Stat 330 - Tutorial 10

This week's tutorial covers two main topics: likelihood and sufficiency. Before getting into the main discussion, we start by defining statistical models and show why the idea of maximum likelihood estimation arises naturally from these models.

Statistical Models

Recall that a probability model is made up of a sample space, a collection of events and a probability measure. A statistical model is defined as a collection of probability measures on a single sample space, where the family is indexed by some parameter, θ. We denote a statistical model by $\{P_\theta : \theta \in \Theta\}$, where Θ is the set of all possible parameter values.

When analyzing data, we are forced to assume ~~some statistical~~ that the true distribution of the data is in some statistical model (even very general nonparametrics usually make structural or smoothness assumptions). The statistician's task is then to choose which value(s) of θ is/are supported by the ~~ob~~ observed data. One of the most popular ways to ~~do this~~ choose θ is called maximum likelihood estimation (MLE). The idea behind MLE is to choose the value of θ that assigns the most probability (or density) to the points we observed. Let's go into some detail about how this works.

Maximum Likelihood Estimation

The MLE procedure is the same for discrete and continuous R.V.s, but our notation doesn't let us talk about both at the same time. We will exclusively discuss continuous R.V.s, but if yours is discrete, just replace densities with PMFs and do exactly the same thing.

Let X be a R.V. with density $f(x;\theta)$ when the parameter value is θ. If we observe $x_1, \dots, x_n$ as iid draws from $f(x;\theta)$, the joint density of our sample is $\prod_{i=1}^{n} f(x_i;\theta)$. For a single θ, this is a function of $\underline{x} = (x_1, \dots, x_n)$. However, after observing data, we can examine how the joint density changes at different θ. When treated as a function of θ, we call this expression the likelihood function.

$$\mathcal{L}(\theta; \underline{x}) := \prod_{i=1}^{n} f(x_i;\theta)$$

②

If we have to choose a value of θ that best ~~represents~~ matches the data, it seems reasonable to choose the one that maximizes the likelihood, $\mathcal{L}$. This gives us a value of θ, $\hat{\theta}_{MLE}$ or just $\hat{\theta}$, which was chosen based on the data. Such a value is called an estimator or a statistic. This particular estimator is called a maximum likelihood estimator, or MLE (we use the acronym MLE to refer to multiple things, ~~the~~ an estimator and the procedure used to obtain that estimator, but you should be able to tell from context which one is meant ~~...~~ (or it won't matter)).

Assuming that you're sold on MLEs being a good idea, now we need to know how to calculate them. Most parameters you will encounter are just real numbers, so we can use machinery from calculus. Specifically, we differentiate the likelihood w.r.t. θ and solve the equation $\frac{\partial \mathcal{L}}{\partial \theta} = 0$. In practice, it is usually easier to work with the log of the likelihood,

$\ell(\theta;\underline{x}) := \log[\mathcal{L}(\theta;\underline{x})]$. Optimizing $\mathcal{L}$ ~~and~~ or ℓ will give the same result because log is an increasing transformation.

e.g. 1: Let $x_1,\dots,x_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Find $\hat{\lambda}_{MLE}$.

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

$$\mathcal{L}(\lambda;\underline{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

$$\ell(\lambda;\underline{x}) = \log[\mathcal{L}(\theta;\underline{x})] = \sum_{i=1}^{n} \log(\lambda e^{-\lambda x_i})$$

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

Differentiating gives

$$\frac{\partial \ell(\lambda;\underline{x})}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Setting this equal to zero and solving gives

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

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

Therefore, $\hat{\lambda}_{MLE} = 1/\bar{x}$.

///

Technically, the procedure we just did isn't complete. We have shown that $\hat{\lambda}_{MLE}$ is a critical point, but we don't yet know that it is a maximizer.

