Bayesian models and inference by sampling

Micha Elsner

August 30, 2012

Trivia

Suggestions of papers to present: by tonight!

- I'll send out papers tomorrow
- First paper will be McMurray, Aslin and Toscano on Tuesday
- Begin discussion on Carmen (at least one comment by Monday night)

Auditors: still encouraged to make it official

Classical methods

Last lecture, described simple mixture model...
and max-likelihood estimation with EM

- Two problems with this approach:
- Local maxima (EM converges to bad solution)
- Overfitting (likelihood always higher for more complex models)

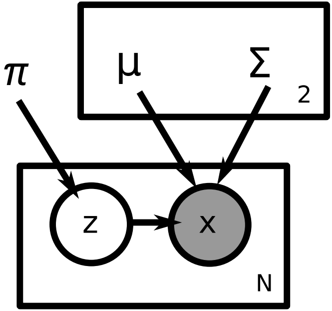
This lecture: Bayesian methods/sampling

Proposed as solutions to both these problems...
Plus a way to impose biases on the model
Warning: not always actual solutions!

Recall from last time

Our standard model of the /a/-/i/ data as a mixture of 2 gaussians

Let $X : x_0 \ldots x_N$ be the list of vowels, with $N = 90$.
Let $Z : z_0 \ldots z_N \in \{0, 1\}$ be class indicators



$$z_i \sim Bernoulli(\pi)$$
$$x_i \sim N(\mu_{z_i}, \Sigma_{z_i})$$

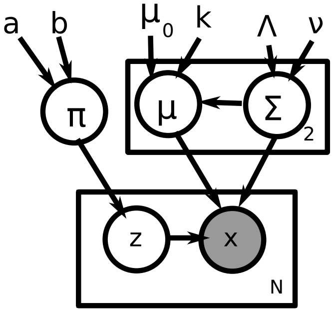
- Bernoulli: coin flip with pr of heads=π

- Recall: circles (z and x) are random variables
- No circle (μ and Σ) are parameters

Bayesian models: priors

Replace parameters with random variables

Let $X : x_0 \ldots x_N$ be the list of vowels, with $N = 90$.
Let $Z : z_0 \ldots z_N \in \{0, 1\}$ be class indicators



$$\pi \sim \mathit{Dirichlet}(a, b)$$
$$\mu, \Sigma \sim \mathit{NIW}(\mu_0, k, \Lambda, \nu)$$
$$z_i \sim \mathit{Bernoulli}(\pi)$$
$$x_i \sim N(\mu_{z_i}, \Sigma_{z_i})$$

- Will discuss specific prior distributions (*Dirichlet*, *NIW*) later

- Recall: circles (z,x, μ, Σ) are random variables
- No circle (a,b etc) are parameters (*hyper*parameters)

Why priors?

Why do priors help?

- Control overfitting (by penalizing more complex models)
- Express explicit biases (“soft” universals, markedness)
- Smooth out estimates based on insufficient data
  - If only one datapoint, frequentist $\hat{\pi}$ either 0 or 1
- Advanced: structured model– allow different clusters to share some information

“Fake data” interpretation

Can think of prior as supplying “fake observations” a priori
For instance, Dirichlet(1, 1) means:

- Pretend, in addition to our data, we have one extra /i/
- ...and one extra /a/
- So $\hat{\pi}$ can never be 0!

Parameters vs hyperparameters

Is choosing hyperparameters any less arbitrary than choosing parameters? Does it introduce bias?

- Choice of hyperparameters generally less important than choice of parameters
  - Priors grow less influential as data increases
- So maybe the same hyperparameters work across languages
  - Even if the same parameter values wouldn't
- Can try to avoid accusations of bias by:
  - Using "uninformative" priors to avoid bias
  - Or "empirical Bayes": priors near data averages

If you want to introduce (theoretically justified) biases, can ignore all this!

Replacing the likelihood

Old model: marginalizing over the latent variables

Probability of data as function of parameters:

$$P(x; \hat{\mu}, \hat{\Sigma}, \hat{\pi}) = \sum_{z} P(x|z; \hat{\mu}, \hat{\Sigma}) P(z|\hat{\pi})$$

Goal is to maximize likelihood by choosing parameters...

