October 17, 2017

CS 361: Probability & Statistics

Inference: Max likelihood & Bayesian

Setting up the inference problem

Inference at its most broad just means drawing conclusions from data

We will be looking at the problem of point inference or making **point estimates**. Which is a fancy way of saying, looking at a dataset and coming up with a number(s)

For example we might have data that we know is coming from some unknown Normal distribution. How do we infer what the parameters of this distribution are (mean, standard deviation) from the data itself?

Alternatively, we might be interested in the average height of people at UIUC. This is not a random number, but if we don't have the means to measure everyone at the university, we may have to look at a sample of population and use it to guess what the average of the whole population is

Estimating model parameters

Suppose we have a dataset *D={x_i}*

Furthermore, suppose we know the type of distribution that models the data. E.g. Bernoulli, Binomial, Poisson, Normal, etc.

Suppose that we do not know the parameters of the model

We will use the letter theta θ to represent the model parameters

We will be trying to answer the following question:
What is a good value for θ, given the data?

Examples

If our data is given by a Bernoulli, Geometric, or Binomial random variable we are trying to find a good $\theta = p$

For the case of multinomial data we are looking for

$$\theta = (p_1, p_2, \ldots, p_k)$$

For a Poisson or exponential distribution, we are looking for a good estimate of

$$\theta = \lambda$$

For a normal distribution we are looking to estimate a good

$$\theta = (\mu, \sigma)$$

Overall, theta is an abbreviated way to refer to the parameters of a model we are trying to figure out

Estimating model parameters via maximum likelihood

Which theta to choose?

When we know θ, and we have a set of data, it is possible to get a numerical answer to the question "what was the probability of seeing that data, given theta"

$$P(\mathcal{D}|\theta)$$

When we have a dataset but don't know theta we can still write this quantity, but it will be a function of theta. We call this expression the **likelihood function**

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$

One procedure for estimating the parameters of the model is to choose the theta that maximizes this expression. We call this the **max likelihood** estimate

Likelihood

Notice that the likelihood function is not a probability distribution over theta

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$

For instance, if we know our data is a single sequence of coin flips, we know that the number of heads may be well-modeled by a binomial distribution, with parameters

$$\theta = (N, p)$$

If the sequence we observe is HHHHH, we know that N=5, so there's no inference to do there. But evaluating the likelihood function for $\theta = p$

$\mathcal{L}(\theta) = \binom{5}{5}\theta^{5}(1-\theta)^{0}$ and we have for example $\mathcal{L}(1) = 1$ and $\mathcal{L}(.9) \approx 0.59$

L(1) + L(.9) > 1, so L is definitely not a probability distribution

Likelihood

We will assume that the data that we are observing are IID — independent and identically distributed. So we will be able to write

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = \prod_{i \in \text{dataset}} P(d_i|\theta).$$

As a notational convention, we will write $\hat{\theta}$ to indicate our actual, calculated estimate of θ

Thus our maximum likelihood estimate can be written as

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

Binomial likelihood

In *N* independent coin flips, we observe *k* heads. What is the maximum likelihood estimate for θ ?

We want to calculate

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

What is the likelihood function in this setup?

$$\mathcal{L}(\theta) = \binom{N}{k} \theta^{k} (1-\theta)^{N-k}$$

Binomial likelihood

We need to take a derivative and set equal to 0

$$\mathcal{L}(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \binom{N}{k} \left[k\theta^{k-1}(1-\theta)^{N-k} - (N-k)\theta^k(1-\theta)^{N-k-1} \right]$$

Set $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ and solve for theta

$$k\theta^{k-1}(1-\theta)^{N-k} = (N-k)\theta^k(1-\theta)^{N-k-1}$$

Binomial likelihood

$$k\,{\cancel{\theta^{k}}}^{-1}(1-\theta)^{\cancel{N-k}} = (N-k)\theta^{\cancel{k}}\cancel{(1-\theta)^{N-k-1}}$$

giving

$$k(1-\theta) = (N-k)\theta$$

or

$$k - \cancel{k}\theta = N\theta - \cancel{k}\theta$$

Thus our max likelihood estimate is given by

$$\hat{\theta} = \frac{k}{N}$$

Geometric likelihood

Suppose we flip a coin, stopping after we see a head. We do this and wind up having to do *N* flips. What is the maximum likelihood estimate of θ ?

Once again, we want to compute

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

This time, we have

$$\mathcal{L}(\theta) = (1 - \theta)^{N-1}\theta$$

Geometric likelihood

We want to choose a theta to maximize L

$$\mathcal{L}(\theta) = (1-\theta)^{N-1}\theta$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = (1-\theta)^{N-1} - (N-1)(1-\theta)^{N-2}\theta$$

Taking $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ we get

$$(N-1)(\cancel{1-\theta)^{N-2}}\theta = (1-\theta)^{\cancel{N-1}}$$

$$\theta N - \cancel{\theta} = 1 - \cancel{\theta}$$

Giving an MLE estimate of $\hat{\theta} = \frac{1}{N}$

Log likelihood

Since the logarithm is a monotonic function, maximizing the log of the likelihood will give us the same theta as if we maximize the likelihood

$$\arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \log \mathcal{L}(\theta)$$

Logarithms allow us to break products up into sums which will make some computations easier

$$\log P(\mathcal{D}|\theta) = \log \prod_{i \in \mathrm{dataset}} P(d_i|\theta) = \sum_{i \in \mathrm{dataset}} \log P(d_i|\theta)$$

Poisson likelihood

Suppose we observe some event of interest over *N* intervals each of the same fixed length and that the number of events that we observe in interval *i* is given by *n_i*. We model this situation with a Poisson random variable and we want to know the max likelihood estimate for λ

