CSE 312: Foundations of Computing II

Central Limit Theorem, Tail Bounds, Maximum Likelihood 9 Solutions

Review of Main Concepts

(a) **Central Limit Theorem (CLT)**: Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $X = \sum_{i=1}^n X_i$, which has $\mathbb{E}[X] = n\mu$ and $\text{Var}(X) = n\sigma^2$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, which has $\mathbb{E}\left[\overline{X}\right] = \mu$ and $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$. $\overline{X}$ is called the *sample mean*. Then, as $n \to \infty$, $\overline{X}$ approaches the normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Standardizing, this is equivalent to $Y = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ approaching $\mathcal{N}(0,1)$. Similarly, as $n \to \infty$, X approaches $\mathcal{N}(n\mu, n\sigma^2)$ and $Y' = \frac{X - n\mu}{\sigma\sqrt{n}}$ approaches $\mathcal{N}(0,1)$.

It is no surprise that $\overline{X}$ has mean μ and variance σ^2/n – this can be done with simple calculations. The importance of the CLT is that, for large n, regardless of what distribution X_i comes from, $\overline{X}$ is *approximately normally distributed with mean* μ *and variance* σ^2/n. Don't forget the continuity correction, only when $X_1, \ldots, X_n$ are discrete random variables.

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

(c) **Chebyshev's Inequality**: Suppose Y is a random variable with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$. Then, for any $\alpha > 0$, $\mathbb{P}(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$.

(d) **Chernoff Bound (for the Binomial)**: This will not be on any homework or exams, but is good to know. It's stronger than the Chebyshev bound. Suppose $X \sim \text{Binomial}(n,p)$ and $\mu = np$. Then, for any $0 < \delta < 1$,

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

(e) **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.

(f) **Strong Law of Large Numbers (SLLN)**: 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, $\mathbb{P}(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$. We say that $\bar{X}_n$ converges almost surely to μ. The SLLN implies the WLLN, but not vice versa.

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

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

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

(i) **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).

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

(j) **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:

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

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

(l) **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.

Bad Computer

Each day, the probability your computer crashes is 10%, independent of every other day. Suppose we want to evaluate the computer's performance over the next 100 days.

(a) Let X be the number of crash-free days in the next 100 days. What distribution does X have? Identify $\mathbb{E}[X]$ and Var(X) as well. Write an exact (possibly unsimplified) expression for $\mathbb{P}(X \geq 87)$.

Solution:

$X \sim$ Binomial$(100, 0.9)$. Hence, $\mathbb{E}[X] = np = 90$ and Var$(X) = np(1-p) = 9$. Finally,

$$\mathbb{P}(X \geq 87) = \sum_{k=87}^{100} \binom{100}{k} (0.9)^k (1-0.9)^{100-k}$$

(b) Approximate the probability of at least 87 crash-free days out of the next 100 days using the Central Limit Theorem. Justify why we can use the CLT here.

Solution:

From the previous part, we know that $\mathbb{E}[X] = 90$ and Var$(X) = 9$.

$$\begin{aligned}
\mathbb{P}(X \geq 87) &= \mathbb{P}(86.5 < X < 100.5) = \mathbb{P}\left(\frac{86.5-90}{3} < \frac{X-90}{3} < \frac{100.5-90}{3}\right) \\
&\approx \mathbb{P}\left(-1.17 < \frac{X-90}{3} < 3.5\right) \approx \Phi(3.5) + \Phi(1.17) - 1 \approx 0.9998 + 0.8790 - 1 = 0.8788
\end{aligned}$$

Notice that, if you had used $86.5 < X$ in place of $86.5 < X < 100.5$, your answer would have been nearly the same, because $\Phi(3.5)$ is so close to 1.

