MATH 2B RECITATION 2/9/12

ALDEN WALKER

1. Generating Random Numbers

Suppose that we would like to sample a random variable following some distribution. How can we do this? I will talk about one method that can be used for continuous distributions. First, we need some source of randomness; basically all computers can generate random bits (this is a whole other subject), so we can sample from the uniform distribution on $[0, 1]$. The objective is to turn this into sampling from some other distribution.

Let F_X be the cumulative distribution function for the random variable X that we would like to sample. Pick a random number $u \in [0, 1]$ under the uniform distribution, and let $x = F_X^{-1}(u)$. It turns out that x is a sample of X. Why is this? Well,

$$\begin{aligned} P(F_X^{-1}(u) < t) &= P(u < F_X(t)) \\ &= F_X(t) \end{aligned}$$

Where the last equality is because u is uniform on $[0, 1]$. Therefore the cumulative distribution function for $F_X^{-1}(u)$ is actually F_X, so it samples X.

2. The Method of Maximum Likelihood

2.1. **Note.** That there are notes on the course website which seem pretty clear to me. Check them out. Also, page 343 in your textbook.

2.2. **Idea.** The idea is that you have some parameter (μ, say) on which the distribution of some random variables depend. You are given these random variables, and you want to estimate μ. The right way to do this is to find the value of μ which maximizes the probability of any observed data. The distribution of the X_i, thought of as a function of μ, is called the likelihood function.

Making it more concrete will be helpful, I think:

2.3. **Example (Discrete).** The discrete case is the most clear. Suppose that you know that you have $X_1 \dots X_n$ iid Poisson variables, and you want to find the mean of the Poisson distribution. Well, since they are independent, their joint density is a product of densities, i.e. $L(\lambda) = \prod_i \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$. We just now have to maximize this as a function of λ. It is usually easier to maximize the log of it when you are dealing with a product:

$$\log(L(\lambda)) = \sum_i (-\lambda) + X_i \log(\lambda) - \log(X_i!)$$

Taking the first derivative yields

$$\frac{d}{d\lambda} \log(L(\lambda)) = \sum_i -1 + X_i \frac{1}{\lambda}$$

We want this to be zero, i.e. $n = \frac{1}{\lambda} \sum_i X_i$, or rather, $\lambda = \frac{1}{n} \sum_i X_i$. This is not a huge surprise, but it is nice reassurance.

2.4. **Another, Stranger Discrete Example.** For some reason this setup is more strange to me: this is kind of like a discrete version of the tilt problem on your homework. Suppose that you have n iid random variables from a distribution which takes one of three values (say -1, 0, 1) assigning probabilities $\frac{1}{3}(1-q)$, $\frac{1}{3}$, and $\frac{1}{3}(1+q)$ to them, respectively. What is the mle for q?

This is a multinomial trial: call the outcomes of the n trials by $X_1, \ldots, X_n$. Really what you want is to deal with sums of indicator functions, so let Z_{-1}, Z_0, and Z_1 be the number of the X_i which are -1, 0, and -1, respectively. The the likelihood function is:

$$L(q) = \frac{n!}{Z_{-1}!Z_0!Z_1!}\left(\frac{1}{3}(1-q)\right)^{Z_{-1}}\left(\frac{1}{3}\right)^{Z_0}\left(\frac{1}{3}(1+q)\right)^{Z_1}$$

To maximize this, we do the usual thing and take the log, giving

$$\log(L(q)) = \log n! - \sum_i \log Z_i + (Z_{-1} + Z_0 + Z_1)\log(1/3) + Z_{-1}\log(1-q) + Z_1\log(1+q)$$

This function has extreme points in the same places that L does, so we can take the derivative: $d/dq \log(L(q)) = \frac{Z_1}{1+q} - \frac{Z_{-1}}{1-q}$. Solving this for zero gives:

$$\begin{aligned} 0 &= \frac{Z_1}{1+q} - \frac{Z_{-1}}{1-q} \\ &= Z_1(1-q) - Z_{-1}(1+q) \\ &= \frac{Z_1}{Z_{-1}} - 1 - \left(\frac{Z_1}{Z_{-1}} + 1\right)q \end{aligned}$$

Therefore $q = \frac{\frac{Z_1}{Z_{-1}}-1}{\frac{Z_1}{Z_{-1}}+1}$. Technically (like, on your homework), we need to verify that this is actually a maximum. Note that the second derivative will always be negative, so it is in fact.

2.5. **Continuous Distributions.** These are the same, except the function that you optimize is the joint density functions. You solve them in the same way: (1) find the density function (perhaps it will be a joint density function if you have, as we did above, many trials (the X_i)) (2) think of it as a function of the parameter and (3) maximize it with respect to the parameter.

2.5.1. *Example.* I think you may have done this in class, but it never hurts to see it twice (the straightforward example I thought of (the exponential distribution) turned out to be on your homework).