Getting the derivative of our likelihood might be tricky since we have

$$\mathcal{L}(\theta) = \prod_{i \in \text{intervals}} P(\{n_i \text{ events}\}|\theta) = \prod_{i \in \text{intervals}} \frac{\theta^{n_i} e^{-\theta}}{n_i!}$$

But we can easily maximize

$$\log \mathcal{L}(\theta) = \sum_i (n_i \log \theta - \theta - \log n_i!)$$

Poisson likelihood

Let’s differentiate the following with respect to theta

$$\log \mathcal{L}(\theta) = \sum_i (n_i \log \theta - \theta - \log n_i!)$$

to get

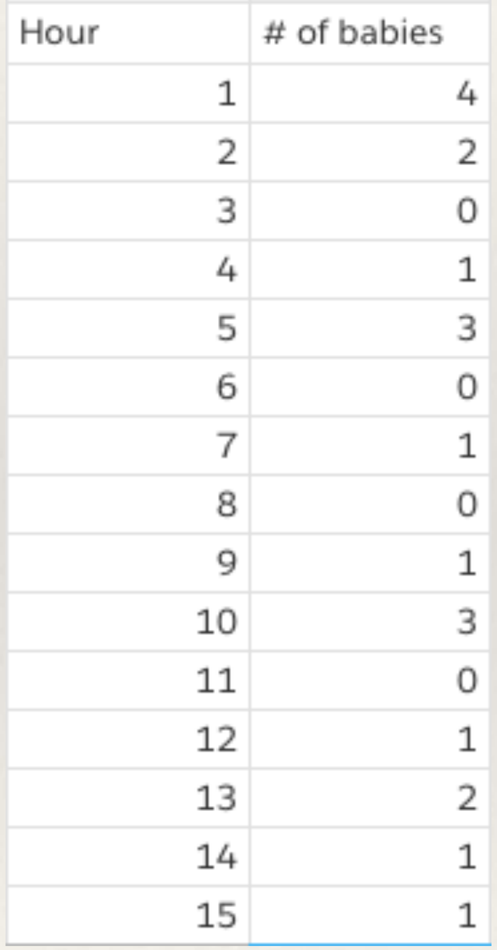
$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \sum_i \left(\frac{n_i}{\theta} - 1 \right)$$

Setting $\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = 0$ we get $\frac{1}{\theta} \left(\sum_i n_i \right) - N = 0$

Solving for theta we get our MLE $\hat{\theta} = \frac{\sum_i n_i}{N}$

Example

Suppose we work in the maternity ward of a large hospital and we begin to write down how many babies are born each hour and get the following dataset



Hour	# of babies
1	4
2	2
3	0
4	1
5	3
6	0
7	1
8	0
9	1
10	3
11	0
12	1
13	2
14	1
15	1

If we suppose the number of babies born each hour is governed by a Poisson distribution, what is the maximum likelihood estimate for its intensity, given the data?

On the last slide we had

$$\hat{\theta} = \frac{\sum_i n_i}{N}$$

So our MLE for λ is 20/15

Normal likelihood

Assume we observe *N* data items x_1, x_2, … , x_N thought to conform to a normal distribution. What is the max likelihood estimate for the mean of this normal distribution given the data?

We have

$$\mathcal{L}(\theta) = P(x_1, x_2, \ldots, x_N | \theta, \sigma)$$

Or

$$\mathcal{L}(\theta) = P(x_1 | \theta, \sigma) P(x_2 | \theta, \sigma) \ldots P(x_N | \theta, \sigma)$$

Which if we use the distribution of the normal, gives us

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2} \right)$$

Normal likelihood

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right)$$

would be a pain to work with, so we look at the log-likelihood given by

$$\log \mathcal{L}(\theta) = \left(\sum_{i=1}^{N} -\frac{(x_i-\theta)^2}{2\sigma^2}\right) + \text{term not depending on } \theta$$

Differentiating with respect to theta, we get

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \sum_{i=1}^{N} \frac{2(x_i-\theta)}{2\sigma^2}$$

Normal likelihood

Setting the derivative of the log likelihood equal to 0 we get

$$0 = \sum_{i=1}^{N} \frac{2(x_i - \theta)}{2\sigma^2}$$

Simplifying

$$0 = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \theta\right)$$

We get

$$N\theta = \sum_{i=1}^{N} x_i$$

Giving an MLE of

$$\hat{\theta} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Normal likelihood

Suppose we have N data items as before but want a maximum likelihood estimate for the standard deviation of the normal distribution our data are coming from

This time we have

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x_i - \mu)^2}{2\theta^2}\right)$$

So our log likelihood is given by

$$\log \mathcal{L}(\theta) = \left(\sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2\theta^2}\right) - N \log \theta + \text{term not depending on } \theta$$

Normal likelihood

$$\log \mathcal{L}(\theta) = \left(\sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2\theta^2} \right) - N \log \theta + \text{term not depending on } \theta$$

differentiating with respect to theta we get

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \frac{-2}{\theta^3} \sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2} - \frac{N}{\theta}$$

Simplifying and setting equal to 0

$$0 = \sum_{i=1}^{N} (x_i - \mu)^2 - N\theta^2$$

For an MLE of

$$\hat{\theta} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Maximum likelihood: drawbacks

A couple of things might trip up max likelihood estimation:

1) Finding the maximum of some functions can be quite hard

2) If we don’t have a large amount of data, we might incorrectly estimate certain model parameters

For example, the MLE for p_i in a multinomial distribution is n_i/N if we have observed n_i instances of face i in N rolls of a die. If we have observed zero 3s in 8 rolls of a die, is it always safe to assume $p\_3=0$?