Lisa Yan and Jerry Cain
CS 109

Section #7
November 3-4, 2020

Section #7 Solutions

Based on the work of many CS109 staffs

1 Warmups

1.1 Parameters and MLE

Suppose $x_1, \ldots, x_n$ are i.i.d. (independent and identically distributed) values sampled from some distribution with density function $f(x|\theta)$, where θ is unknown. Recall that the likelihood of the data is

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

Recall we solve an optimization problem to find $\hat{\theta}$ which maximizes $L(\theta)$, i.e., $\hat{\theta} = \arg\max_\theta L(\theta)$.

1. Write an expression for the log-likelihood, $LL(\theta) = \log L(\theta)$.
2. Why can we optimize $LL(\theta)$ rather than $L(\theta)$?
3. Why do we optimize $LL(\theta)$ rather than $L(\theta)$?

1. $LL(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$
2. The logarithm (for bases > 1) is a monotonically increasing function. This means that if $f(a) > f(b)$, then $\log(f(a)) > \log(f(b))$, so the arg max function is preserved across a logarithmic transformation, i.e., $\arg\max L(\theta) = \arg\max LL(\theta)$.
3. Logs turn products into sums, which makes taking the derivative for maximization or minimization much simpler.

1.2 Beta

1. Suppose you have a coin where you have no prior belief on its true probability of heads p. How can you model this belief as a Beta distribution?
2. Suppose you have a coin which you believe is fair, with "strength" α. That is, pretend you've seen α heads and α tails. How can you model this belief as a Beta distribution?
3. Now suppose you take the coin from the previous part and flip it 10 times. You see 8 heads and 2 tails. How can you model your posterior belief of the coin's probability of heads?

1. Beta(1, 1) is a uniform prior, meaning that prior to seeing the experiment, all probabilities of heads are equally likely.
2. Beta($\alpha + 1, \alpha + 1$). This is our prior belief about the distribution.
3. Beta($\alpha + 9, \alpha + 3$)

1.3 Maximum A Posteriori

1. Intuitively, what is MAP? What problem is it trying to solve? How does it differ from MLE?
2. Given a 6-sided die (possibly unfair), you roll the die N times and observe the counts for each of the 6 outcomes as $n_1, ..., n_6$. What is the maximum a posteriori estimate of this distribution, using Laplace smoothing? Recall that the die rolls themselves follow a multinomial distribution.

1. From the course notes: The paradigm of MAP is that we should choose the value for our parameters that is **the most likely given the data**. At first blush this might seem the same as MLE; however, remember that MLE chooses the value of parameters that **makes the data most likely**. One of the disadvantages of MLE is that it best explains data we have seen and makes no attempt to generalize to unseen data. In MAP, we incorporate prior belief about our parameters, and then we update our posterior belief of the parameters based on the data we have seen.
2. Using a prior which represents one imagined observation of each outcome is called Laplace smoothing and it guarantees that none of your probabilities are 0 or 1. The Laplace estimate for a Multinomial RV is $p_i = \frac{n_i+1}{N+6}$ for $i = 1, ..., 6$.

2 The Honor Code

We have decided that automated tools should be used to identify if two submissions are suspiciously similar. (N.B. these tools are never used as a basis for community standards cases — those always require professional human opinion.) However, automated tools are never perfect. As active CS109 students, we would like to explore the chance of a false positive in automated tools. For this task, we will consider Breakout, a CS106A assignment where students implement Breakout.

This problem combines our knowledge of Beta, the Central Limit Theorem, and Maximum Likelihood Estimation. Exciting!

2.1 Beta Sum

What is the distribution of the sum of 100 i.i.d. Betas? Let X be the sum

$$X = \sum_{i=1}^{100} X_i \qquad \text{where each } X_i \sim \text{Beta}(a = 3, b = 4)$$

Note the variance of a Beta:

$$\text{Var}(X_i) = \frac{ab}{(a+b)^2(a+b+1)} \qquad \text{where } X_i \sim \text{Beta}(a, b)$$

By the Central Limit Theorem, the sum of equally weighted IID random variables will be Normally distributed. We calculate the expectation and variance of X_i using the beta formulas:

$$\begin{aligned} E(X_i) &= \frac{a}{a+b} && \text{Expectation of a Beta} \\ &= \frac{3}{7} \approx 0.43 \\ \text{Var}(X_i) &= \frac{ab}{(a+b)^2(a+b+1)} && \text{Variance of a Beta} \\ &= \frac{3 \cdot 4}{(3+4)^2(3+4+1)} \\ &= \frac{12}{49 \cdot 8} \approx 0.03 \end{aligned}$$

$$\begin{aligned} X &\sim N(\mu = n \cdot E[X_i], \sigma^2 = n \cdot \text{Var}(X_i)) \\ &\sim N(\mu = 43, \sigma^2 = 3) \end{aligned}$$

2.2 Single Match

Say there are 1000 decision points when writing Breakout. Assume: Each decision point is binary. Each student makes all 1000 decisions. For each decision there is a probability p that a student takes the more popular choice. What is the probability distribution for the number of matching decisions (we are going to refer to this as the "score")? If possible, could you approximate this probability?