New model: marginalizing over the latent variables

Probability of data given hyperparameters:
The *posterior probability*:

$$P(x|a, b, \mu_0, \Lambda, k, \nu) = \int_{\mu,\Sigma,\pi} \sum_{z} P(x|z, \mu, \Sigma) P(z|\pi)$$
$$P(\mu, \Sigma|\mu_0, \Lambda, k, \nu) P(\pi|a, b)$$

How we use the posterior

Maximum a posteriori (MAP) inference

Replace integral with maximization (“max-likelihood with priors”)

$$\hat{\mu}, \hat{\Sigma}, \hat{\pi} = argmax_{\mu,\Sigma,\pi} \sum_z P(x|z;\mu,\Sigma)P(z|\pi)$$
$$P(\mu|\mu_0,\Lambda,k,\nu)P(\pi|a,b)$$

Use entire posterior distr.

For instance, what is expected mean of /i/ category given our data?

$$E[\mu_{/i/}|x,a,\ldots] =$$
$$\int_{\mu,\Sigma,\pi} \mu_{/i/} \sum_z P(x|z,\mu,\Sigma)P(\mu|\ldots$$

Requires us to approximate (complicated) integral

Why look at the posterior instead of MAP?

Statisticians will give you reasons:

- Look only at what we really care about
  - What is the average mean (over all variances)...
  - vs the mean at a *particular* variance
- Sometimes MAP solution is an outlier (requires very specific parameter setting; typical solution more robust)
- Or posterior could be multimodal (different good clusterings)
  - (Difficult to represent this: mean won't do)

Honestly I think statisticians care more than we do...
But the sampling technique we will see is very popular
...and can also be used for MAP estimates.

Approximating an integral by sampling

Want to know the mean of /i/ category:

$$E[\mu_{/i/}|x, a, \ldots] = \int_{\mu,\Sigma,\pi} \mu_{/i/} \sum_{z} P(x|z, \mu, \Sigma)P(\mu| \ldots$$

Can't evaluate analytically, so:

- Take many (M) samples from $P(\mu|x, a \ldots)$
- Use mean over samples
  - $\frac{1}{M} \sum_{m=1}^{M} \mu_{/i/}^{m}$
- As M grows, this approaches true expectation
- Expectation also usually close to MAP (not always!)

Pause for deep breath

- Bayesian model has prior distributions on parameters
  - Deal with overfitting, express prior beliefs
- *Integrate* over all possible parameters to get expected values
- Can take integral by *sampling* from posterior given data

New goal: take samples from model posterior

How do we sample?

Basic distributions

Easy to sample from most textbook distributions:
Library functions in your favorite language:

x = random.normalvariate(mu, sigma)

(Typically by inverting cumulative distr. fn)

To sample from more complex distribution, build on these basic functions

Gibbs sampling

Gibbs sampling

To sample from joint distribution over many RVs:

$$P(z_1 \dots z_n, \mu, \Sigma, \pi | x_1 \dots x_n, a, b \dots)$$

- Initialize at random
- Do forever:
  - For each RV in turn, compute distribution conditioned on other RVs:

$$P(z_1 | z_2 \dots z_n, \mu, \Sigma, \pi, x_1 \dots)$$

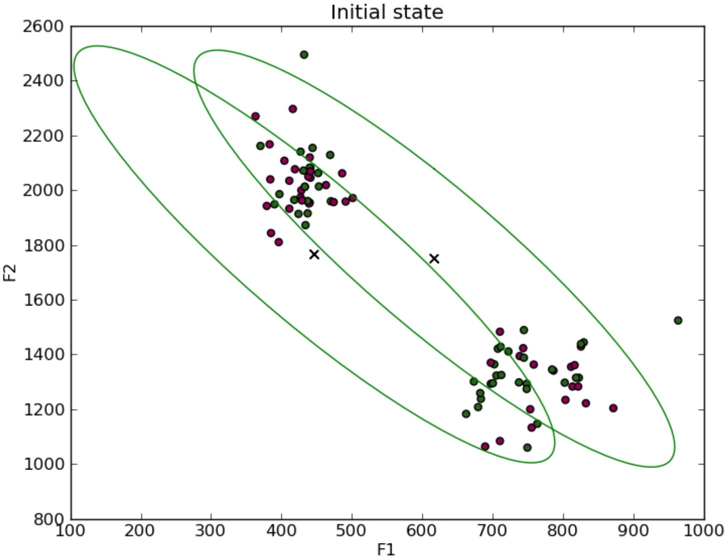
  - Sample new value from that distribution:

