Naman Jain Vimal Kumar
1001670153

Problem 1)

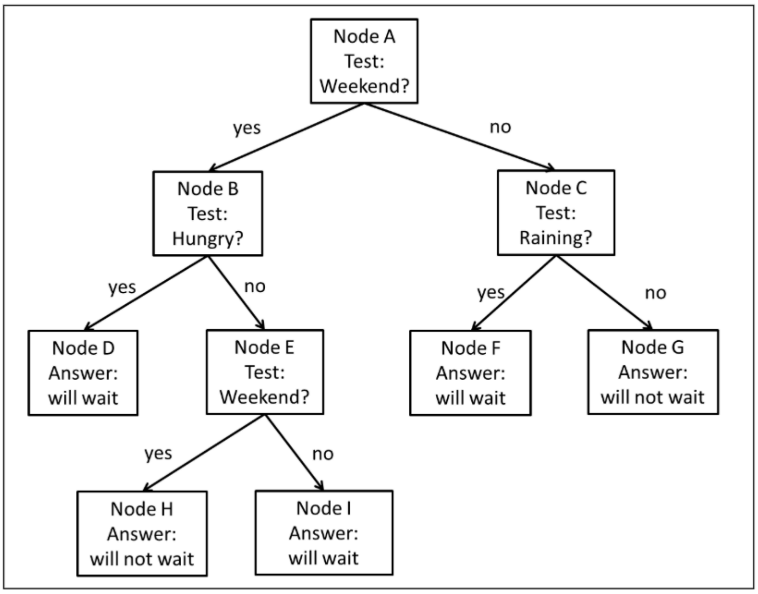


Figure 1 A decision tree for estimating whether the patron will be willing to wait for a table at a restaurant

A. *Suppose that, on the entire set of training samples available for constructing the decision tree of Figure 1, 80 people decided to wait, and 20 people decided not to wait. What is the initial entropy at node A (before the test is applied)?*

Let
$K_1 - Number\ of\ people\ decided\ to\ wait = 80$
$K_2 - Number\ of\ people\ decided\ not\ to\ wait = 20$
$K = K_1 + K_2 = (80 + 20) = 100$

We know that, entropy is given by:

$$H\left(\frac{K_1}{K}, \frac{K_2}{K}\right) = \sum_{i=1}^{N} -\frac{K_i}{K}\left(\log_2 \frac{K_i}{K}\right)$$

Computing initial entropy,

$$\mathrm{H(A)} = \left(-\frac{80}{100} * \log_2 \frac{80}{100}\right) + \left(-\frac{20}{100} * \log_2 \frac{20}{100}\right)$$

$$\mathrm{H(A)} = (-0.8 * \log_2 0.8) + (-0.2 * \log_2 0.2)$$

$$\mathrm{H(A)} = (-0.8 * -0.3219) + (-0.2 * -2.3219)$$

$$\mathrm{H(A)} = 0.25752 + 0.46438$$

$$\mathbf{H(A) = 0.7219}$$

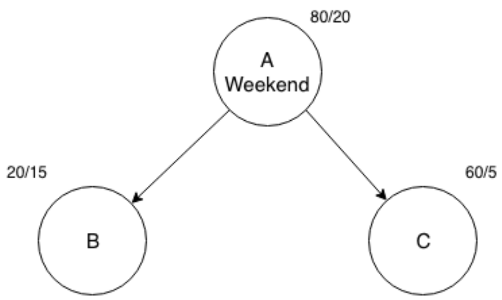
B. *As mentioned in the previous part, at node A 80 people decided to wait, and 20 people decided not to wait.*
- ***Out of the cases where people decided to wait, in 20 cases it was weekend and in 60 cases it was not weekend.***
- ***Out of the cases where people decided not to wait, in 15 cases it was weekend and in 5 cases it was not weekend.***

What is the information gain for the weekend test at node A?