Let A_i be the event that decision point i is matched. We note that a match occurs when both students make the more popular choice or when both students make the less popular choice. $P(A_i) = P(\text{Both more popular}) + P(\text{Both less popular}) = p^2 + (1-p)^2$.

Let M be a random variable for the number of matches. It is easy to see that each of the 1000 decisions is an independent Bernoulli experiment with probability of success $p' = p^2 + (1-p)^2$.

Therefore $M \sim Bin(1000, p')$.

We can use a Normal distribution to approximate a binomial. We approximate $M \sim Bin(1000, p')$ with Normal random variable $Y \sim N(1000p', 1000p'(1-p'))$.

2.3 Maximum Match

When we look at two assignments, the probability of a false match is exceedingly small. What would the max similarity score look like when we compare one student to 5000 historical breakout submissions? Let X_i be the similarity score between a student who worked on their own and student i. Let Y be the highest match score between the student and all other submissions:

$$Y = \max_i X_i$$

The Central Limit Theorem tells us about the distribution of the sum of IID random variables. A more obscure theorem, the Fisher-Tippett-Gnedenko theorem, tells us about the *max* of IID random variables. It says that the max of IID exponential or normal random variables will be a "Gumbel" random variable.

$$Y \sim \text{Gumbel}(\mu, \beta) \qquad \text{The max of IID vars}$$

$$f(Y = k) = \frac{1}{\beta} e^{-(z+e^{-z})} \text{ where } z = \frac{k-\mu}{\beta} \qquad \text{The Gumbel PDF}$$

You are given data of 1300 students' max scores from three quarters (we believe they all worked independently): $y^{(1)} \ldots y^{(1300)}$. Set up (but do not solve) simultaneous equations we could solve to find the values of μ and β.

For this problem, we use Maximum Likelihood Estimator (MLE) to estimate the parameters $\theta = (\mu, \beta)$.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(Y^{(i)} = y^{(i)} \mid \theta) \\
LL(\theta) &= \log \prod_{i=1}^{n} f(Y^{(i)} = y^{(i)} \mid \theta) \\
&= \sum_{i=1}^{n} \log f(Y^{(i)} = y^{(i)} \mid \theta) \\
&= \sum_{i=1}^{n} \log \frac{1}{\beta} e^{-(z_i + e^{-z_i})} && \text{where } z_i = \frac{y^{(i)} - \mu}{\beta} \\
&= \sum_{i=1}^{n} \log \frac{1}{\beta} + \sum_{i=1}^{n} -(z_i + e^{-z_i}) \\
&= -n \log(\beta) + \sum_{i=1}^{n} -(z_i + e^{-z_i})
\end{aligned}
$$

Now we must choose the values of $\theta = (\mu, \beta)$ that maximize our log-likelihood function. First, we need to find the first derivative of the log-likelihood function with respect to our parameters.

$$
\begin{aligned}
\frac{\partial LL(\theta)}{\partial \mu} &= \frac{\partial}{\partial \mu} \Big[-n \log(\beta) + \sum_{i=1}^{n} -(z_i + e^{-z_i}) \Big] \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \mu} \Big[-(z_i + e^{-z_i}) \Big] \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial z_i} \Big[-(z_i + e^{-z_i}) \Big] \frac{\partial z_i}{\partial \mu} && \text{By the Chain Rule} \\
&= \sum_{i=1}^{n} \Big[-1 + e^{-z_i} \Big] \Big[-\frac{1}{\beta} \Big] \\
&= \frac{1}{\beta} \sum_{i=1}^{n} 1 - e^{-z_i}
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial LL(\theta)}{\partial \beta} &= \frac{\partial}{\partial \beta}\Big[- n \log(\beta) + \sum_{i=1}^{n} -(z_i + e^{-z_i})\Big] \\
&= -\frac{n}{\beta} + \sum_{i=1}^{n} \frac{\partial}{\partial \beta}\Big[- (z_i + e^{-z_i})\Big] \\
&= -\frac{n}{\beta} + \sum_{i=1}^{n} \frac{\partial}{\partial z_i}\Big[- (z_i + e^{-z_i})\Big]\frac{\partial z_i}{\partial \beta} && \text{By the Chain Rule} \\
&= -\frac{n}{\beta} + \sum_{i=1}^{n} \Big[- 1 + e^{-z_i}\Big]\Big[\frac{\mu - y^{(i)}}{\beta^2}\Big] && \text{Where the last term equals } \frac{\partial z_i}{\partial \beta}
\end{aligned}$$

We want to find a simultaneous solution for both, but this is algebraically not possible. We will instead use an approximate method (gradient ascent) to solve for these, which will be taught next week.