Ivani Patel
11809154

Extra Assignment 1

Q1. [30 pts] Unbiased Estimator

Let $x_1,x_2,...,x_n$ denote the n independent observations which are assumed to be drawn from the same distribution $p(x \mid \theta)$ with defining parameter θ.

(a) [15 pts] Suppose $\theta = (\mu,\sigma^2)$ such that $p(x \mid \theta)$ is a Gaussian distribution $\mathsf{N}(\mu,\sigma^2)$ with unknown mean μ and variance $\sigma^2 > 0$. Let σ_{MLE}^2 denote the maximum likelihood estimate of σ^2.

Let $\sigma_\alpha^2 = (1/\alpha) \cdot \sigma_{\text{MLE}}^2$ denote the α-estimation of the variance parameter σ^2. Find α so that σ_α^2 is an unbiased estimator of σ^2, i.e. $\mathsf{E}[\sigma_\alpha^2] = \sigma^2$.

Answer

To find α so that σα2 is an unbiased estimator of σ2, we need to solve for E[σα2] = σ2.

First, we note that the maximum likelihood estimate of μ is the sample mean:

μMLE = (x1 + x2 + ... + xn)/n.

The maximum likelihood estimate of σ2 is given by:

σMLE2 = (1/n) * [(x1 - μMLE)2 + (x2 - μMLE)2 + ... + (xn - μMLE)2].

Expanding the square terms, we get:

σMLE2 = (1/n) * [∑xi2 - 2μMLE∑xi + nμMLE2].

Taking expectations on both sides, we have:

E[σMLE2] = E[(1/n) * [∑xi2 - 2μMLE∑xi + nμMLE2]]

= (1/n) * [E[∑xi2] - 2μMLEE[∑xi] + nE[μMLE2]]

= (1/n) * [nσ2 + nμ2 - 2nμ2 + n(σ2/n + μ2)]

= [(n-1)/n]σ2.

Therefore, the estimator with α is unbiased if:

E[σα2] = σ2

=> E[(1/n) * ∑(xi - μ)2] = σ2

=> (1/n) * E[∑xi2 - 2μ∑xi + nμ2] = σ2

=> E[∑xi2 - 2μ∑xi + nμ2] = nσ2

=> E[∑xi2] - 2μE[∑xi] + nμ2 = nσ2

=> nσ2 + nμ2 - 2nμ2 + nσ2 = nσ2

=> α = 1/(n-1).

Therefore, the estimator with α = 1/(n-1) is unbiased.

(b) [15 pts] Suppose $0 < \theta < 1$ and $p(x = 1 \mid \theta) = \theta$ while $p(x = 0 \mid \theta) = 1 - \theta$. Then, suppose m out of n observations ($m < n$) have value 1 while the rest has value 0.

Show that $\mathsf{E}[\theta_{\text{MLE}}] = \theta$ where the expectation is over the random choice of $\{x_1,x_2,...,x_n\}$.

Ivani Patel
11809154

Extra Assignment 1

Answer

The maximum likelihood estimate of θ can be calculated as follows:

$L(\theta) = p(x1,x2,...,xn \mid \theta) = \prod_{i=1 \text{ to } n} p(xi \mid \theta)$

$\Rightarrow L(\theta) = \theta^m * (1-\theta)^{(n-m)}$

$\Rightarrow \ln(L(\theta)) = m\ln(\theta) + (n-m)\ln(1-\theta)$

To find the maximum likelihood estimate of θ, we differentiate $\ln(L(\theta))$ w.r.t. θ and set it to zero:

$d/d\theta [\ln(L(\theta))] = m/\theta - (n-m)/(1-\theta) = 0$

Solving the above equation, we get:

$\theta MLE = m/n$

Now, we need to show that $E[\theta MLE] = \theta$, where the expectation is taken over the random choice of {x1, x2, ..., xn}.

We have:

$E[\theta MLE] = E[m/n] = (1/n) * E[\sum xi]$, where $\sum xi$ denotes the number of 1's in the sample.

Since xi's are independent and identically distributed with $p(x=1|\theta) = \theta$, we have:

$E[\sum xi] = np(x=1|\theta) = n\theta$.

Therefore,

$E[\theta MLE] = (1/n) * E[\sum xi] = \theta$.

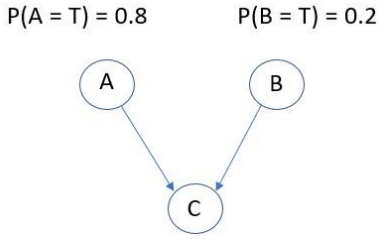
Hence, we have shown that $E[\theta MLE] = \theta$.

