CSE 312: Foundations of Computing II

Section 9: Concentration Inequalities and Maximum Likelihood Solutions

0. Concentration Inequalities

Suppose $X \sim \text{Binomial}(6, 0.4)$. We will bound $\Pr(X \geq 4)$ using the tail bounds we've learned, and compare this to the true result.

(a) Give an upper bound for this probability using Markov's inequality. Why can we use Markov's inequality?

Solution:

$\Pr(X \geq 4) \leq \frac{\mathbb{E}[X]}{4} = \frac{2.4}{4} = 0.6$. We can use it since X is nonnegative.

(b) Give an upper bound for this probability using Chebyshev's inequality. You may have to rearrange algebraically and it may result in a weaker bound.

Solution:

$\Pr(X \geq 4) = \Pr(X - 2.4 \geq 1.6) \leq \Pr(\ |X - 2.4| \geq 1.6) \leq \frac{\text{Var}(X)}{1.6^2} = \frac{1.44}{1.6^2} = 0.5625$

(c) Give an upper bound for this probability using the Chernoff bound.

Solution:

$\Pr(X \geq 4) = \Pr\big(X \geq (1 + \frac{2}{3})2.4\big) \leq e^{-(\frac{2}{3})^2 \mathbb{E}[X]/3} = e^{-4 \times 2.4/27} \approx 0.7$

(d) Give the exact probability.

Solution:

$\Pr(X \geq 4) = \Pr(X = 4) + \Pr(X = 5) + \Pr(X = 6) = \binom{6}{4}(0.4)^4(0.6)^2 + \binom{6}{5}(0.4)^5(0.6) + \binom{6}{6}0.4^6 \approx 0.1792$

1. Laplace MLE

Suppose $x_1, \ldots, x_{2n}$ are iid realizations from the Laplace density (double exponential density): for $x \in \mathbb{R}$,

$$f_X(x \mid \theta) = \frac{1}{2} e^{-|x-\theta|}$$

Find the MLE for θ. For this problem, you need not verify that the MLE is indeed a maximizer. You may find the **sign** function useful:

$$\text{sgn}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

Solution:

$$\begin{aligned} L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \prod_{i=1}^{2n} \frac{1}{2} e^{-|x_i - \theta|} \\ \ln L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \sum_{i=1}^{2n} \left[-\ln 2 - |x_i - \theta|\right] \\ \frac{\partial}{\partial \theta} \ln L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \sum_{i=1}^{2n} \operatorname{sgn}\left(x_i - \theta\right) = 0 \\ \hat{\theta} &= \text{any value in } [x_n^{'}, x_{n+1}^{'}] \end{aligned}$$

where $x_i^{'}$ is the i^{th} order statistic: the i^{th} smallest observation.
If you wanted to argue that this is a global maximizer, note that the log likelihood is the sum of concave functions, so every critical point is a global maximizer.

2. MAP Estimation

Let $x_1, ..., x_n$ be iid realizations from a distribution with common pmf $p_X(x; \theta)$ where θ is an unknown but **fixed** parameter. Let's call the event $\{X_1 = x_1, ..., X_n = x_n\} = \mathcal{D}$ for data. You may wonder why in MLE, we seek to maximize the likelihood $L(\mathcal{D} \mid \theta)$, rather than $\Pr(\theta \mid \mathcal{D})$. This is because it doesn't make sense to compute $\Pr(\theta)$, since θ is fixed. However, in **Maximum a Posteriori (MAP) estimation**, we assume the parameter is a random variable (denoted Θ), and attempt to maximize $\pi_\Theta(\theta \mid \mathcal{D})$, where π_Θ is the pmf or pdf of Θ, depending on whether Θ is continuous or discrete. Using Bayes Theorem, we get $\pi_\Theta(\theta \mid \mathcal{D}) = \frac{L(\mathcal{D}|\theta)\pi_\Theta(\theta)}{L(\mathcal{D})}$. To maximize the LHS with respect to θ, we may ignore the denominator on the RHS since it is constant with respect to θ. Hence MAP seeks to maximize $\pi_\Theta(\theta \mid \mathcal{D}) \propto L(\mathcal{D} \mid \theta)\pi_\Theta(\theta)$. We call $\pi_\Theta(\theta)$ the **prior** distribution on the parameter Θ, and $\pi_\Theta(\theta \mid \mathcal{D})$ the **posterior** distribution on Θ. MLE maximizes the likelihood, and MAP maximizes the product of the likelihood and the prior. If the prior is uniform, we will see that MAP is the same as MLE (since $\pi_\Theta(\theta)$ won't depend on θ).

(a) Suppose we have the samples $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0$ from the Bernoulli(θ) distribution, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0, 1)$. What is $\hat{\theta}_{MLE}$?

Solution:

$$\begin{aligned} L\left(x_1, \ldots, x_5 \mid \theta\right) &= \theta^2(1-\theta)^3 \\ \ln L\left(x_1, \ldots, x_5 \mid \theta\right) &= 2\ln(\theta) + 3\ln(1-\theta) \\ \frac{\partial}{\partial \theta} \ln L\left(x_1, \ldots, x_5 \mid \theta\right) &= \frac{2}{\theta} - \frac{3}{1-\theta} = 0 \\ & \quad 2 - 2\theta = 3\theta \\ \hat{\theta}_{MLE} &= \boxed{\frac{2}{5}} \end{aligned}$$

(b) Suppose we impose that $\theta \in \{0.2, 0.5, 0.7\}$. What is $\hat{\theta}_{MLE}$?

