CSE 312: Foundations of Computing II
Quiz Section #9: Maximum Likelihood Estimation, Chernoff Bound (solutions)

Review: Main Theorems and Concepts

Weak Law of Large Numbers (WLLN): Let $X_1, \ldots, X_n$ be iid random variables with common mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean for a sample of size n. Then, for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0$. We say that $\bar{X}_n$ converges in probability to μ.

Chernoff Bound: Let $X_1, ..., X_n$ be independent random variables with $X_i \sim Ber(p_i)$, and let $X = \sum_{i=1}^{n} X_n$ have $\mathbb{E}[X] = \mu$. Then, for any $\delta > 0$, we have the following

- $\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(\frac{-\delta^2\mu}{2+\delta}\right)$
- $\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$
- $\mathbb{P}(|X-\mu| > \delta\mu) \leq 2\exp\left(\frac{-\delta^2\mu}{2+\delta}\right)$

Sample/Realization: A sample (or realization) x of a random variable X is the value that is actually observed.

Likelihood: Let $x_1, \ldots x_n$ be iid samples from probability mass function $p_X(\mathrm{x} \mid \theta)$ (if X discrete) or density $f_X(\mathrm{x} \mid \theta)$ (if X continuous), where θ is a parameter (or a vector of parameters). We define the likelihood function to be the probability of seeing the data.

If X is discrete:

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

If X is continuous:

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

Maximum Likelihood Estimator (MLE): We denote the MLE of θ as $\hat{\theta}_{\text{MLE}}$ or simply $\hat{\theta}$, the parameter (or vector of parameters) that maximizes the likelihood function (probability of seeing the data).

$$\hat{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\theta} L(x_1, \ldots, x_n \mid \theta) = \operatorname*{argmax}_{\theta} \ln L(x_1, \ldots, x_n \mid \theta)$$

Log-Likelihood: We define the log-likelihood as the natural logarithm of the likelihood function. Since the logarithm is a strictly increasing function, the value of θ that maximizes the likelihood will be exactly the same as the value that maximizes the log-likelihood.

If X is discrete:

$$\ln L(x_1, \ldots, x_n \mid \theta) = \sum_{i=1}^{n} \ln p_X(x_i \mid \theta)$$

If X is continuous:

$$\ln L(x_1, \ldots, x_n \mid \theta) = \sum_{i=1}^{n} \ln f_X(x_i \mid \theta)$$

Steps to find the maximum likelihood estimator, $\hat{\theta}$:

1. Find the likelihood and log-likelihood of the data.
2. Take the derivative of the log-likelihood and set it to 0 to find a candidate for the MLE, $\hat{\theta}$.
3. Take the second derivative and show that $\hat{\theta}$ indeed is a maximizer, that $\frac{d^2L}{d\theta^2} < 0$ at $\hat{\theta}$. Also ensure that it is the global maximizer: check points of non-differentiability and boundary values.

Exercises

1. Suppose $x_1, \ldots, x_n$ are iid samples from a distribution with density

$$f_X(x \mid \theta) = \begin{cases} \dfrac{\theta x^{\theta-1}}{3^\theta}, & 0 \le x \le 3 \\ 0, & \text{otherwise} \end{cases}$$

Find the MLE for θ.

$$\begin{aligned} L(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} \frac{\theta x_i^{\theta-1}}{3^\theta} \\ \ln L(x_1, \ldots, x_n \mid \theta) &= \sum_{i=1}^{n} (\ln\theta + (\theta-1)\ln x_i - \theta \ln 3) \\ \frac{\partial}{\partial\theta} \ln L(x_1, \ldots, x_n \mid \theta) &= \sum_{i=1}^{n} \left(\frac{1}{\theta} + \ln x_i - \ln 3\right) = 0 \\ \frac{n}{\hat{\theta}} + \sum_{i=1}^{n} \ln x_i - n \ln 3 &= 0 \\ \frac{n}{\hat{\theta}} &= n\ln 3 - \sum_{i=1}^{n} \ln x_i \\ \hat{\theta}_{\text{MLE}} &= \frac{n}{n\ln 3 - \sum_{i=1}^{n} \ln x_i} \end{aligned}$$

Check that it is a maximum by showing the second derivative is negative for all values of θ.

$$\frac{\partial^2}{\partial\theta^2} \ln L(x_1, \ldots, x_n \mid \theta) = \sum_{i=1}^{n} \left(-\frac{1}{\theta^2}\right) = -\frac{n}{\theta^2} < 0$$

so $\ln L(x_1, \ldots, x_n \mid \theta)$ is concave downward everywhere.

2. Suppose $x_1, \ldots, x_{2n}$ are iid samples from the Laplace density (double exponential density)

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

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:

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

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

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.

3. Suppose $X \sim Bin(6, 0.4)$. We will bound $\mathbb{P}(X \geq 4)$ using the concentration inequalities we've learned, and compare this to the true result.

   (a) Give an upper bound for this probability using Markov's inequality.

   $\mathbb{P}(X \geq 4) \leq \frac{\mathbb{E}[X]}{4} = \frac{2.4}{4} = 0.6$

   (b) Give an upper bound for this probability using Chebyshev's inequality.

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

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

   $\mathbb{P}(X \geq 4) = \mathbb{P}\left(X \geq (1 + \frac{2}{3})2.4\right) \leq e^{-\delta^2 \mu/(2+\delta)} = e^{-0.4} \approx 0.6703$

   (d) Give the exact probability.

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

4. (**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 $\mathbb{P}(\theta \mid \mathcal{D})$. This is because it doesn't make sense to compute $\mathbb{P}(\theta)$, 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 $\pi_\Theta(\theta \mid \mathcal{D}) = \dfrac{L(\mathcal{D} \mid \theta)\pi_\Theta(\theta)}{L(\mathcal{D})}$. 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 $Ber(\theta)$ distribution, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0, 1)$. What is $\hat{\theta}_{MLE}$?

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

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

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 $\hat{\theta}_{MAP}$?

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.

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 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 $W \sim Beta(\alpha, \beta)$ 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 $Beta(k + \alpha, n - k + \beta)$ density, and find the MAP

estimator for Θ. (Hint: use the mode given). Notice that $Beta(1,1) \equiv Unif(0,1)$. If we had this prior, how would the MLE and MAP estimates compare?

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 $\sim Beta(k+\alpha, n-k+\beta)$. 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 $Beta(1,1) \equiv Unif(0,1)$, 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 distribution. Intepret what the parameters α, β mean as to the prior.

$\alpha - 1$ 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 $\dfrac{k}{n}$ (heads/tails), and the MAP estimate is $\dfrac{k+\alpha-1}{n+(\alpha+\beta-2)} = \dfrac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$. 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 $\mathbb{P}(word \mid spam)$ and $\mathbb{P}(word \mid ham)$ with a $Beta(2,2)$ prior when we did smoothing for Naive Bayes.

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

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.