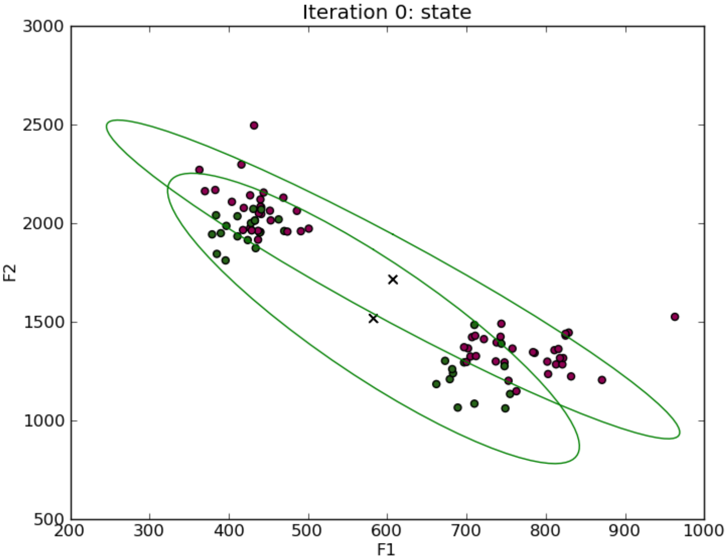
$$z_1 \leftarrow P(z_1 | z_2 \dots)$$

  - Entire state $(z_1 \dots z_n, \mu, \Sigma, \pi)$ is sample from P

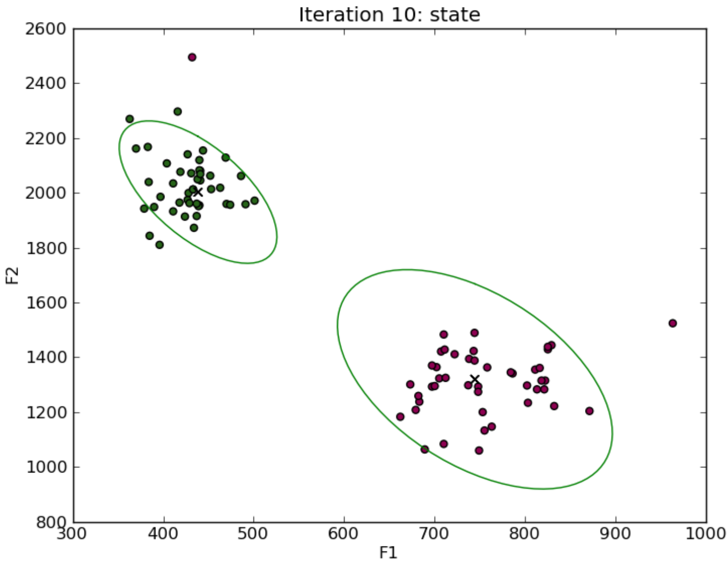
Gibbs vs EM

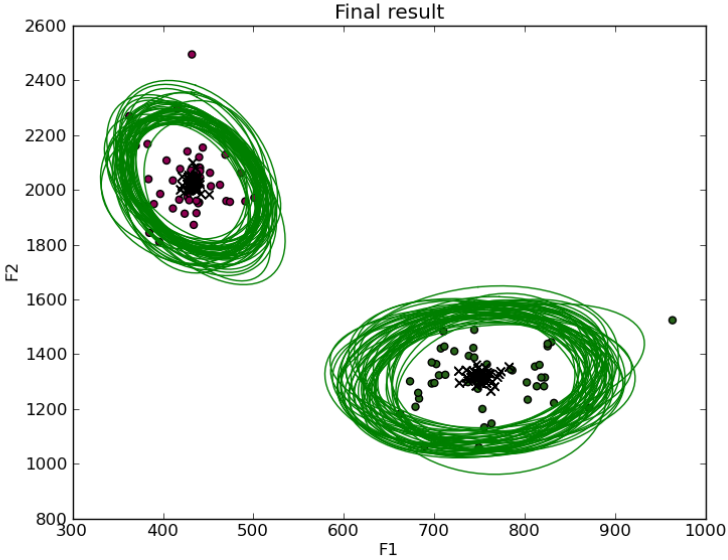
- Like EM, Gibbs spends most of time computing $P(z_i = 0 | x_i, \mu, \Sigma, \pi)$
  - Unlike EM, Gibbs also needs conditionals on the parameters, eg: $P(\pi | z, a, b)$
  - Unlike EM, Gibbs *samples* instead of maximizing or taking expectations
- Posterior usually increases on avg. (similar to likelihood)
  - But can decrease slightly due to random chance...
- No explicit test for convergence
- Gibbs can escape local maxima
  - But this happens with very low probability


Initial state
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
F2
100
200
300
400
500
600
700
800
900
1000
F1



Iteration 0: state
3000
2500
2000
1500
1000
500
200
300
400
500
600
700
800
900
1000
F1
F2


Iteration 10: state
F1
F2
300
400
500
600
700
800
900
1000
800
1000
1200
1400
1600
1800
2000
2200
2400
2600


Final result
F2
F1
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
300
400
500
600
700
800
900
1000

Choosing a prior

Besides the *model-specific* question of which prior to choose, there are *mathematical* issues

- Choosing the right kind of prior distribution makes programs more efficient...
