CSE 312: Foundations of Computing II

Section 9: Tail Bounds, Maximum Likelihood, Markov Chains Solutions

1. Review of Main Concepts

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

(b) **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 $\alpha > 0$, $\mathbb{P}(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$.

(c) **Chernoff Bound (for the Binomial)**:(We did 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 $0 < \delta < 1$,

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

(d) **Weak Law of Large Numbers (WLLN)**: (We have not covered this in class, but good to know.) 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.

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

(f) **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:

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

If X is continuous:

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

(g) **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).

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

(h) **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:

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

(i) **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 $\mathbb{E}[\hat{\theta}] = \theta$.

(j) **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.

(k) (To be covered in class 12/4) A **discrete-time stochastic process (DTSP)** is a sequence of random variables $X_0, X_1, X_2, ...$, 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.

(l) (To be covered in class 12/4) 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. ...has **stationary transition probabilities**. Meaning, if we are at some state s_i, we transition to another state s_j with probability *independent* of the current time. Due to this property and the previous, the transitions are governed by n^2 probabilities: the probability of transitioning from one of n current states to one of n next states. These are stored in a square $n \times n$ **transition probability matrix (TPM)** P, 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 for any/every value of t.

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, $\hat{\theta}$.

Solution:

The data tells us, for each student in the class, what their grade was. We begin by computing the likelihood of seeing the given data given our parameter θ. Because each student is assigned a grade independently, the likelihood is equal to the product over students of the chance they got the particular grade they got, which gives us:

$$L(x|\theta) = 0.5^{x_A} \theta^{x_B} (2\theta)^{x_C} (0.5 - 3\theta)^{x_F}$$

From there, we just use the MLE process to get the log-likelihood, take the first derivative, set it equal to 0, and solve for $\hat{\theta}$.

$$\ln L(x|\theta) = x_A \ln(0.5) + x_B \ln(\theta) + x_C \ln(2\theta) + x_F \ln(0.5 - 3\theta)$$

$$\frac{\partial}{\partial \theta} \ln L(x|\theta) = \frac{x_B}{\theta} + \frac{x_C}{\theta} - \frac{3x_F}{0.5 - 3\theta} = 0$$

Solving yields $\hat{\theta} = \frac{x_B + x_C}{6(x_B + x_C + x_F)}$.

3. A Red Poisson

(This problem was done in class.) 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.

Solution:

Because each Poisson RV is i.i.d., the likelihood of seeing that data is just the PMF of the Poisson distribution multiplied together for every x_i. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}
L\left(x_1, \ldots, x_n \mid \theta\right) &= \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \\
\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-\theta - \ln(x_i!) + x_i \ln(\theta)\right] \\
\frac{\partial}{\partial \theta} \ln L\left(x_1, \ldots, x_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta}\right] = 0 \\
-n + \frac{\Sigma_{i=1}^{n} x_i}{\hat{\theta}} &= 0 \\
\hat{\theta} = \frac{\Sigma_{i=1}^{n} x_i}{n}
\end{aligned}$$

4. Independent Shreds, You Say?

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?

Solution:

Note that $\Sigma_{i \in [n]} x_i = 30$, as given in the problem spec. Therefore, there are 30 **1**s and 70 **0**s. (Note that they come in some specific order.) Therefore, we can setup L as follows, because there is a θ chance of getting a 1, and a $(1-\theta)$ chance of getting a 0 and they are each i.i.d. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}
L\left(x_1, \ldots, x_n \mid \theta\right) &= (1-\theta)^{70} \theta^{30} \\
\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= 70 \ln(1-\theta) + 30 \ln \theta \\
\frac{\partial}{\partial \theta} \ln L\left(x_1, \ldots, x_n \mid \theta\right) &= -\frac{70}{1-\theta} + \frac{30}{\theta} = 0 \\
\frac{30}{\hat{\theta}} &= \frac{70}{1-\hat{\theta}} \\
30 - 30\hat{\theta} &= 70\hat{\theta} \\
\hat{\theta} &= \frac{30}{100}
\end{aligned}$$

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

Solution:

An estimator is unbiased if the expectation of the estimator is equal to the original parameter, i.e.: $E[\hat{\theta}] = \theta$. Setting up the expectation of our estimator and plugging it in for the generic case, we get the

following, which we can then reduce with linearity of expectation:

$$\begin{aligned}\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{1}{100}\sum_{i=1}^{100} X_i\right] \\ &= \frac{1}{100}\sum_{i=1}^{100}\mathbb{E}[X_i] \\ &= \frac{1}{100}\cdot 100\theta = \theta.\end{aligned}$$

so it is unbiased.

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.