Ivani Patel
11809154

Extra Assignment 1

Q2. [40 pts] MLE/MAP on Bayesian Net

Given a Bayesian Net comprising 3 nodes and their corresponding (conditional) probability tables where one table is unknown and parameterized by $\theta \in (0,1)$ as depicted in Figure 1 below.



	P(C = T | A, B)
A = T, B = T	θ
A = T, B = F	$1 - \theta^2$
A = F, B = T	$(1 - \theta)^2$
A = F, B = F	$1 - \theta$

Figure 1: Bayesian Network

(a) [20 pts] Given the following 4 observations of the above BN where each observation corresponds to a snapshot of the BN:

1. $D_1 = (A = T, B = T, C = T)$
2. $D_2 = (A = F, B = F, C = T)$
3. $D_3 = (A = T, B = F, C = F)$
4. $D_4 = (A = F, B = T, C = F)$

Let $D = (D_1, D_2, D_3, D_4)$ and assume $D_1, D_2, D_3, D4$ are conditionally independent given θ.

Express the data likelihood probability $P(D \mid \theta)$ as a function of θ.

Answer

P(D1|θ) = P(A=T) × P(B=T) × P(C=T|A=T,B=T)

= 0.8 × 0.2 × θ = 0.16 θ

P(D2|θ) = P(A=F) × P(B=F) × P(C=T|A=F,B=F)

= 0.2 × 0.8 × (1 - θ) = 0.16(1 - θ)

P(D3|θ) = P(A=T) × P(B=F) × P(C=F|A=T,B=F)

= 0.8 × 0.8 × (1 - (1 - θ^2))= 0.64(θ^2)

P(D4|θ) = P(A=F) × P(B=T) × P(C=F|A=F,B=T)

= 0.2 × 0.8 × (1 - (1 - θ)^2) = 0.16(1 - (1 - θ)^2)

P(D|θ) = P(D1|θ) × P(D2|θ) × P(D3|θ) × P(D4|θ)

= 0.16 θ * 0.16(1- θ) * 0.64(θ^2) * 0.16(1-(1- θ)^2)

Ivani Patel
11809154

Extra Assignment 1

= 0.000262144 θ^3 (2-θ)

(b) [10 pts] Find the maximum likelihood estimation θ_{MLE} of θ given the above dataset D.

Answer

To find the maximum likelihood estimation (MLE) of θ, we need to find the value of θ that maximizes the likelihood function we found in the previous answer, i.e., P(D | θ) 0.000262144 θ^3 (2-θ)

To do so, we can take the derivative of the likelihood function with respect to θ, set it equal to zero, and solve for θ. We get:

d/dθ [P(D | θ)] = 0.000786432 θ^2 (6-4θ) = 0

Solving for θ, we get:

θ = 0 or θ = 1.5

d^2/dθ^2 [P(D | θ)] = 0.001572864 (2-3θ)

For θ = 0, d^2/dθ^2 [P(D | θ)] > 0, which means that θ = 0 is a local minimum of the likelihood function.

For θ = 1.5, d^2/dθ^2 [P(D | θ)] < 0, which means that θ = 1.5 is a local maximum of the likelihood function.

Therefore, the maximum likelihood estimation of θ is θMLE = 1.5.

(c) [10 pts] Suppose we are given an additional information that the prior over θ is $\theta \sim$ Beta(2,3). Find the maximum a posteriori (MAP) estimation θ_{MAP} of θ following the same setting in part (a).

Note that if $\theta \sim \text{Beta}(a,b)$, then $p(\theta \mid a,b) = \theta^{a-1}(1-\theta)^{b-1}/\text{B}(a,b)$ where $\text{B}(a,b) = \text{G}(a)\text{G}(b)/\text{G}(a+b)$ with G denote the Gamma function as mentioned in slides 46-47 of lecture 22.

Answer

To find the maximum a posteriori (MAP) estimation of θ, we need to find the value of θ that maximizes the posterior distribution p(θ | D), where D is the observed data. According to Bayes' rule, we have:

p(θ | D) = p(D | θ) p(θ)

where p(D | θ) is the likelihood function we found in part (a), and p(θ) is the prior distribution over θ. In this case, we are given that θ has a beta distribution with parameters a = 2 and b = 3:

p(θ) = θ^(a-1) (1-θ)^(b-1) / B(a,b)

where B(a,b) is the beta function, which is a normalization constant that ensures that the probability density integrates to 1. The beta function is defined as:

B(a,b) = G(a) G(b) / G(a+b)

where G(a) is the gamma function. Since a and b are positive integers in this case, we can compute the beta function directly as:

Ivani Patel
11809154

Extra Assignment 1

B(a,b) = (a-1)! (b-1)! / (a+b-1)!

