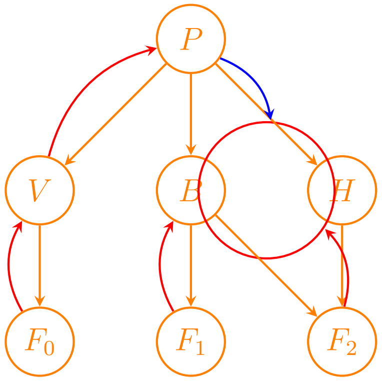
For example, adding variable for height w. dependencies from P, to F_2, creates a ‘diamond’:

$$\begin{aligned} M_{Sp} = \langle\, & \langle P, V, B, H, F_0, F_1, F_2 \rangle, \\ & \{ \langle P, \{\text{/i/}, \text{/u/}\} \rangle, \langle V, \{+,-\} \rangle, \langle B, \{+,-\} \rangle, \langle H, \{+,-\} \rangle, \langle F_0, \mathbb{I}_0^{99} \rangle, \langle F_1, \mathbb{I}_0^{99} \rangle, \langle F_2, \mathbb{I}_0^{99} \rangle \}, \\ & \{ \langle P, \emptyset \rangle, \langle V, \{P\} \rangle, \langle B, \{P\} \rangle, \langle H, \{P\} \rangle, \langle F_0, \{V\} \rangle, \langle F_1, \{B\} \rangle, \langle F_2, \{B, H\} \rangle \}, \\ & \{ \langle P, \theta_P \rangle, \langle V, \theta_V \rangle, \langle B, \theta_B \rangle, \langle H, \theta_H \rangle, \langle F_0, \theta_{F_0} \rangle, \langle F_1, \theta_{F_1} \rangle, \langle F_2, \theta_{F_2} \rangle \} \,\rangle \end{aligned}$$

This means some marginals will have multiple free variables (which makes them larger):

$$\begin{aligned}\mathsf{P}_{\theta_{Sp}}(b) &= \sum_{p,v,h,f_0,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p,v,b,h,f_0,f_1,f_2)\\ &\overset{\text{def}}{=} \sum_{p} \left(_{p} \mathsf{P}(p)\cdot \left(_{p} \sum_{v} \mathsf{P}(v \mid p)\cdot\ldots\right)\right)\cdot \mathsf{P}(b \mid p)\cdot \left(_{b} \sum_{f_1} \mathsf{P}(f_1 \mid b)\right)\cdot \sum_{h} \mathsf{P}(h \mid p)\cdot \left(_{b,h} \sum_{f_2} \mathsf{P}(f_2 \mid b,h)\right)\end{aligned}$$

Graphically, messages must pass through ‘junctions’ of joint variables:



Well, they’re not full joints at least.