2️⃣

Assignment - 2

Pradyoth Singenahalli Prabhu – 02071847

Woodi Raghavendra Varun – 02070206

Bharath Anand – 02044023

1. What are typical reasons for overfitting? Please list at least three reasons.

Solution: A typical issue in machine learning is overfitting, in which a model does well on the training data but fails to generalize on new, untried data.

1. **Complex models:** Overfitting can occur when a model is excessively complex, meaning it has an excessive number of parameters compared to the available training data. In such cases, the model may end up fitting the noise present in the data, rather than the actual underlying patterns.
2. **Insufficient training data:** Overfitting may also occur due to inadequate training data, especially in deep learning, where models can contain millions of parameters and demand extensive data for effective training. In such cases, the model may be compelled to fit the noise present in the limited data, instead of capturing the essential underlying patterns, leading to overfitting.
3. **Lack of regularization:** To avoid overfitting, regularization methods like L1 or L2 regularization, early stopping, or dropout can be utilized to limit the model's complexity or halt the training process before overfitting occurs. Failure to apply these techniques may cause the model to keep improving its performance on the training data but perform poorly on new data due to a lack of generalization, which results in overfitting.

2. “In fruit example, if each box contained the same fraction of apples and oranges then P($F|B$) = P(F)” Can you prove that?

Solution: Assuming that "P(F|B)" refers to the probability of selecting an orange given that the fruit comes from a particular box B, and "P(F)" refers to the probability of selecting an orange without any information about which box it came from, we can prove that P(F|B) = P(F) under the given conditions.

Below we have proved:

```
P(A)= P(O) = 1/2

P(F/B) = (P(B/F) * P(F)) / P(B)

P(B/F) = 1/3

P(F) = 1/2

P(B) = P(F) * P(B/F) + P(F) * P(B/F)
     = (1/2 * 1/3) + (1/2 * 1/3)
     = 1/3

P(F/B) = (P(B/F) * P(F)) / P(B)
       = (1/3 * 1/2) / 1/3
       = 1/2
       = P(F)
```

Therefore, `P(F|B) = P(F)`,under the given conditions.

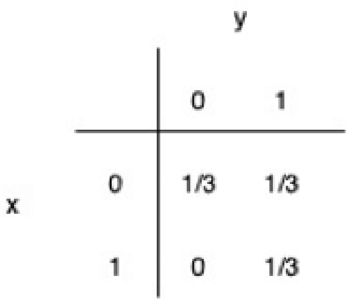
3. True or False

1. From a Bayesian perspective, the probability is a quantification of uncertainty. - **True**
2. Uniform distribution will provide the smallest entropy for discrete variables. - **False**
3. With more nodes included in the decision tree, the testing error will become smaller and

   smaller. - **False**
4. If we toss a coin three times and always get “head”, then from Bayesian perspective, we

   believe Pr(x is head) = 1. - **False**
5. There might be more than one decision trees that fit the same data. - **True**
6. In decision tree, higher impurity means lower Gini values. - **False**

4. Consider the below joint probability distribution between *x* and *y*.



		y	
		0	1
x	0	1/3	1/3
	1	0	1/3

a. **Please compute the marginal probabilities *p*(*x*) and *p*(*y*) for each value `x` and `y` take on.**

Solution: We must add the joint probabilities over the other variable in order to get the marginal probabilities p(X) and p(Y).

We must add the probabilities for each row in order to determine the marginal probability p(X):

```
p(x0) = 1/3 + 1/3 = 2/3
p(x1) = 0 + 1/3 = 1/3
```

To compute the marginal probability *p*(Y), we need to sum the probabilities of each column:

```
p(y1) = 1/3 + 1/3 = 2/3
p(y0) = 1/3 +0 = 1/3
```

b. Are `x` and `y` independent? Prove your result.

Solution: To determine the independence of X and Y, we need to compare the marginal probabilities p(X, Y) with the product of the joint probabilities p(X, Y). If the joint probabilities are equivalent to the product of the marginal probability for all values of X and Y, then we can conclude that X and Y are independent..

Lets take X = x1 and Y = y1,

```
p(X = x1, Y = y1) = 1/3
p(X = x1) * p(Y = y1) = 2/3 * 1/3 = 2/9
p(X = x1, Y = y1) ≠  p(X = x1) * p(Y = y1).

Now after doing the same thing for X = x1 and Y = y2 , X = x2 and Y = y1 and X = x2 and Y = y2.
```

Upon comparing the marginal probabilities p(X,Y) and the product of the joint probabilities p(X,Y) in the above calculations, it is observed that the marginal probabilities are not equal to the joint probabilities. Hence, it can be concluded that X and Y are **not independent.**

c. Compute conditional probabilities *p*(*x*|*y*) and *p*(*y*|*x*) for each value *x* and *y* take on.