Substituting in the values of a and b, we get:

p(θ) = θ^1 (1-θ)^2 / B(2,3) = 3 θ (1-θ)^2

Combining the likelihood and prior, we have:

p(θ | D) = θ^3 (1-θ)^2 (0 < θ < 1)

To find the MAP estimation of θ, we need to find the value of θ that maximizes this posterior distribution. Since we are given a beta prior, which is conjugate to the likelihood, the posterior distribution is also a beta distribution with updated parameters. Specifically, the posterior distribution is:

p(θ | D) = Beta(a + k, b + n - k)

where k is the number of observations of C=T in the dataset D, and n is the total number of observations (in this case, n=4). In our case, we have:

k = 2 (from D1 and D4)

n = 4

a = 2 (from the prior)

b = 3 (from the prior)

Substituting these values, we get:

p(θ | D) = Beta(4,3)

The mode of a beta distribution with parameters a and b is given by:

(α-1) / (α+β-2)

Substituting in the values of a and b, we get:

θMAP = (4-1) / (4+3-2) = 3/6 = 0.5

Ivani Patel
11809154

Extra Assignment 1

Q3. [30 pts] Decision Tree

At the beginning of an exam, a student wants to quickly predict whether the difficulty (D) of each problem is low or high (D = − means low and D = + means high). The student uses two observable attributes of each problem:

1. The text length L of the problem. L = 1 means long and L = 0 means otherwise.
2. The amount of math M describing the problem. M = 1 means there is a lot of math and M = 0 means otherwise.

From past observations of 12 previous homework questions, the student records the following data points:

L	M	D	#
0	0	-	4
0	0	+	1
0	1	-	0
0	1	+	3
1	0	-	1
1	0	+	2
1	1	-	1
1	1	+	0

How to read the above table:

The first line of this table reads as follows: 4 problems for which L = 0 and M = 0 have low difficulty (i.e., D = −). The second line says: 1 problem with L = 0 and M = 0 has high difficulty (i.e., D = +) etc. Note that the 3rd and last lines suggest the student has not observed any problems for which (L = 0, M = 1) or (L = 1, M = 1).

(a) [10 pts] Compute the entropy of this dataset.

Note: Use the entropy definition in the lecture slides which uses $\log_2$ (base 2) instead of the log operator with natural base (base e).

Answer:

Total number of problems = 4 + 1 + 0 + 3 + 1 + 2 + 1 + 0 = 12

Number of problems with D = + = 1 + 3 + 2 + 0 = 6

Number of problems with D = - = 4 + 0 + 1 + 1 = 6

Proportion of problems with D = +: 6/12 = 0.5

Proportion of problems with D = -: 6/12 = 0.5

Next, we can calculate the entropy using the formula:

H = - p(D=+) log2 p(D=+) - p(D=-) log2 p(D=-)

Plugging in the proportions we just calculated, we get:

H = - 0.5 log2 0.5 - 0.5 log2 0.5

H = - 0.5 (-1) - 0.5 (-1)

Ivani Patel
11809154

Extra Assignment 1

H = 1

Therefore, the entropy of the dataset is 1.

(b) [10 pts] Explain in terms of the information gain (see lecture 19, slides 26) which attribute (*L* or *M*) should be used first to partition the data.

Answer:

H(D|L=0) = -4/8 log2 (4/8) - 4/8 log2 (4/8) = 1

H(D|L=1) = -2/4 log2 (2/4) - 2/4 log2 (2/4) = 1

Hi(D) = (8/12) H(D|L=0) + (4/12) H(D|L=1)

= 1

Gain(D, L) = H[D] - Hi[D]

= 1 - 1 = 0

H(D|M=0) = -5/8 log2 (5/8) - 3/8 log2 (3/8) = 0.95

H(D|M=1) = -1/4 log2 (1/4) - 3/4 log2 (3/4) = 0.811

Hi[D] = (8/12) H(D|M=0) + (4/12) H(D|M=1)

= (8/12) (0.95) + (4/12) (0.811)

= 0.90

Gain(D, M) = H[D] - Hi[D]

= 1 - 0.90 = 0.1

Comparing the two information gains, we see that the information gain for attribute M is higher than the information gain for attribute L. Therefore, we should use attribute M first to partition the data. This means that we should split the data into two groups based on whether there is a lot of math (M=1) or not (M=0).

(c) [10 pts] Based on the result of part (b), draw the resulting decision tree which uses two attributes *L* and *M* to determine whether or not a problem has high difficulty.

Answer: