Maximum Likelihood Estimation

Given a collection of data, what is the underlying distribution
→ Assuming a distribution, what are the likeliest parameters for that distribution.

Suppose we've conducted an experiment and our collected data is a list of values $x_1, x_2, x_3, \ldots x_n$.

Suppose there is a function f for the probability of these data to be measured. This function is parametrized by θ:

$$\text{Prob}(\text{data}) = f(x_1, x_2, x_3, \ldots, x_n \mid \theta)$$ ← joint prob of all the data simultaneously given a (set of) parameters θ.

Now let's make things easier: Assume each data point can occur independently from all others and each follows the same distribution (independent, identically distributed or "iid").

$$\hookrightarrow f(x_1, x_2, \ldots x_n \mid \theta) = f(x_1 \mid \theta) \cdot f(x_2 \mid \theta) \cdots f(x_n \mid \theta)$$
$$= \prod_{i=1}^{n} f(x_i \mid \theta)$$

Ok, our goal is to find θ, we're given the data. We need to flip this around: what is the likelihood of θ, given the data? No problem, flip!

$$\mathcal{L}(\theta \mid x_1, x_2, \ldots x_n) = f(x_1, x_2, \ldots x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

(whoa are you allowed to just do that?)

Moving forward, we want to find the θ that maximizes $\mathcal{L}(\theta \mid \text{data})$. (called the maximum likelihood estimator, $\hat{\theta}$)

Often it's easier to maximize the log-likelihood ℓ which is maximized at the same place:
$$\ell(\theta \mid \text{data}) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$$

MLE cont

Some times we can find the $\hat{\theta}$ easily with calculus, other times we need a numerical method.

Using calculus, compute the derivative $\frac{\partial \ell}{\partial \theta}$, set equal to 0, solve.

Let's do a specific example:

Poisson distribution:

Prob for a certain number of events to occur if the average rate of events is known and the prob. for the next event to occur does not depend on the time since the previous event.

$$P(X=k;\lambda) = P(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Let's assume the x_i's are drawn from a poisson. Then we need to find $\hat{\lambda}$:

$$f(x_i \mid \theta) \Rightarrow f(x_i \mid \lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Compute the (log)-likelihood:

$$\mathcal{L}(\theta \mid \text{data}) = \prod_{i=1}^{n} f(x_i \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}$$

$$\ell(\theta \mid \text{data}) = \ln \mathcal{L}(\theta \mid \text{data}) = \ln e^{-n\lambda} + \sum_{i=1}^{n} \ln \lambda^{x_i} - \sum_{i=1}^{n} \ln(x_i!)$$

$$= -n\lambda + \sum_{i=1}^{n} x_i \ln \lambda - (\ \ )$$

$$= -n\lambda + \ln\lambda \sum_{i=1}^{n} x_i - (\ \ )$$

Now we find an equation for the maximum as a function of λ, and solve it for $\hat{\lambda}$:

$$\frac{\partial \ell}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left(-n\lambda + \ln\lambda \sum x_i - (\ )\right) = -n + \underbrace{\ln\lambda \cdot \frac{\partial}{\partial\lambda}\sum x_i}_{0!} + \sum x_i \cdot \frac{1}{\lambda} - 0$$

$\boxed{2}$

Solving this for $\hat{\lambda}$

$$-n + \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The MLE $\hat{\lambda}$ for poisson is the <u>average</u> of the data.

So if we are given a bunch of integer X-data, we can compute the mean, and draw $P(X=k,\hat{\lambda})$ on top of the histogram and see if it's close.

Recall I did something sneaky: $\mathcal{L}(\theta | data) = f(data | \theta)$

If these are both probabilities you can't just do that, you need to use <u>Bayes Theorem</u>

<u>Bayes</u>:

- The joint prob of two events A and B both occurring is commutative:

$$P(A \cap B) = P(B \cap A) \quad \text{(b/c the sets } A \cap B \text{ and } B \cap A \text{ are the same)}.$$

- We can write the joint prob: $P(A \cap B) = P(A|B) \cdot P(B)$ (intuitive...)

Combining those and rearranging gives us Bayes thm:

$$P(A \cap B) = P(B \cap A)$$

$$P(A|B)P(B) = P(B|A)P(A)$$

$$\boxed{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}} \quad \text{Bayes}$$

This is how one can "flip around" a prob.

Now what does this have to do w/ MLE?

Instead of $\mathcal{L}(\theta \mid \text{data}) = f(\text{data} \mid \theta)$ we use Bayes:

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)\, P(\theta)}{P(\text{data})}$$

$$= \frac{f(x_1, x_2, \cdots x_n \mid \theta)\, P(\theta)}{P(x_1, x_2, \ldots, x_n)}$$

Now the denominator is indep. of θ, meaning it's just a constant that won't change the maximum

So the difference really is the $P(\theta)$. Now, if we assume that is also a constant, what happens? It won't affect the location of the max either, and

$\mathcal{L}(\theta \mid \text{data}) = P(\theta \mid \text{data})$ except for a mult. const.

Assuming $P(\theta) = \text{const}$ means all θ's are equally likely.

And that is the assumption behind MLE.

"The ML estimator is equal to the Bayesian estimator given a uniform prior distribution of the parameters."

$\rightarrow$ MLE is a special case.