Solution: To compute conditional probabilities, we use the following formula: Bayes theorem

```
P(x|y) = P(x,y) / P(y)
P(y|x) = P(x,y) / P(x)
```

In the probability distribution table provided, let's calculate the conditional probabilities for each value of x and y:

```
For x = x1 and y = y1:

P(x1,y1) = 1/3
P(y1) = 1/3 + 0 = 1/3

P(x1|y1) = P(x1,y1) / P(y1) = (1/3) / (1/3) = 1
P(y1|x1) = P(x1,y1) / P(x1) = (1/3) / (1/3) = 1

For x = x1 and y = y2:

P(x1,y2) = 1/3
P(y2) = 1/3 + 1/3 = 2/3

P(x1|y2) = P(x1,y2) / P(y2) = (1/3) / (2/3) = 1/2
P(y2|x1) = P(x1,y2) / P(x1) = (1/3) / (1/3) = 1

For x = x2 and y = y1:

P(x2,y1) = 0
P(y1) = 1/3 + 0 = 1/3

P(x2|y1) = P(x2,y1) / P(y1) = 0
P(y1|x2) = P(x2,y1) / P(x2) = undefined (division by zero)

For x = x2 and y = y2:

P(x2,y2) = 1/3
P(y2) = 1/3 + 1/3 = 2/3

P(x2|y2) = P(x2,y2) / P(y2) = (1/3) / (2/3) = 1/2
P(y2|x2) = P(x2,y2) / P(x2) = (1/3) / 0 = undefined (division by zero)
```

5. Can you find the probability of $p(B = b|F = a)$? Please also indicate the prior, likelihood, and posterior terms in your solutions.

To find the probability of `p(B=b|F=a)`, we can use Bayes' theorem, which states that:

```
p(B=b|F=a) = p(F=a|B=b) * p(B=b) / p(F=a)
```

where `p(F=a|B=b)` is the likelihood term, `p(B=b)` is the prior term, and `p(F=a)` is the marginal likelihood or evidence term.

Let's calculate each term one by one:

Prior term:
The probability of selecting box b, denoted as p(B=b), is independent of any knowledge regarding the fruit. As there are only two boxes, the prior probability of selecting either box is 1/2.

```
p(B=b) = 1/2
```

Likelihood term:
p(F=a|B=b) is the probability of selecting an apple from box b. From the problem statement, we know that box blue has 3 apples and 1 orange, and box red has 2 apples and 6 oranges. Therefore, the probability of selecting an apple from box blue is 3/4, and the probability of selecting an apple from box red is 2/8 (which can be simplified to 1/4). So, the likelihood term is:

```
p(F=a|B=b) = 3/4, if b=blue
p(F=a|B=b) = 1/4, if b=red
```

Evidence term:
p(F=a) is the probability of selecting an apple, regardless of which box it came from. To calculate this, we need to consider all the possible ways of selecting an apple:

```
p(F=a) = p(F=a|B=blue) * p(B=blue) + p(F=a|B=red) * p(B=red)
p(F=a) = (3/4 * 1/2) + (1/4 * 1/2)
p(F=a) = 1/2
```

Now we can put everything together to get the posterior probability:

```
p(B=b|F=a) = p(F=a|B=b) * p(B=b) / p(F=a)
p(B=b|F=a) = (3/4 * 1/2) / (1/2)
p(B=b|F=a) = 3/4
```

Given that the fruit is an apple, the probability of choosing box blue is 3/4, and the probability of choosing box red is 1/4. Hence, it can be inferred that the most probable box that the apple came from is box blue.

6. Given N independent and identically distributed(i.i.d.)observations(x1,x2,...xN)∈ R1×N that follow the Gaussian distribution N(μ, σ2). Please answer the following questions:

a. Formulation of joint probability Pr(x1, x2, ... xN)

Since the observations are independent and identically distributed, the joint probability distribution of the observations can be expressed as the product of the individual probability distributions:

$$\Pr(x_1, x_2, \ldots, x_N) = \Pr(x_1) \cdot \Pr(x_2) \cdot \ldots \cdot \Pr(x_N)$$

Since each observation follows a Gaussian distribution with mean μ and variance σ^2, we can express the probability density function of xi as:

$$\Pr(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Therefore, the joint probability distribution can be written as:

$$\Pr(x_1, x_2, \ldots, x_N) = \frac{1}{\sqrt{(2\pi\sigma^2)^{N/2}}} \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right)$$

where Σ represents the sum over all observations from 1 to N.