Solution:

We can compute $L(\mathcal{D} \mid \theta)$ for each value of θ, and take the largest.
$L(\mathcal{D} \mid 0.2) = (1 - 0.2)^3(0.2)^2 = 0.02048$
$L(\mathcal{D} \mid 0.5) = (1 - 0.5)^3(0.5)^2 = 0.03125$

$L(\mathcal{D} \mid 0.7) = (1-0.7)^3(0.7)^2 = 0.01323$

So $\hat{\theta}_{MLE} = \boxed{0.5}$.

(c) Assume Θ is restricted as in part (b) (now a random variable for MAP). Assume a (discrete) prior of $\pi_\Theta(0.2) = 0.1, \pi_\Theta(0.5) = 0.01, \pi_\Theta(0.7) = 0.89$. What is $\hat{\theta}_{MAP}$?

Solution:

We compute the objective to maximize for MAP:

$\pi_\Theta(0.2 \mid \mathcal{D}) \propto L(\mathcal{D} \mid 0.2)\pi_\Theta(0.2) = 0.02048 \cdot 0.1 = 0.002048$

$\pi_\Theta(0.5 \mid \mathcal{D}) \propto L(\mathcal{D} \mid 0.5)\pi_\Theta(0.5) = 0.03125 \cdot 0.01 = 0.0003125$

$\pi_\Theta(0.7 \mid \mathcal{D}) \propto L(\mathcal{D} \mid 0.7)\pi_\Theta(0.7) = 0.01323 \cdot 0.89 = 0.0117747$

Hence $\hat{\theta}_{MAP} = \boxed{0.7}$.

(d) Show that we can make the MAP estimator whatever we want it to be. That is, for each of the three candidate parameters above, find a prior distribution on Θ such that the MAP estimate is that parameter.

Solution:

Just assign a prior of 1 to the desired parameter. If you don't want something degenerate, assign a prior extremely close to 1, and give uniform probability to the other parameters.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\theta \in (0,1)$ (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign θ the **Beta distribution** with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$ and 0 otherwise as a prior, where c is a normalizing constant which has a complicated form. The **mode** of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is given as $\frac{\alpha-1}{\alpha+\beta-2}$ (the mode is the value with the highest density $= \arg\max_{w \in (0,1)} f_W(w)$). Suppose $x_1, ..., x_n$ are iid samples from the Bernoulli distribution with unknown parameter, where $\sum_{i=1}^n x_i = k$. Recall that the MLE is k/n. Show that the posterior $\pi_\Theta(\theta \mid \mathcal{D})$ has a $\text{Beta}(k+\alpha, n-k+\beta)$ density, and find the MAP estimator for Θ. (Hint: use the mode given). Notice that $\text{Beta}(1,1) \equiv \text{Uniform}(0,1)$. If we had this prior, how would the MLE and MAP estimates compare?

Solution:

We want to maximize $\pi_\Theta(\theta \mid \mathcal{D}) \propto L(\mathcal{D} \mid \theta)\pi_\Theta(\theta) \propto \left(\theta^k(1-\theta)^{n-k}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) = \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$. Hence the posterior $\sim \text{Beta}(k+\alpha, n-k+\beta)$. We are given the mode of any beta distribution, so our estimate is $\hat{\theta}_{MAP} = \boxed{\dfrac{k+\alpha-1}{n+\alpha+\beta-2}}$. If $\alpha = \beta = 1$, then this is exactly the MLE, and $\text{Beta}(1,1) \equiv \text{Uniform}(0,1)$, so having a uniform prior causes the MLE to equal the MAP estimate.

(f) Since the posterior is also a Beta distribution, we call the Beta distribution the **conjugate prior** to the Bernoulli/Binomial distribution. Intepret what the parameters α, β mean as to the prior.

Solution:

$\alpha - 1$ is the number of heads you pretend to see beforehand, and $\beta - 1$ is the number of tails you pretend to see beforehand. Why is this? Because our MLE was $\frac{k}{n}$ (heads/tails), and the MAP estimate is $\frac{k+\alpha-1}{n+(\alpha+\beta-2)} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$. Hence we add $\alpha+\beta-2$ "fake" trials, $\alpha-1$ which are heads (numerator), and the other $\beta-1$ which are tails. This should look familiar as our estimates for $\Pr(word \mid spam)$ and $\Pr(word \mid ham)$ with a $\text{Beta}(2,2)$ prior when we did smoothing for Naive Bayes.

(g) Which do you think is "better", MLE or MAP?

Solution:

There is no right answer. There are two main schools in statistics: Bayesians and Frequentists. Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a fixed quantity. On the other hand, Bayesians prefer MAP, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a random variable, and we seek the mode - the value with the highest probability or density. An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not? If so, what distribution should we put on the parameter space?

Anyway, in the long run, the prior "washes out", and the only thing that matters is the likelihood; the observed data. For small sample sizes like this, the prior significantly influences the MAP estimate. However, as the number of samples goes to infinity, the MAP and MLE are equal.