Solution:

Since the samples are i.i.d., the likelihood of seeing n samples of them is just their PDFs multiplied together. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}L\left(y_1,\ldots,y_n \mid \theta\right) &= \prod_{i=1}^{n}\frac{1}{2\theta}\exp(-\frac{|y_i|}{\theta}) \\ \ln L\left(y_1,\ldots,y_n \mid \theta\right) &= \sum_{i=1}^{n}\left[-\ln 2 - \ln\theta - \frac{|y_i|}{\theta}\right] \\ \frac{\partial}{\partial\theta}\ln L\left(y_1,\ldots,y_n \mid \theta\right) &= \sum_{i=1}^{n}\left[-\frac{1}{\theta} + \frac{|y_i|}{\theta^2}\right] = 0 \\ -\frac{n}{\hat{\theta}} + \frac{\Sigma_{i=1}^{n}|y_i|}{\hat{\theta}^2} &= 0 \\ \hat{\theta} = \frac{\Sigma_{i=1}^{n}|y_i|}{n}\end{aligned}$$

6. A biased estimator

In class, we showed that the maximum likelihood estimate of the variance θ_2 of a normal distribution (when both the true mean μ and true variance σ^2 are unknown) is what's called the *sample variance*. That is

$$\hat{\theta}_2 = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_1)^2)\right)$$

where $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the MLE of the mean. Is $\hat{\theta}_2$ unbiased?

Solution:

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$E(\hat{\theta}_2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right)$$

which by linearity of expectation is

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E\left(\frac{2}{n}\overline{X}\sum_{i=1}^{n} X_i\right) + E(\overline{X}^2) \qquad (*)$$
$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - 2E(\overline{X}^2) + E(\overline{X}^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2). \qquad (**)$$

We know that for any random variable Y, since $Var(Y) = E(Y^2) - (E(Y))^2$ it holds that

$$E(Y^2) = Var(Y) + (E(Y))^2.$$

Also, we have $E(X_i) = \mu,\ Var(X_i) = \sigma^2\ \forall i$ and $E(\overline{X}) = \mu,\ Var(\overline{X}) = \frac{\sigma^2}{n}$. Combining these facts, we get

$$E(X_i^2) = \sigma^2 + \mu^2 \quad \forall i \quad \text{and} \quad E(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2.$$

Substituting these equations into (**) we get

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= \left(1 - \frac{1}{n}\right)\sigma^2.$$

Thus $\hat{\theta}_2$ is not unbiased.

7. Faulty Machines

(Markov Chains discussed in class 12/4). 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).

Solution:

Formally, a Markov chain is defined by a state space $\mathcal{S}$ and a transition probability matrix. The two possible states of the machine are "working" and "broken". So, $\mathcal{S} = \{W, B\}$. Let X_t be the state of the process at time t. Then we can define the following transition probabilities:

$$\mathbb{P}(X_{t+1} = W \mid X_t = W) = 1 - b \qquad \mathbb{P}(X_{t+1} = B \mid X_t = W) = b$$
$$\mathbb{P}(X_{t+1} = W \mid X_t = B) = r \qquad \mathbb{P}(X_{t+1} = B \mid X_t = B) = 1 - r$$

We can also represent the TPM with the following matrix:

$$P = \begin{bmatrix} 1-b & b \\ r & 1-r \end{bmatrix}$$

where the ijth entry is probability that the machine is in the jth state at time $t+1$ given it was in state i at time t. (Here state 1 is working and state 2 is broken.)

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

Solution:

We are trying to find $\mathbb{P}(X_3 = W \mid X_1 = W)$. From the law of total probability, and then plugging in the values from our transition matrix:

$$\begin{aligned}
P(X_3 = W \mid X_1 = W) &= \sum_{i \in \mathcal{S}} \mathbb{P}(X_3 = W \mid X_1 = W, X_2 = i)\, \mathbb{P}(X_2 = i \mid X_1 = W) \\
&= \mathbb{P}(X_3 = W \mid X_2 = W)\, \mathbb{P}(X_2 = W \mid X_1 = W) + \mathbb{P}(X_3 = W \mid X_2 = B)\, \mathbb{P}(X_2 = B \mid X_1 = W) \\
&= \mathbb{P}(X_3 = W \mid X_2 = W)\, (1-b) + \mathbb{P}(X_3 = W \mid X_2 = B)\, b \\
&= (1-b)(1-b) + rb \\
&= (1-b)^2 + rb
\end{aligned}$$

(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*.

Solution:

The stationary distribution is the row vector $\pi = [\pi_W \ \pi_B]$ such that $\pi P = \pi$. The entries in the vector π_W and π_B can be interpreted as the probabilities that the machine works or is broken converge to. As such, $\pi_W + \pi_B = 1$. Additionally, multiplying the stationary distribution by the TPM gives us the following two equations:

$$\pi_W = \pi_W(1-b) + \pi_B r \qquad \pi_B = \pi_W b + \pi_B(1-r)$$

Solving each for π_W and π_B gives us the following solutions for the stationary distribution:

$$\pi_W = \frac{r}{b+r} \qquad \pi_B = \frac{b}{b+r}$$

So, as $n \to \infty$ the probability that the machine works on day n is $\pi_W = \frac{r}{b+r}$

8. 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.