b. Log-likelihood function of the above joint probability

The log-likelihood function for the joint probability of the observations can be derived as follows:

$$\Pr(x_1, x_2, \ldots, x_N) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right)$$

Taking the natural logarithm of both sides, we get:

$$\ln Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}N) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^2$$

This is the log-likelihood function of the joint probability. We can use this function to estimate the parameters of the Gaussian distribution using maximum likelihood estimation (MLE). Specifically, we can differentiate the log-likelihood function with respect to the parameters (mean and variance) and set the derivatives to zero to obtain the MLE estimates.

c. Maximum likelihood solution to μ and σ^2

To obtain the maximum likelihood estimates of the parameters μ and σ^2, we need to maximize the log-likelihood function with respect to these parameters.

Maximizing with respect to μ:

Taking the derivative of the log-likelihood function with respect to μ, we get:

$$\frac{d}{d\mu}\ln Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}N) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)$$

Setting this derivative to zero, we get:

$$\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i - \mu) = 0$$

Solving for μ, we get:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Therefore, the MLE estimate for μ is the sample mean of the observations.

Maximizing with respect to σ^2:

Taking the derivative of the log-likelihood function with respect to σ^2, we get:

$$\frac{d}{d(\sigma^2)}\ln Pr(x_1, x_2, ..., x_N) = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

Setting this derivative to zero, we get:

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

Therefore, the MLE estimate for σ^2 is the sample variance of the observations.

In summary, the MLE estimates for μ and σ^2 are:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} xi$$

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

d. Are the solutions to question (c) biased estimations or not, and why?

The maximum likelihood estimates (MLE) of μ and σ^2 derived in question (c) are unbiased estimators.

To see why, we can use the properties of expectation and variance of the sample mean and sample variance.

Unbiasedness of μ estimator:

$$E[\mu] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$$

Therefore, the MLE estimator for μ is unbiased.

Unbiasedness of σ^2 estimator:

`E[σ^2] = E[1/N * Σi=1 to N (xi - μ)^2] = 1/N * Σi=1 to N E[(xi - μ)^2] = 1/N * Σi=1 to N Var[xi] = σ^2`

Therefore, the MLE estimator for σ^2 is unbiased.

In summary, the MLE estimates for μ and σ^2 are unbiased estimators.

e. Consider the curve fitting problem $y = \mathbf{w}x$. Following Slides-5, we assume the observations of yi are means for Gaussians N(ti|yi = $\mathbf{w}xi$, σ2) that will be able to model the probability for different target values ti, and different Gaussians share σ2 =1/β.Please provide the maximum likelihood estimation for $\mathbf{w}$(derivations and details of each step are required).

Let's assume we have N observations {t1, t2, ..., tN} with corresponding inputs {x1, x2, ..., xN}. We model each observation as a Gaussian distribution with mean yi=wxi and variance σ2=1/β.
The probability density function for each observation is:

```
p(ti|w,β) = N(ti|yi,σ2) = N(ti|wxi,1/β)
```

The joint probability density function for all the observations is the product of individual probability density functions:

```
p(t1,t2,...,tN|w,β)=∏Ni=1N(ti|wxi,1/β)
```

Taking the negative logarithm of the joint probability density function, we get the negative log-likelihood function:

```
−ln(p(t1,t2,...,tN|w,β))=1/2σ2∑(ti−wxi)2+N/2ln(σ2)+constant
```

We can simplify this by substituting σ2=1/β:

```
−ln(p(t1,t2,...,tN|w,β))=β/2∑(ti−wxi)2+N/2ln(β)+constant
```

To find the maximum likelihood estimate of w, we differentiate the negative log-likelihood function with respect to w and set it to zero:

```
∂/∂w(−ln(p(t1,t2,...,tN|w,β))) = −β∑(ti−wxi)xi = 0
```

Solving for w, we get:

$$w = \sum N_i = 1 x_i t_i / \sum N_i = 2xi$$

Therefore, the maximum likelihood estimate for w is:

$$w^* = \sum Ni = 1 x_i t_i / \sum N_i = 2xi$$

7. a. First, please compute the: (1) entropy; (2) Gini; (3) classification error of cases

a. C0: 8 and C1: 12

Total instances = 20

1. **Entropy:**
   Entropy is defined as

   $$H(S) = -\sum_{x} p(x) \log_2(p(x))$$

   where p(x) is the proportion of instances of class x in the dataset.