In general, when ~~doing MLEs~~ finding MLEs, it isn't enough to find a critical point of $\mathcal{L}$ (or ℓ). We also need to use the second derivative test to check that our candidate MLE is actually a maximizer. Specifically, we need to check that the second derivative of $\mathcal{L}$ (or ℓ) is negative at $\hat{\theta}_{MLE}$.

e.g. 1 cont.

$$\frac{\partial^2 \ell(\lambda;\underline{X})}{\partial\lambda^2} = \frac{\partial}{\partial\lambda}\left[\frac{\partial \ell(\lambda;\underline{X})}{\partial\lambda}\right] = \frac{-n}{\lambda^2}$$

~~Notice that $\hat{\lambda}_{MLE} = \frac{1}{\bar{X}} > 0$ for all possible samples ($\bar{X} \geq 0$ since each $X_i \geq 0$), so~~

$$\cancel{\left.\frac{\partial^2 \ell(\lambda;\underline{X})}{\partial\lambda^2}\right|_{\lambda=\hat{\lambda}_{MLE}} = \frac{-n}{\hat{\lambda}_{MLE}^2} = -n\bar{X}^2 < 0}$$

Notice that $\bar{X} > 0$ with probability 1, and we don't need to worry about anything that happens with probability zero. Therefore,

$$\left.\frac{\partial^2 \ell(\lambda;\underline{X})}{\partial\lambda^2}\right|_{\lambda=\hat{\lambda}_{MLE}} = \frac{-n}{(\hat{\lambda}_{MLE})^2} = -n\bar{X}^2 < 0$$

So, by the second derivative test, $\hat{\lambda}_{MLE}$ ~~maximizer~~ is a maximizer of ℓ (and also $\mathcal{L}$). ///

The function $\frac{\partial \ell(\theta;\underline{X})}{\partial\theta}$ is called the score function, and the equation $\frac{\partial \ell(\theta;\underline{X})}{\partial\theta} = 0$ is called the score equation.

You have seen likelihood and MLEs before. Now we move onto something you're probably less familiar with: sufficiency.

//

Sufficiency and Sufficient Statistics

When we estimate the mean in a normal model, with σ^2 known, we ~~we~~ don't actually use the whole dataset. Specifically, we only really need to know $\bar{X}$. [redacted] Similarly, when estimating θ in a $\text{Unif}(0,\theta)$ model, we use $X_{(n)}$, the maximum of our sample, ~~that~~ and ignore all the other points. These are both examples of a general phenomenon called sufficiency. ~~In general~~ Informally, a statistic is called sufficient for a statistical model if, once we know the value of the statistic, the rest of the data doesn't help us choose a value of θ. In math, a statistic, T, is called a sufficient statistic for a statistical model with parameter θ and likelihood function $\mathcal{L}(\theta;\underline{X})$ if, for any two samples $\underline{X}_1, \underline{X}_2$ with $T(\underline{X}_1)=T(\underline{X}_2)$,

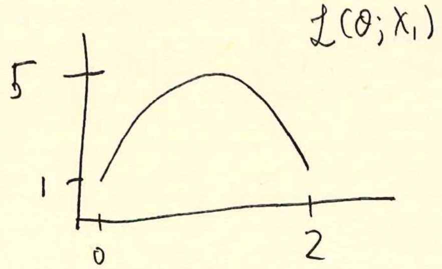
$$\mathcal{L}(\theta;\underline{X}_1)=c(\underline{X}_1,\underline{X}_2)\,\mathcal{L}(\theta;\underline{X}_2) \quad (I)$$

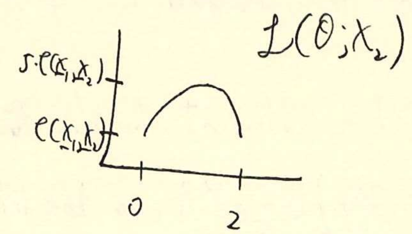
for all $\theta\in\Theta$, where c is a function of the two samples that does not depend on θ.