- And easier to code

Consider π in our model (π is prior $p(z = 0)$)
Standard Gibbs sweep requires:

$$\pi \sim P(\pi | z_1 \dots z_n, a, b)$$

If this is a toolbox function (eg `random.beta`) we are done...
Otherwise, can be very difficult

Conjugacy

What is $P(\pi|z_1 \dots z_n)$? Use Bayes' rule....

$$P(\pi|z_1 \dots z_n, a, b) \propto P(z_1 \dots z_n|\pi)P(\pi|a, b)$$

($\propto$: read "proportional to".)
Since π is $p(z_i = 0)$...

$$P(z_1 \dots z_n|\pi)P(\pi|a, b) = \pi^{\#(z_i=0)}(1-\pi)^{\#(z_i=1)}P(\pi|a, b)$$

Dirichlet distribution (for a two-component vector)

Distribution over $0 < \pi < 1$ and $0 < 1 - \pi < 1$, such that:

$$P(\pi|a, b) \propto \pi^{a-1}(1-\pi)^{b-1}$$

There's a toolbox routine for sampling from $P(\pi|a, b)$
(Note: this special-case of the k-probability Dirichlet is also called the Beta distribution)

Conjugacy (2)

$$P(z_1 \dots z_n|\pi)P(\pi|a,b) = \pi^{\#(z_i=0)}(1-\pi)^{\#(z_i=1)}P(\pi|a,b)$$

If we pick $P(\pi|a,b)$ as Dirichlet, so $P(\pi|a,b) \propto \pi^{a-1}(1-\pi)^{b-1}$

$$\pi^{\#(z_i=0)}(1-\pi)^{\#(z_i=1)}P(\pi|a,b) =$$
$$\pi^{\#(z_i=0)}(1-\pi)^{\#(z_i=1)}\pi^{a-1}(1-\pi)^{b-1} =$$
$$\pi^{\#(z_i=0)+a-1}(1-\pi)^{\#(z_i=1)+b-1}$$

Thus, $P(\pi|z_1 \dots z_n, a, b) \sim Dirich(\pi|u,v)$ for $u = \#(z_i = 0) + a$ and $v = \#(z_i = 1) + b$
And we use the toolbox routine!

