CSE 312: Foundations of Computing II

Section 8: Maximum Likelihood, Markov Chains 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 $p_X(x ; \theta)$ (if X discrete) or density $f_X(x ; \theta)$ (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:

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

If X is continuous:

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

(c) **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).

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

(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:

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

If X is continuous:

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

(e) **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter θ is defined as Bias $\left(\hat{\theta}, \theta\right) = \mathbb{E}[\hat{\theta}] - \theta$. An estimator $\hat{\theta}$ of θ is unbiased iff 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) A **discrete-time stochastic process (DTSP)** is a sequence of random variables $X_0, X_1, X_2, \ldots$, where X_t is the value at time t. For example, the temperature in Seattle or stock price of TESLA each day, or which node you are at after each time step on a random walk on a graph.

(h) A **Markov Chain** is a DTSP, with the additional following three properties:

I. ...has a finite (or countably infinite) **state space** $\mathcal{S} = \{s_1, \ldots, s_n\}$ which it bounces between, so each $X_t \in \mathcal{S}$.

II. ...satisfies the **Markov property**. A DTSP satisfies the Markov property if the future is (conditionally) independent of the past given the present. Mathematically, it means, $P(X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x_t) = P(X_{t+1} = x_{t+1} | X_t = x_t)$.

III. ...is defined by a **transition probability matrix (TPM)** P whose entries are **transition probabilities**. P is a square $n \times n$ matrix, where $P_{ij} = P(X_{t+1} = s_j | X_t = s_i)$ is the probability of transitioning from state s_i to state s_j. Note that these values are independent of the time t.

(i) (To be covered in class 11/29) The **stationary distribution** π of a Markov chain is an n-dimensional row vector $\pi = [\pi_1 \ \ldots, \pi_n]$ such that $\pi P = \pi$ and $\sum_{i=1}^{n} \pi_i = 1$. The interpretation of π is the following: If at some time t, the probability of being in state i is π_i for all i, then the same holds true at time $t+1$ (and at every step thereafter).

(j) (To be covered in class 11/29) The **fundamental theorem of Markov chains** says that $\lim_{t\to\infty} P_{ij}^t = \pi_j$ for all i, j, where P_{ij}^t is the probability that, starting from state i, the chain is in state j after t steps.

(k) (To be covered in class 12/1) **Markov's Inequality**: Let X be a non-negative random variable, and $\alpha > 0$. Then, $\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$.

(l) (To be covered in class 12/1) **Chebyshev's Inequality** (we did not cover this in class): Suppose Y is a random variable with $\mathbb{E}[Y] = \mu$ and Var$(Y) = \sigma^2$. Then, for any $\alpha > 0$, $\mathbb{P}(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$.

(m) **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 $X \sim$ Binomial(n, p) and $\mu = np$. Then, for any $0 < \delta < 1$,

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

(n) (To be covered in class 12/1) **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.

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:

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

7. Faulty Machines

You are trying to use a machine that only works on some days. If on a given day, the machine is working it will break down the next day with probability $0 < b < 1$, and works on the next day with probability $1 - b$. If it is not working on a given day, it will work on the next day with probability $0 < r < 1$ and not work the next day with probability $1 - r$.

(a) In this problem we will formulate this process as a Markov chain. First, let X_t be a random variable that denotes the state of the machine at time t. Then, define a state space $\mathcal{S}$ that includes all the possible states that the machine can be in. Lastly, for all $A, B \in \mathcal{S}$ find $\mathbb{P}(X_{t+1} = A \mid X_t = B)$ (A and B can be the same state).

(b) Suppose that on day 1, the machine is working. What is the probability that it is working on day 3?

(c) As $n \to \infty$, what does the probability that the machine is working on day n converge to? To get the answer, solve for the *stationary distribution*.

8. Another Markov chain

Suppose that the following is the transition probability matrix for a 4 state Markov chain (states 1,2,3,4).

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 0 & 2/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/5 & 2/5 & 2/5 & 0 \end{bmatrix}$$

(a) What is the probability that $X_2 = 4$ given that $X_0 = 4$?

(b) Write down the system of equations that the stationary distribution must satisfy and solve them.

9. Three tails

You flip a fair coin until you see three tails in a row. Model this as a Markov chain with the following states:

- S: start state, which we are only in before flipping any coins.
- H: We see a heads, which means no streak of tails currently exists.
- T: We've seen exactly one tail in a row so far.
- TT: We've seen exactly two tails in a row so far.
- TTT: We've accomplished our goal of seeing three tails in a row and stop flipping.

(a) Write down the transition probability matrix.

(b) Write down the system of equations whose variables are $D(s)$ for each state $s \in \{S, H, T, TT, TTT\}$, where $D(s)$ is the expected number of steps until state TTT is reached starting from state s. Solve this system of equations to find $D(S)$.

(c) Write down the system of equations whose variables are $\gamma(s)$ for each state $s \in \{S, H, T, TT, TTT\}$, where $\gamma(s)$ is the expected number of heads seen before state TTT is reached. Solve this system to find $\gamma(S)$, the expected number of heads seen overall until getting three tails in a row.

10. 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?

11. 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.

12. Poisson CLT practice

Suppose $X_1, \ldots, X_n$ are iid Poisson(λ) random variables, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the sample mean. How large should we choose n to be such that $\mathbb{P}(\frac{\lambda}{2} \leq \overline{X}_n \leq \frac{3\lambda}{2}) \geq 0.99$? Use the CLT and give an answer involving $\Phi^{-1}(\cdot)$. Then evaluate it exactly when $\lambda = 1/10$ using the Φ table on the last page.

13. Tail bounds

Suppose $X \sim$ Binomial$(6, 0.4)$. We will bound $\mathbb{P}(X \geq 4)$ using the tail bounds above, 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?

(b) (optional) 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.

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

(d) Give the exact probability.

14. MAP Estimation (bonus material that we will not cover)

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?