Let's unpack this definition a little. Provided that $\mathcal{L}(\theta;\underline{X}_2)\neq 0$, (I) says that the likelihood ratio,

$$\frac{\mathcal{L}(\theta;\underline{X}_1)}{\mathcal{L}(\theta;\underline{X}_2)}=c(\underline{X}_1,\underline{X}_2)$$

does not depend on θ. ~~Note that we can only use a function to~~ This means that, as a f'n of θ, the only difference between the likelihoods ~~of~~ based on $\underline{X}_1$ and $\underline{X}_2$ is the scale of the vertical axis.





In particular, both likelihood functions have the same maximizer, although the values at which they are maximized is different.

From the definition, it's not at all obvious how to find sufficient statistics. Fortunately, there is a theorem that makes this much easier.

Th'm 6.1.1 (Factorization Th'm):

~~If the density or probab~~ If the likelihood function of a statistical model factors ~~and~~ in the following way: $\mathcal{L}(\theta; \underline{X}) = h(\underline{X})\, g(\theta, T(\underline{X}))$, then the statistic $T(\underline{X})$ is sufficient. Said differently, if the likelihood factors into a part that doesn't depend on θ, & a part that depends on θ, but only on the data ~~the~~ through a function, T, then the function T is a sufficient statistic.

E.g. 2: Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$. Find a sufficient statistic for this model.

First, notice that our statistical model has two parameters. This suggests that we will probably need at least two terms in our sufficient statistic ~~([redacted] not need to be one-dimensional).~~ (sufficient statistics can be multivariate). The density of our R.V.s is

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

The likelihood function ~~is therefore based on a sum~~ is therefore

$$\mathcal{L}(\alpha, \beta; \underline{X}) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-x_i/\beta}$$

$$= \frac{1}{[\Gamma(\alpha)\beta^\alpha]^n} \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} x_i\right)$$

The ~~Fact~~ Factorization Th'm then says that the function $T(\underline{X}) = \left(\sum_{i=1}^{n} x_i, \prod_{i=1}^{n} x_i\right)$ is sufficient in this model. Notice that the $h(\underline{X})$ term ~~in this~~ doesn't occur in this model. ◇

E.g. 3: Let $X_1, \dots, X_n \sim N(\mu, 1)$. Find a sufficient statistic.

The density here is $f(x; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2}\right]$, so the likelihood is

$$\mathcal{L}(\mu; X) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

By the factorization Theorem, the statistic $T(\underline{X}) = (x_1, \dots, x_n)$ is sufficient (check that this works!). While this is true, it's not very satisfying or useful. This example motivates the following example of minimal sufficiency.

A statistic is called minimal sufficient for a statistical model if we can always compute the value of this statistic given the likelihood function. Minimal sufficient statistics will often be MLEs, but this doesn't always have to be the case.

E.g. 3 cont. Find a minimal sufficient statistic for ~~the no~~ the normal model with $\sigma^2=1$.

Recall that the likelihood function is

$$\mathcal{L}(\mu;\underline{X}) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

The corresponding log-likelihood is

$$\ell(\mu;\underline{X}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

The score f'n is

$$\nabla_\mu \ell(\mu;\underline{X}) = \sum_{i=1}^{n}(x_i-\mu)$$

Solving the score equation, we get

$$\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0$$

$$\sum_{i=1}^{n} x_i = n\hat{\mu}$$

$$\hat{\mu} = \bar{X}$$

We computed $\bar{X}$ by solving the score equation. No special assumptions were required to do so in this model, so we can always compute $\bar{x}$ from the likelihood. However, we have not yet shown that $\bar{x}$ is sufficient.

$$\mathcal{L}(\mu;\underline{X}) = (2\pi)^{-n/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}x_i^2 + \mu\sum_{i=1}^{n}x_i - \frac{1}{2}\mu^2\right]$$

$$= \left[(2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right)\right]\left[\exp(\mu n\bar{X})\exp\left(-\frac{1}{2}\mu^2\right)\right]$$

$$= h(\underline{X})\, g(\mu, \bar{X})$$

Therefore, by the factorization Theorem, $\bar{X}$ is sufficient for this model. We checked ~~either~~ for minimal sufficiency above, so we now have that $\bar{X}$ is minimal sufficient for the normal model with σ^2 known.

◇