Let's try to find the mle for the variance from n iid normal variables. Suppose that we know the mean μ. In practice, if you wanted to find μ and σ^2, you would first find the mle for μ because that won't depend on σ. Then you could use that to find σ. Anyway we know the mean is μ. Then the likelihood function is $L(\sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}}\exp(\frac{-1}{2\sigma^2}(X_i-\mu)^2)$. We would then take logs to get $\log L(\sigma^2) = \sum_i -(1/2)\log(2\pi\sigma^2) + \frac{-1}{2\sigma^2}(X_i-\mu)^2$. The derivative of this is $\sum_i \frac{-1}{2\sigma^2} + \frac{1}{2\sigma^4}(X_i-\mu)^2$, so we should set $\sigma^2 = (1/n)\sum_i(X_i-\mu)^2$, as we expected.

2.6. **Important Note So As Not To Lose Points.** If you want to maximize a function of one variable over $\mathbb{R}$ (or maybe $(0,\infty)$), you need to make sure that it doesn't take its maximum at zero, or as it tends to zero, or as it tends to ∞. Anyway just make sure you take care of all the cases.

3. Estimators

See page 344 in the textbook. The idea here is that you actually want the value of some function g of a parameter in your distribution (to use the example on your homework, you want $g(\lambda) = e^{-\lambda}$, where λ is the parameter to a sort-of-Poisson distribution). You're going to use an estimator function of random variables from that distribution to estimate the value of the function g (in relation to the above, that was figuring out what a good estimator was for estimating the value of the identity function of the parameter).

3.1. **Definitions.** Of course, you want to be able to measure how good a job your estimator is doing. For that you would want the **mean squared error**. If $T(X)$ is your estimator for $g(\Theta)$, where Θ is the unknown parameter, then this is defined to be $E[(T(X) - g(\Theta))^2]$. It makes sense to take the expected value of this because it is a function of the random variable X (note that X may actually be a vector $(X_i)_i$ if you have multiple observations).

Another quantity that has a name is $E[T(X) - g(\Theta)]$. This is called the **bias**. If this is zero no matter what Θ is, we call the estimator **unbaised**.
Important mildy counterintuitive fact: it is not always true that the "best" estimator (i.e. the smallest MSE) is unbiased!

3.2. **Example 1.** It is always true that the sample mean $\bar{X}_i$ is an unbiased estimator of the true mean. This is because

$$\begin{aligned} E[(1/n)\sum_i X_i - \mu] &= (1/n)\sum_i E[X_i] - \mu \\ &= \mu - \mu = 0 \end{aligned}$$

However, the MSE of this estimator can change based on the distribution. For example, in the Poisson problem above,

$$\begin{aligned} E[((1/n)\sum_i X_i - \mu)^2] &= E[\frac{1}{n^2}\left(\sum_i X_i\right)^2 - 2(1/n)\mu\sum_i E[X_i] + \mu^2 \\ &= E[\left((1/n)\sum_i X_i\right)^2] - \mu^2 \\ &= E[Y^2] - E[Y]^2 \\ &= \mathrm{Var}(Y) \\ &= \frac{\lambda}{n} \end{aligned}$$

Where $Y = (1/n)\sum_i X_i$. Note that the variance of a sum (of ind vars) is the sum of the variances, and multiplying by a constant multiplies the variance by the square, so we are left with a $1/n$. This is good, since clearly the MSE of the sample mean should decrease with n. Also note that the MSE changes as the true value of λ changes.

3.3. **Example 2.** This is from the notes, but I think it is interesting. Suppose that we have n iid variables from the uniform distribution on $[0, \Theta]$. It turns out that the mle for Θ is $\max_i X_i$. To see this, note the likelihood function is $L(\Theta) = (1/\Theta)^n$ if $\Theta > \max_i X_i$, and 0 otherwise. Since maximizing the likelihood mean minimizing Θ, we make it as small as possible.

Note that this estimator is **not** unbiased, because it always guesses high!

It turns out you can make it unbiased by multiplying the max by $n/(n+1)$, but let's take all estimators of the form $c \max_i X_i$ and try to minimize the MSE.

The relevant expected value is $E[c^2(\max_i X_i)^2 - 2c\Theta \max_i Xi + \Theta^2] = c^2 E[(\max_i X_i)^2] - 2c\Theta E[\max_i X_i] + \Theta^2$. Finding the first quantity requires finding the density of $\max_i X_i$. This is $d/dt P(\max_i X_i < t) = d/dt(t/\Theta)^n = (n/\Theta)(t/\Theta)^{n-1}$. This only makes sense up to Θ, of course. It's interesting that it works out so nicely.

$$\begin{aligned} E[(\max_i X_i)^2] &= \int_0^\Theta t^2 \frac{nt^{n-1}}{\Theta^n} dt \\ &= \frac{n}{\Theta^n}\int_0^\Theta t^{n+1} dt \\ &= \frac{n}{\Theta^n}\frac{\theta^{n+2}}{n+2} \\ &= \frac{n}{n+2}\Theta^2 \end{aligned}$$

Similarly, $E[\max_i X_i] = \frac{n}{n+1}\Theta$. Therefore the mse of the estimator $c\max_i X_i$ is

$$c^2\frac{n}{n+2}\Theta^2 - 2c\Theta\frac{n}{n+1}\Theta + \Theta^2$$
$$= \left(c^2\frac{n}{n+2} - 2c\frac{n}{n+1} + 1\right)\Theta^2$$

Therefore, we just minimize this with respect to c to get the estimator with the smallest mse. A simple calculation yields $\frac{n+2}{n+1}$. Note that we can't get rid of the multiplication by Θ, so it's ok to minimize the rest.