Solution:

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

(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 $D(S)$.

Solution:

Using the law of total expectation and the TPM above we can set up and solve the following system of equations:

$$\begin{aligned} D(TTT) &= 0 \\ D(TT) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(TTT) = \frac{1}{2}D(H) + 1 \\ D(T) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(TT) = \frac{3}{4}D(H) + \frac{3}{2} \\ D(H) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(T) = \frac{7}{8}D(H) + \frac{7}{4} \\ D(S) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(T) = \frac{7}{8}D(H) + \frac{7}{4} \end{aligned}$$

Solving for $D(H)$ gives us that $D(H) = 14$, which allows as to solve for the rest of the expected number of steps, $D(TT) = 8$, $D(T) = 12$, $D(S) = 14$. So, we expect to flip 14 coins before we flip three tails in a row.

(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.

Solution:

Like in the previous part we can use the LoTE and the TPM to set up and solve the follwoing system of equations:

$$\begin{aligned} \gamma(TTT) &= 0 \\ \gamma(TT) &= 0.5\gamma(H) + 0.5\gamma(TTT) = 0.5\gamma(H) \\ \gamma(T) &= 0.5\gamma(H) + 0.5\gamma(TT) = 0.75(H) \\ \gamma(H) &= 1 + 0.5\gamma(H) + 0.5\gamma(T) = 0.875\gamma(H) + 1 \\ \gamma(S) &= 0.5\gamma(H) + 0.5\gamma(T) = 0.875\gamma(H) \end{aligned}$$

Solving for $\gamma(H)$ gives us $\gamma(H) = 8$. This allows us to solve for the other expected values which are $\gamma(TT) = 4$, $\gamma(T) = 6$, $\gamma(S) = 7$. So, we expect to see 7 heads before we flip three tails in a row.

9. 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 $X_0 = 4$?

Solution:

Let's denote the state space $\mathcal{S} = \{1, 2, 3, 4\}$. Using the law of total probability we can determine that

$$\begin{aligned}
\mathbb{P}(X_2 = 4 \mid X_0 = 4) &= \sum_{i \in \mathcal{S}} \mathbb{P}(X_2 = 4 \mid X_0 = 4, X_1 = i)\, \mathbb{P}(X_1 = i \mid X_0 = 4) \\
&= \sum_{i \in \mathcal{S}} \mathbb{P}(X_2 = 4 \mid X_1 = i)\, \mathbb{P}(X_1 = i \mid X_0 = 4) \\
&= 0 + \frac{2}{5} \cdot \frac{2}{3} + \frac{2}{5} \cdot \frac{1}{3} + 0 \\
&= \frac{2}{5}
\end{aligned}$$

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

Solution:

The stationary distribution is the row vector $\pi = [\pi_1 \ \pi_2 \ \pi_3 \ \pi_4]$ such that $\pi P = \pi$. We know that $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$. Additionally, multiplying the stationary distribution by the TPM gives us the following equations:

$$\begin{aligned}
\pi_1 &= \frac{1}{3}\pi_2 + \frac{1}{3}\pi_3 + \frac{1}{5}\pi_4 \\
\pi_2 &= \frac{1}{2}\pi_1 + \frac{1}{3}\pi_3 + \frac{2}{5}\pi_4 \\
\pi_3 &= \frac{1}{2}\pi_1 + \frac{2}{5}\pi_4 \\
\pi_4 &= \frac{2}{3}\pi_2 + \frac{1}{3}\pi_3
\end{aligned}$$

Solving for each π_i gives us the following solutions for the stationary distribution:

$$\pi_1 = \frac{46}{206} \qquad \pi_2 = \frac{60}{206} \qquad \pi_3 = \frac{45}{206} \qquad \pi_4 = \frac{55}{206}$$

10. 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 $\mathbb{P}(H)$?

