A Brief Introduction to Directed Graphical Models
Probabilistic Models in the Study of Language
Day 3

Roger Levy

UC San Diego
Department of Linguistics

August 8, 2012
Bayesian parameter estimation

The scenario: you are a native English speaker in whose experience passivizable constructions are passivized with frequency $q$.

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

You encounter a new dialect of English and hear data $y$ consisting of $n$ passivizable utterances, $m$ of which were passivized:

$$X \sim \text{Bern}(\pi)$$

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- **Or** place a probability distribution on the number of passives in the next $N$ passivizable utterances.
Anatomy of Bayesian inference

Simplest possible scenario:

$I \rightarrow \theta \rightarrow Y$
Anatomy of Bayesian inference

Simplest possible scenario:

The corresponding Bayesian inference:

\[ P(\theta | y, I) = \frac{P(y | \theta, I)P(\theta | I)}{P(y | I)} \]
Anatomy of Bayesian inference

Simplest possible scenario:

\[ \begin{array}{c}
I \rightarrow \theta \rightarrow Y
\end{array} \]

The corresponding Bayesian inference:

\[
P(\theta|y, I) = \frac{P(y|\theta, I)P(\theta|I)}{P(y|I)}
\]

\[
\text{Likelihood for } \theta \quad \text{Prior over } \theta
\]

\[
= \frac{P(y|\theta)}{P(y|I)}
\]

\[
\text{Likelihood marginalized over } \theta
\]

(because \( y \perp I | \theta \))
Anatomy of Bayesian inference

Simplest possible scenario:

The corresponding Bayesian inference:

\[
P(\theta | y, I) = \frac{P(y | \theta, I)P(\theta | I)}{P(y | I)}
\]

Likelihood for \( \theta \)  
Prior over \( \theta \)

\[
= \frac{\underbrace{P(y | \theta)}}{P(y | I)} \frac{\underbrace{P(\theta | I)}}{P(y | I)} \quad \text{(because } y \perp I | \theta)\]

Likelihood marginalized over \( \theta \)

- At the “bottom” of the graph, our model is the binomial distribution:

\[
P(y | \theta) \sim Binom(n, \theta)
\]

- But to get things going we have to set the prior \( P(\theta | I) \).
Priors for the binomial distribution

- For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$
  - Because the prior is often supposed to account for “knowledge we bring to the table”, we often write $P(\theta|I)$ to be explicit
Priors for the binomial distribution

For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$

- Because the prior is often supposed to account for “knowledge we bring to the table”, we often write $P(\theta|I)$ to be explicit

- Model parameters are nearly always real-valued, so $P(\theta)$ is generally a multivariate continuous distribution
For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$

Because the prior is often supposed to account for “knowledge we bring to the table”, we often write $P(\theta|I)$ to be explicit.

Model parameters are nearly always real-valued, so $P(\theta)$ is generally a multivariate continuous distribution.

In general, the sky is the limit as to what you choose for $P(\theta)$.
Priors for the binomial distribution

- For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$
  - Because the prior is often supposed to account for “knowledge we bring to the table”, we often write $P(\theta|I)$ to be explicit

- Model parameters are nearly always real-valued, so $P(\theta)$ is generally a multivariate continuous distribution

- In general, the sky is the limit as to what you choose for $P(\theta)$

- But in many cases there are useful priors that will make your life easier
The beta distribution

The beta distribution has two parameters $\alpha_1, \alpha_2 > 0$ and is defined as:

$$P(\pi | \alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1}$$

$$(0 \leq \pi \leq 1, \alpha_1 > 0, \alpha_2 > 0)$$

where the beta function $B(\alpha_1, \alpha_2)$ serves as a normalizing constant:

$$B(\alpha_1, \alpha_2) = \int_0^1 \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1} d\pi$$
Some beta distributions

If $X \sim B(\alpha_1, \alpha_2)$:

- $E[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}$
- If $\alpha_1, \alpha_2 > 1$, then $X$ has a mode at $\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence $I = \langle \alpha_1, \alpha_2 \rangle$. 
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence

\[ I = \langle \alpha_1, \alpha_2 \rangle. \]

\[
P(\pi | y, \alpha_1, \alpha_2) = \frac{P(y | \pi)P(\pi | \alpha_1, \alpha_2)}{P(y | \alpha_1, \alpha_2)}
\]  

(1)
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence $I = \langle \alpha_1, \alpha_2 \rangle$.

$$P(\pi|y, \alpha_1, \alpha_2) = \frac{P(y|\pi)P(\pi|\alpha_1, \alpha_2)}{P(y|\alpha_1, \alpha_2)} \quad (1)$$

Since the denominator is not a function of $\pi$, it is a normalizing constant. Ignore it and work in terms of proportionality:

$$P(\pi|y, \alpha_1, \alpha_2) \propto P(y|\pi)P(\pi|\alpha_1, \alpha_2)$$
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence $I = \langle \alpha_1, \alpha_2 \rangle$.

$$P(\pi|y, \alpha_1, \alpha_2) = \frac{P(y|\pi)P(\pi|\alpha_1, \alpha_2)}{P(y|\alpha_1, \alpha_2)}$$  \hspace{1cm} (1)

Since the denominator is not a function of $\pi$, it is a normalizing constant. Ignore it and work in terms of proportionality:

$$P(\pi|y, \alpha_1, \alpha_2) \propto P(y|\pi)P(\pi|\alpha_1, \alpha_2)$$

Likelihood for the binomial distribution is

$$P(y|\pi) = \binom{n}{m} \pi^m (1 - \pi)^{n-m}$$
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence $I = \langle \alpha_1, \alpha_2 \rangle$.

$$P(\pi | y, \alpha_1, \alpha_2) = \frac{P(y | \pi) P(\pi | \alpha_1, \alpha_2)}{P(y | \alpha_1, \alpha_2)}$$  \hspace{1cm} (1)

Since the denominator is not a function of $\pi$, it is a normalizing constant. Ignore it and work in terms of proportionality:

$$P(\pi | y, \alpha_1, \alpha_2) \propto P(y | \pi) P(\pi | \alpha_1, \alpha_2)$$

Likelihood for the binomial distribution is

$$P(y | \pi) = \binom{n}{m} \pi^m (1 - \pi)^{n-m}$$

Beta prior is

$$P(\pi | \alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}$$
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi | y, \alpha_1, \alpha_2) \propto \pi^m (1 - \pi)^{n-m} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}
\]
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
\mathcal{P}(\pi|y, \alpha_1, \alpha_2) \propto \pi^m (1 - \pi)^{n-m} \pi^{\alpha_1 - 1} (1 - \pi)^{\alpha_2 - 1}
\]

\[
\propto \pi^{m + \alpha_1 - 1} (1 - \pi)^{n - m + \alpha_2 - 1}
\]
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi|y, \alpha_1, \alpha_2) \propto \pi^m(1-\pi)^{n-m} \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1} \\
\propto \pi^{m+\alpha_1-1}(1-\pi)^{n-m+\alpha_2-1}
\]

Crucial trick: this is itself a beta distribution! Recall that if \( \theta \sim \text{Beta}(\alpha_1, \alpha_2) \) then

\[
P(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}
\]
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi|y, \alpha_1, \alpha_2) \propto \pi^m (1-\pi)^{n-m} \pi^{\alpha_1-1} (1-\pi)^{\alpha_2-1}
\]

\[
\propto \pi^{m+\alpha_1-1} (1-\pi)^{n-m+\alpha_2-1}
\]

**Crucial trick:** this is itself a beta distribution! Recall that if
\( \theta \sim \text{Beta}(\alpha_1, \alpha_2) \) then

\[
P(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1-\pi)^{\alpha_2-1}
\]

Hence \( P(\theta|y, \alpha_1, \alpha_2) \) is distributed as \( \text{Beta}(\alpha_1 + m, \alpha_2 + n - m) \).
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi | y, \alpha_1, \alpha_2) \propto \pi^m (1 - \pi)^{n-m} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1} \\
\propto \pi^{m+\alpha_1-1} (1 - \pi)^{n-m+\alpha_2-1}
\]

**Crucial trick:** this is itself a beta distribution! Recall that if \( \theta \sim Beta(\alpha_1, \alpha_2) \) then

\[
P(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}
\]

Hence \( P(\theta | y, \alpha_1, \alpha_2) \) is distributed as \( Beta(\alpha_1 + m, \alpha_2 + n - m) \).

