Problem 1.

(a) As the data set has half positive classes and half negative classes, the initial entropy E(S) = 1.

(1) Consider X1, it has

|  |  |
| --- | --- |
| Four 1s | Three +1 |
| One -1 |
| Six 0s | Two +1 |
| Four -1 |

Thus, E(X1) = 4/10(E1) + 6/10(E0)

= 4/10 \* (-3/4 \* log2(3/4) – 1/4 \* log2(1/4)) + 6/10 \* (-2/6 \* log2(2/6) – 4/6 \* log2(4/6))

= 0.8754887502163469

Gain(S, X1) = 1 - 0.8754887502163469 = 0.12451124978365313

(2) Consider X2, it has

|  |  |
| --- | --- |
| Three 1s | Three +1 |
|  |
| Seven 0s | Two +1 |
| Five -1 |

Thus, E(X2) = 3/10(E1) + 7/10(E0)

= 0(pure) + 7/10 \* (-2/7 \* log2(2/7) – 5/7 \* log2(5/7))

= 0.6041843979966417

Gain(S, X2) = 1 - 0.6041843979966417= 0.3958156020033583

(3) Consider X3, it has

|  |  |
| --- | --- |
| Six 1s | Three +1 |
| Three -1 |
| Four 0s | Two +1 |
| Two -1 |

Thus, E(X3) = 6/10(E1) + 4/10(E0)

= 6/10 \* 1 + 4/10 \* 1 # both half-half

= 1

Gain(S, X3) = 1 - 1= 0

(4) Consider X4, it has

|  |  |
| --- | --- |
| Four 1s | Two +1 |
| Two -1 |
| Six 0s | Three +1 |
| Three -1 |

Thus, E(X4) = 4/10(E1) + 6/10(E0)

= 4/10 \* 1 + 6/10 \* 1 # both half-half

= 1

Gain(S, X4) = 1 - 1= 0

Therefore, X2 has the highest information gain, we first split on X2.

X2

/ \

1 0

**Within X2(1),**

The data set has all positive classes, so the initial entropy E(X2(1)) = 0.

(1) Consider X1, it has

|  |  |
| --- | --- |
| Two 1s | Two +1 |
|  |
| One 0s | One +1 |
|  |

Thus, E(X1) = 0 # two pure sets

Gain(X2, X1) = 0 – 0 = 0

(2) Same thing goes for X3 and X4, all pure sets, information gain is 0.

**Within X2(0),**

The data set has 2 positive classes and 5 negative classes, so

E(X2(0)) = -(2/7)\* log2(2/7)-(5/7)\*log2(5/7) = 0.863120568566631

(1) Consider X1, it has

|  |  |
| --- | --- |
| Two 1s | One +1 |
| One -1 |
| Five 0s | One +1 |
| Four -1 |

Thus, E(X2, X1) = 0.8013772106338303

Gain(X2, X1) = 0.061743357932800724

(1) Consider X3, it has

|  |  |
| --- | --- |
| Four 1s | One +1 |
| Three -1 |
| Three 0s | One +1 |
| Two -1 |

Thus, E(X2, X1) = 0.6

Gain(X2, X1) = 0.4390359525563188