Conjugacy (3)

Conjugate prior

A prior distribution on a parameter, which, multiplied by the probability of a set of data, yields a posterior distribution in the same family

- The Dirichlet is the conjugate prior for the Bernoulli (coin-flip) data likelihood
- A large family of useful distributions (the exponential family) all have conjugate priors
  - For the Bernoulli (coin flip), categorial (die roll), multinomial (many die rolls): Dirichlet
  - For the Gaussian: Gaussian mean, inv. Wishart covariance
  - Others in any stats textbook

Integrating out parameters

Choosing a conjugate prior lets us do another trick:
Instead of computing $P(\pi|z_1 \dots z_n, a, b)$...
And then $P(z_1|\pi)$ etc...
We can compute directly:

$$P(z_1 = 0|z_2 \dots z_n, a, b) = \int_{\pi} P(z_1 = 0|\pi) P(\pi|z_2 \dots z_n, a, b)$$

$$= \int_{\pi} \pi \mathit{Dirich}(\pi|\#(z_{2\dots n} = 0) + a, \#(z_{2\dots n} = 1) + b)$$
$$= \mathit{mean}(\mathit{Dirich}(\pi|\#(z_{2\dots n} = 0) + a, \#(z_{2\dots n} = 1) + b))$$
$$= \frac{\#(z_{2\dots n} = 0) + a}{\#(z_{2\dots n} = 0) + a + \#(z_{2\dots n} = 1) + b}$$

(Often people use notation like z_{-i} or $z_{/i}$ for "all z excluding z_i")

Dirichlet processes

The integration trick lets us deal with distributions with infinitely many parameters...
The two-dimensional Dirichlet

$$Dirich(\pi|a, b) \propto \pi^{a-1}(1-\pi)^{b-1}$$

The k-dimensional Dirichlet

$$Dirich(\pi_1, \pi_2 \ldots \pi_k|\alpha_1 \ldots \alpha_k) \propto \pi_1^{\alpha_1 - 1}\pi_2^{\alpha_2 - 1} \ldots$$

Suppose we let all the α_i be equal fractions of A (ie $\alpha_i = \frac{A}{k}$)...
and then $k \to \infty$

Dirichlet process

The limiting distribution is defined as $DP(\pi_1 \ldots |A)$

Dirichlet processes (2)

We can't work directly with $DP(A)$, but...
Recall we used integration to get:

$$P(z_i = 0 | z_{-i}, a, b) \propto \#(z_{-i} = 0) + a$$

The equivalent still holds.
Since we set $\alpha_1 = \frac{A}{k}$, the limit of α_1 is 0...
So under a DP, if group 1 is represented in z_{-i}...

$$P(z_1 = 0 | z_{-1}, A) \propto \#(z_{-1} = 0) + \alpha_1 (= 0)$$

The α_i of the (infinite) groups unrepresented in z_{-i} sum up to A, so:

$$P(z_i \text{ unrepresented}) \propto A$$

Chinese restaurant process

Chinese restaurant process

This representation of the posterior $p(z_i|z_{-i})DP(A)$ is called the Chinese Restaurant process $CRP(z_1 \ldots z_n|A)$
At any time, there are k occupied "tables" (groups such that some z_i "customers" are in that group)
And an infinite number of unoccupied "tables" with total probability $\propto A$

Conditioned on z_{-i} with k groups:

$$
\begin{aligned}
P(z_i = g) &\propto \#(z_{-i} = g) \qquad\qquad g \leq k \\
P(z_i = k + 1) &\propto A
\end{aligned}
$$

Properties of CRP

- A controls the dispersion or diversity
  - Larger A, more groups on average
- A priori average of $A\,log(n)$ clusters for n observations
  - Can be overridden by observed data
- “Rich-get-richer” dynamics
  - Large groups attract new observations

The CRP is a principled way of comparing models with more versus fewer clusters...
Unlike max-likelihood

So...

- Bayesian models offer control of overfitting
  - And a way to specify prior beliefs
- Popular inference method is Gibbs sampling
  - Randomized iterative algorithm
  - Computes expected values of things
  - Theoretically escapes local maxima, but not practically
- Choosing conjugate priors leads to efficient algorithms
- The Dirichlet process is a prior over category indicators that allows an unbounded number of categories

Code is online. Remember to send paper presentation preferences!