- With a beta prior and a binomial likelihood, the posterior is still beta-distributed. This is called conjugacy.
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivable utterances.
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

- To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability
Using our beta-binomial model

Goal:

▸ Estimate the success parameter $\pi$ associated with passivization in the new English dialect;

▸ Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

▸ To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability

▸ $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (?)
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

- To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability

  - $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (?)

- The mode of a beta distribution is $\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability

- $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (?)
- The mode of a beta distribution is $\frac{\alpha_1-1}{\alpha_1+\alpha_2-2}$
- Hence we might use $\alpha_1 = 3, \alpha_2 = 24$ (note that $\frac{2}{25} = 0.08$)
Using our beta-binomial model

Goal:

▷ Estimate the success parameter $\pi$ associated with passivization in the new English dialect;

▷ **Or** place a probability distribution on the number of passives in the next $N$ passivizable utterances.

▷ To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability

▷ $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (?)

▷ The mode of a beta distribution is $\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$

▷ Hence we might use $\alpha_1 = 3$, $\alpha_2 = 24$ (note that $\frac{2}{25} = 0.08$)

▷ Suppose that $n = 7$, $m = 2$: our posterior will be $\text{Beta}(5, 29)$, hence $\hat{\pi} = \frac{4}{32} = 0.125$
Beta-binomial posterior distributions
Fully Bayesian density estimation

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.
Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

In the fully Bayesian view, we don’t summarize our posterior beliefs into a point estimate; rather, we marginalize over them in predicting the future:
Fully Bayesian density estimation

Goal:
- Estimate the success parameter \( \pi \) associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next \( N \) passivizable utterances.

In the fully Bayesian view, we don’t summarize our posterior beliefs into a point estimate; rather, we marginalize over them in predicting the future:

\[
P(y_{new} | y, I) = \int_{\theta} P(y_{new} | \theta) P(\theta | y, I) \, d\theta
\]
Fully Bayesian density estimation

Goal:

◮ Estimate the success parameter $\pi$ associated with passivization in the new English dialect;

◮ Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

In the fully Bayesian view, we don’t summarize our posterior beliefs into a point estimate; rather, we marginalize over them in predicting the future:

$$P(y_{new} | y, I) = \int_{\theta} P(y_{new} | \theta) P(\theta | y, I) \, d\theta$$

This leads to the beta-binomial predictive model:

$$P(r | k, I, y) = \binom{k}{r} \frac{B(\alpha_1 + m + r, \alpha_2 + m - n + k - r)}{B(\alpha_1 + m, \alpha_2 + n - m)}$$
Fully Bayesian density estimation

\[ P(k \text{ passives out of 50 trials} | \mathbf{y}, \mathbf{I}) \]

- Binomial
- Beta–Binomial
Fully Bayesian density estimation

- In this case (as in many others), marginalizing over the model parameters allows for greater dispersion in the model’s predictions.
Fully Bayesian density estimation

- In this case (as in many others), marginalizing over the model parameters allows for greater dispersion in the model’s predictions.
- This is because the new observations are only conditionally independent given $\theta$—with uncertainty about $\theta$, they are linked!
Fully Bayesian density estimation

- In this case (as in many others), marginalizing over the model parameters allows for greater dispersion in the model’s predictions.
- This is because the new observations are only conditionally independent given $\theta$—with uncertainty about $\theta$, they are linked!

\[ y^{(1)}_{new} \quad y^{(2)}_{new} \quad \ldots \quad y^{(N)}_{new} \]