Solution:

We can use the law of total probability, conditioning on $U = \frac{k}{n}$ for $k = 0, ..., n$. Note that the probability of getting heads conditioning on a fixed U value is U, and that the probability of U taking on any value in its range is $\frac{1}{n+1}$ since it is discretely uniform.

$$\mathbb{P}(H) = \sum_{k=0}^{n} \mathbb{P}(H|U = \frac{k}{n})\mathbb{P}(U = \frac{k}{n}) = \sum_{k=0}^{n} \frac{k}{n} \cdot \frac{1}{n+1} = \frac{1}{n(n+1)} \sum_{k=0}^{n} k = \frac{1}{n(n+1)} \frac{n(n+1)}{2} = \frac{1}{2}$$

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

Solution:

We do the same thing, this time using the continuous law of total probability. Note, this time, that we're conditioning on $U = u$ and taking the integral with respect to u, and that the density of U for any value in its range is 1 because it is uniformly random.

$$\mathbb{P}(H) = \int_{-\infty}^{\infty} \mathbb{P}(H|U = u) f_U(u) du$$

We can take the integral from 0 to 1 instead because outside of that range the density of U is 0.

$$= \int_0^1 \mathbb{P}(H|U = u) f_U(u) du = \int_0^1 u \cdot 1 du = \frac{1}{2}[u^2]_0^1 = \frac{1}{2}$$

(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.

Solution:

We use the continuous law of total probability again, this time not deriving it any further and sticking with negative infinity to infinity because we don't know the range of the RV X.

$$\mathbb{P}(E) = \int_{-\infty}^{\infty} \mathbb{P}(E|X = x) f_X(x) dx$$

11. 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.

Solution:

We know $\mathbb{E}[X_i] = \text{Var}(X_i) = \lambda$. By the CLT, $\overline{X}_n \approx \mathcal{N}\left(\lambda, \frac{\lambda}{n}\right)$, so we can standardize this normal approximation.

$$\mathbb{P}(\frac{\lambda}{2} \leq \overline{X}_n \leq \frac{3\lambda}{2}) \approx \mathbb{P}(\frac{-\lambda/2}{\sqrt{\lambda/n}} \leq Z \leq \frac{\lambda/2}{\sqrt{\lambda/n}}) = \Phi\left(\frac{\lambda/2}{\sqrt{\lambda/n}}\right) - \Phi\left(\frac{-\lambda/2}{\sqrt{\lambda/n}}\right)$$

$$= \Phi\left(\frac{\lambda/2}{\sqrt{\lambda/n}}\right) - \left(1 - \Phi\left(\frac{\lambda/2}{\sqrt{\lambda/n}}\right)\right) = 2\Phi\left(\frac{\lambda/2}{\sqrt{\lambda/n}}\right) - 1 \geq 0.99 \to \Phi\left(\frac{\lambda/2}{\sqrt{\lambda/n}}\right) \geq 0.995$$

$$\to \frac{\sqrt{\lambda}}{2}\sqrt{n} \geq \Phi^{-1}(0.995) \to n \geq \frac{4}{\lambda}\left[\Phi^{-1}(0.995)\right]^2$$

We have $\lambda = \frac{1}{10}$ and from the table, $\Phi^{-1}(0.995) \approx 2.575$ so that $n \geq \frac{4}{1/10} \cdot 2.575^2 = 265.225$. So $n = 266$ is the smallest value that will satisfy the condition.

12. Tail bounds

Suppose $X \sim$ Binomial$(6, 0.4)$. We will bound $\mathbb{P}(X \geq 4)$ 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:

We know that the expected value of a binomial distribution is np, so: $\mathbb{P}(X \geq 4) \leq \frac{\mathbb{E}[X]}{4} = \frac{2.4}{4} = 0.6$. We can use it since X is nonnegative.

(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.

Solution:

$\mathbb{P}(X \geq 4) = \mathbb{P}(X - 2.4 \geq 1.6) \leq \mathbb{P}(|X - 2.4| \geq 1.6)$ we can add those absolute value signs because that only adds more possible values, so it is an upper bound on the probability of $X - 2.4 \geq 1.6$. Then, using Chebyshev's inequality we get:
$\mathbb{P}(|X - 2.4| \geq 1.6) \leq \frac{\text{Var}(X)}{1.6^2} = \frac{1.44}{1.6^2} = 0.5625$

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

Solution:

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

(d) Give the exact probability.

Solution:

Since X is a binomial, we know it has a range from 0 to n (or in this case 0 to 6). Thus, the possible values to satisfy $X \geq 4$ are 4, 5, or 6. We plug in the PMF for each to get: $\mathbb{P}(X \geq 4) = \mathbb{P}(X = 4) + \mathbb{P}(X = 5) + \mathbb{P}(X = 6) = \binom{6}{4}(0.4)^4(0.6)^2 + \binom{6}{5}(0.4)^5(0.6) + \binom{6}{6}0.4^6 \approx 0.1792$

13. MAP Estimation (optional)

I recommend you 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 $\hat{\theta}_{MLE}$?

Solution:

We begin with finding the likelihood by multiplying the probabilities of seeing each of the independent realizations from the Ber(θ) distribution. From there, we just use the MLE process to get the log-likelihood, take the first derivative, set it equal to 0, and solve for $\hat{\theta}_{MLE}$.

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

(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/trials), 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 $\mathbb{P}(word \mid spam)$ 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?

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.