CSE 312: Foundations of Computing II

Section 8: Maximum Likelihood and more

1. Review of Main Concepts

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

(b) **Likelihood**: Let $x_1, \ldots x_n$ be iid realizations from probability mass function $p_X(\mathrm{x}\,;\theta)$ (if X discrete) or density $f_X(\mathrm{x}\,;\theta)$ (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:

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

If X is continuous:

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

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

$$\hat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta} L\left(x_1, \ldots, x_n \mid \theta\right) = \arg\max_{\theta} \ln L\left(x_1, \ldots, x_n \mid \theta\right)$$

(d) **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:

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

If X is continuous:

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

(e) **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter θ is defined as $\text{Bias}\left(\hat{\theta}, \theta\right) = \mathbb{E}[\hat{\theta}] - \theta$. An estimator $\hat{\theta}$ of θ is unbiased iff $\text{Bias}\left(\hat{\theta}, \theta\right) = 0$, or equivalently $\mathbb{E}[\hat{\theta}] = \theta$.

(f) **Steps to find the maximum likelihood estimator, $\hat{\theta}$:**

(a) Find the likelihood and log-likelihood of the data.

(b) Take the derivative of the log-likelihood and set it to 0 to find a candidate for the MLE, $\hat{\theta}$.

(c) Take the second derivative and show that $\hat{\theta}$ indeed is a maximizer, that $\frac{\partial^2 L}{\partial \theta^2} < 0$ at $\hat{\theta}$. Also ensure that it is the global maximizer: check points of non-differentiability and boundary values.

(g) **Markov's Inequality**: Let X be a non-negative random variable, and $\alpha > 0$. Then, $\mathbb{P}\left(X \geq \alpha\right) \leq \frac{\mathbb{E}[X]}{\alpha}$.

(h) **Chebyshev's Inequality** (we did not cover this in class): Suppose Y is a random variable with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$. Then, for any $\alpha > 0$, $\mathbb{P}\left(|Y - \mu| \geq \alpha\right) \leq \frac{\sigma^2}{\alpha^2}$.

(i) **Chernoff Bound (for the Binomial)**: (We will not cover this in class, but it's good to know.) It's stronger than the Chebyshev bound. Suppose $X \sim \text{Binomial}(n, p)$ and $\mu = np$. Then, for any $0 < \delta < 1$,

- $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}}$
- $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}}$

2. 312 Grades

Suppose Professor Karlin loses everyones grades for 312 and decides to make it up by assigning grades randomly according to the following probability distribution, and hoping the n students wont notice: give an A with probability 0.5, a B with probability θ, a C with probability 2θ, and an F with probability $0.5 - 3\theta$. Each student is assigned a grade independently. Let x_A be the number of people who received an A, x_B the number of people who received a B, etc, where $x_A + x_B + x_C + x_F = n$. Find the MLE for θ.

3. A Red Poisson

Suppose that $x_1, \ldots, x_n$ are i.i.d. samples from a Poisson(θ) random variable, where θ is unknown. Find the MLE of θ.

4. Independent Shreds, You Say?

(Covered in class.) You are given 100 independent samples $x_1, x_2, \ldots, x_{100}$ from Bernoulli(θ), where θ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. You would like to estimate the distribution's parameter θ. Give all answers to 3 significant digits.

(a) What is the maximum likelihood estimator $\hat{\theta}$ of θ?

(b) Is $\hat{\theta}$ an unbiased estimator of θ?

5. Y Me?

Let $y_1, y_2, ...y_n$ be i.i.d. samples of a random variable with density function

$$f_Y(y|\theta) = \frac{1}{2\theta} \exp(-\frac{|y|}{\theta})$$

.

Find the MLE for θ in terms of $|y_i|$ and n.

6. 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:

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

7. What if we lose ?

[This is practice with earlier material] Suppose 59 percent of voters favor Proposition 600. Use the Normal approximation to estimate the probability that a random sample of 100 voters will contain:

(a) at most 50 in favor. Mention any assumption that you make.

(b) more than 100 voters in favor or fewer than 0 voters in favor (again based on this normal approximation). Will the probability be non zero?

8. Law of Total Probability Review

(a) (Discrete version) Suppose we flip a coin with probability U of heads, where U is equally likely to be one of $\Omega_U = \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$ (notice this set has size $n+1$). Let H be the event that the coin comes up heads. What is $\mathbb{P}(H)$?

(b) (Continuous version) Now suppose $U \sim$ Uniform(0,1) has the *continuous* uniform distribution over the interval $[0, 1]$. What is $\mathbb{P}(H)$?

(c) Let's generalize the previous result we just used. Suppose E is an event, and X is a continuous random variable with density function $f_X(x)$. Write an expression for $\mathbb{P}(E)$, conditioning on X.

9. MAP Estimation*

(Optional: depending on if we have covered this in lecture; Read sections 7.4 and 7.5, if you're interested) 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 $\pi_\Theta(\theta \mid \mathcal{D}) = \frac{L(\mathcal{D}|\theta)\pi_\Theta(\theta)}{L(\mathcal{D})}$. 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 $\hat{\theta}_{MLE}$?

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

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

(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.

(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 $W \sim$ Beta(α, β) 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$(1,1) \equiv$ Uniform$(0,1)$. If we had this prior, how would the MLE and MAP estimates compare?

(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.

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