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 $0.5-3\theta$. 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:

$$L(x|\theta) \propto 0.5^{x_A}\theta^{x_B}(2\theta)^{x_C}(0.5-3\theta)^{x_F}$$
$$\ln L(x|\theta) = x_A\ln(0.5)+x_B\ln(\theta)+x_C\ln(2\theta)+x_F\ln(0.5-3\theta)$$
$$\frac{\partial}{\partial\theta}\ln L(x|\theta) = \frac{x_B}{\theta}+\frac{x_C}{\theta}-\frac{3x_F}{0.5-3\theta}=0$$

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

Continuous Law of Total Probability Review

(a) 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 $k=0,...,n$.

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

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

Solution:

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

(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:

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

Independent Shreds, You Say?

You are given 100 independent samples $x_1,x_2,\ldots,x_{100}$ from Bernoulli(p), where p is unknown. These 100 samples sum to 30. You would like to estimate the distribution's parameter p. Give all answers to 3 significant digits.

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

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. Therefore, we can setup L as follows,

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

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

Solution:

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

so it is unbiased.

What if we lose ?

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.

Solution:

We will make an assumption here. We will assume that the i^{th} person is in favor of the proposition with probability $\frac{59}{100}$. We define $X_i \sim$ Bernoulli($\frac{59}{100}$) representing whether the i^{th} person is in favor or not. We define $X = \sum_{i=1}^{100} X_i$ representing the number of people who are in favor of the proposition. We can approximate X by $Y \sim N(100\cdot 0.59, 100\cdot 0.242)$. We need to find $\mathbb{P}(\frac{Y-59}{\sqrt{(24.2)}} < \frac{50.5-59}{\sqrt{(24.2)}})$(after continuity correction and standardization) which is equal to $\Phi(-1.729)$.

(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?

Solution:

We will use our normal approximation Y from part(a). We are interested in $\mathbb{P}(Y < -0.5) + \mathbb{P}(Y > 100.5)$(after continuity correction) which is the same as

$$\mathbb{P}(\frac{Y-59}{\sqrt{24.2}} < \frac{-0.5-59}{\sqrt{24.2}}) + \mathbb{P}(\frac{Y-59}{\sqrt{24.2}} > \frac{100.5-59}{\sqrt{24.2}}) = \Phi(-12.09) + 1 - \Phi(8.436)$$

. Yes, the probability will be non -zero because the density of the normal distribution is non-zero everywhere. Note that this result is acceptable because the normal distribution is an approximation.

Y Me?

Let $Y_1, Y_2, ...Y_n$ be i.i.d. random variables with density function

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

.

Find the the MLE for σ in terms of $|y_i|$.

Solution:

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

It Means Nothing

(a) Suppose $x_1, x_2, \ldots, x_n$ are samples from a normal distribution whose mean is known to be zero, but whose variance is unknown. What is the maximum likelihood estimator for its variance?

Solution:

Before we begin, we should note that this derivation will have to be with respect to σ^2, not σ. Therefore, we want to analyze the function $L(x_1, \ldots, x_n \mid \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$.

$$\begin{aligned}
L(x_1, \ldots, x_n \mid \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2} \\
\ln L(x_1, \ldots, x_n \mid \sigma^2) &= \sum_{i=1}^{n} -\ln\sqrt{2\pi\sigma^2} - \frac{x_i^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{x_i^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{x_i^2}{2\sigma^2} \\
&= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{\Sigma_{i=1}^{n} x_i^2}{2\sigma^2} \\
\frac{\partial}{\partial\sigma^2}\ln L(x_1, \ldots, x_n \mid \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{\Sigma_{i=1}^{n} x_i^2}{2\sigma^4} = 0 \\
\frac{\Sigma_{i=1}^{n} x_i^2}{2\sigma^4} &= \frac{n}{2\sigma^2} \\
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2
\end{aligned}$$

(b) Suppose the mean is known to be μ but the variance is unknown. How does the maximum likelihood estimator for the variance differ from the maximum likelihood estimator when both mean and variance are unknown?

Solution:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

vs.

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

(The former turns out to be unbiased, the latter biased.)