   For case (a), we have:

   ```
   p(C0) = 8/20 = 0.4
   p(C1) = 12/20 = 0.6
   ```

   Therefore, the entropy is:

   ```
   H(S) = -[(0.4 * log2(0.4)) + (0.6 * log2(0.6))] = 0.971
   ```

2. **Gini:**
   Gini index is defined as

$$Gini(S) = 1 - \sum_{x}(p(x)^2)$$

   where p(x) is the proportion of instances of class x in the dataset.

   For case (a), we have:

```
p(C0) = 8/20 = 0.4
p(C1) = 12/20 = 0.6
```

   Therefore, the Gini index is:

```
Gini(S) = 1 - [(0.4^2) + (0.6^2)] = 0.480
```

3. **Classification error:**
   Classification error is defined as

$$Error(S) = 1 - max(p(x))$$

   where p(x) is the proportion of instances of class x in the dataset.

   For case (a), we have:

```
p(C0) = 8/20 = 0.4
p(C1) = 12/20 = 0.6
```

   Therefore, the classification error is:

```
Error(S) = 1 - max(0.4, 0.6) = 0.4
```

b. C0: 6 and C1: 14

Total instances = 20

1. **Entropy:**
   For case (b), we have:

```
p(C0) = 6/20 = 0.3
p(C1) = 14/20 = 0.7
```

Therefore, the entropy is:

```
H(S) = -[(0.3 * log2(0.3)) + (0.7 * log2(0.7))] = 0.881
```

2. **Gini:**
   For case (b), we have:

```
p(C0) = 6/20 = 0.3
p(C1) = 14/20 = 0.7
```

Therefore, the Gini index is:

```
Gini(S) = 1 - [(0.3^2) + (0.7^2)] = 0.420
```

3. **Classification error:**
   For case (b), we have:

```
p(C0) = 6/20 = 0.3
p(C1) = 14/20 = 0.7
```

Therefore, the classification error is:

```
Error(S) = 1 - max(0.3, 0.7) = 0.3
```

b. Second, assume we plan to split the data for case (a) and (b) in the follow ways. Could you calculate the information gain for each? Do you suggest a split based on these, and why?

a. C0:1 and C1:9 + C0:7 and C1:3

Total instances = 20

1. **Entropy of the parent node:**
   The entropy of the parent node is the same as that calculated in the previous question for case (a), i.e., `0.971`.

2. **Entropy of the child nodes:**
   To calculate the entropy of the child nodes, we first need to calculate the weighted average of the entropies based on the proportion of instances in each child node.

For the left child node (C0:1 and C1:9), we have:

```
Total instances = 10

p(C0) = 1/10 = 0.1
p(C1) = 9/10 = 0.9
```

Entropy of the left child node:

```
H(left) = -[(0.1 * log2(0.1)) + (0.9 * log2(0.9))] = 0.468
```

For the right child node (C0:7 and C1:3), we have:

```
Total instances = 10

p(C0) = 7/10 = 0.7
p(C1) = 3/10 = 0.3
```

Entropy of the right child node:

```
H(right) = -[(0.7 * log2(0.7)) + (0.3 * log2(0.3))] = 0.881
```

Weighted average entropy of the child nodes:

```
H(child) = [(10/20) * H(left)] + [(10/20) * H(right)] = 0.674
```

1. **Information gain:**
   Information gain is defined as the difference between the entropy of the parent node and the weighted average entropy of the child nodes.

Information gain for the given split is:

```
IG(S, C0:1 and C1:9 + C0:7 and C1:3) = H(S) - H(child) = 0.971 - 0.674 = 0.297
```

b. C0:3 and C1:3 + C0:3 and C1:11

Total instances = 20

1. **Entropy of the parent node:**
   The entropy of the parent node is the same as that calculated in the previous question for case (b), i.e., 0.881.

2. **Entropy of the child nodes:**
   For the left child node (C0:3 and C1:3), we have:

   ```
   Total instances = 6
   p(C0) = 0.5
   p(C1) = 0.5
   ```

Entropy of the left child node:

```
H(left) = -[(0.5 * log2(0.5)) + (0.5 * log2(0.5))] = 1
```

For the right child node (C0:3 and C1:11), we have:

```
Total instances = 14
p(C0) = 3/14 = 0.214
p(C1) = 11/14 = 0.786
```

Entropy of the right child node:

```
H(right) = -[(0.214 * log2(0.214)) + (0.786 * log2(0.786))] = 0.749
```

Weighted average entropy of the child nodes:

```
H(child) = [(6/20) * H(left)] + [(14/20) * H(right)] = 0.824
```

1. Information gain:
   Information gain for the given split is:

```
IG(S, C0:3 and C1:3 + C0:3 and C1:11) = H(S) - H(child) = 0.881 - 0.824 = 0.057
```