Information Gain is given by,

$$I(N,L) = H(E) - \sum_{i=1}^{L} \frac{K_i}{K} H(E_i)$$

From the graph,



Computing required terms,

$\boldsymbol{H(A) = 0.7219}$

$H(B) = \left(-\frac{20}{35}\log_2\frac{20}{35}\right) + \left(-\frac{15}{35} * \log_2\frac{15}{35}\right)$

$H(B) = (-0.5714 * \log_2 5714) + (-0.4286 * \log_2 0.4286)$

$H(B) = (-0.5714 * \ -0.8073) + (-0.4286 * -1.2223)$

$H(B) = 0.4613 + 0.5239$

$\boldsymbol{H(B) = 0.9852}$

$H(C) = \left(-\frac{60}{65}\log_2\frac{60}{65}\right) + \left(-\frac{5}{65} * \log_2\frac{5}{65}\right)$

$H(C) = (-0.5714 * \log_2 0.5714) + (-0.0769 * \log_2 0.0769)$

$H(C) = (-0.5714 * 0.1154) + (-0.0769 * \ -3.7009)$

$H(C) = 0.1065 + 0.2846$

$\boldsymbol{H(C) = 0.3911}$

Therefore,

$$I(E,L) = \ 0.7219 - \left(\frac{35}{100} * 0.9852 + \frac{65}{100} * 0.3911\right)$$

$$I(E,L) = 0.7219 - 0.34482 - 0.254215$$

$$\boldsymbol{I(E,L) = \ 0.122865}$$

C. ***In the decision tree of Figure 1, node E uses the exact same test (whether it is weekend or not) as node A. What is the information gain, at node E, of using the weekend test?***

Information Gain will be **0** as there are no participating records for Node I, that is when there is no weekend.

D. We have a test case of a hungry patron who came in on a rainy Tuesday. Which leaf node does this test case end up in? What does the decision tree output for that case?

Current Node	***Question***	***Answer***	***Next Node***
A	*Weekend?*	*No*	*C*
C	*Raining?*	*Yes*	*F*

Output: The patron will wait.

E. We have a test case of a not hungry patron who came in on a sunny Saturday. Which leaf node does this test case end up in? What does the decision tree output for that case?

Current Node	***Question***	***Answer***	***Next Node***
A	*Weekend?*	*Yes*	*B*
B	*Hungry?*	*No*	*E*
E	*Weekend?*	*Yes*	*H*

Output: The patron will not wait.

Problem 2)

Class	A	B	C
X	1	2	1
X	2	1	2
X	3	2	2
X	1	3	3
X	1	2	2
Y	2	1	1
Y	3	1	1
Y	2	2	2
Y	3	3	1
Y	2	1	1

We want to build a decision tree that determines whether a certain pattern is of type X or type Y. The decision tree can only use tests that are based on attributes A, B, and C. Each attribute has 3 possible values: 1, 2, 3 (we do not apply any thresholding). We have the 10 training examples, shown on the table (each row corresponds to a training example). What is the information gain of each attribute at the root? Which attribute achieves the highest information gain at the root?

$$H(E) = \left(-\frac{5}{10} * \log_2 \frac{5}{10}\right) + \left(-\frac{5}{10} * \log_2 \frac{5}{10}\right)$$
$$H(E) = 2 * (-0.5 * -1\ )$$

$$\boldsymbol{H(E) = 1}$$

Calculating Information Gain for attribute 'A'

$$H(E_1) = (-\frac{3}{3}\log_2\frac{3}{3}) + (-\frac{0}{3}\log_2\frac{0}{3}) = \mathbf{0}$$

$$H(E_2) = (-\frac{1}{4}\log_2\frac{1}{4}) + (-\frac{3}{4}\log_2\frac{3}{4}) = 0.5 + 0.3113 = \mathbf{0.8113}$$

$$H(E_3) = (-\frac{1}{3}\log_2\frac{1}{3}) + (-\frac{2}{3}\log_2\frac{2}{3}) = 0.5283 + 0.3899 = \mathbf{0.9182}$$

$$I_A = 1 - \left(\left(\frac{3}{10} * 0\right) + \left(\frac{4}{10} * 0.8113\right) + \left(\frac{3}{10} * 0.9182\right)\right) = 1 - (0 + 0.324 + 0.276)$$

$$\boldsymbol{I_A = 0.4}$$

Calculating Information Gain for attribute 'B'

$$H(E_1) = (-\frac{1}{4}\log_2\frac{1}{4}) + (-\frac{3}{4}\log_2\frac{3}{4}) = 0.5 + 0.3113 = \mathbf{0.8113}$$

$$H(E_2) = (-\frac{3}{4}\log_2\frac{3}{4}) + (-\frac{1}{4}\log_2\frac{1}{4}) = 0.3113 + 0.5 = \mathbf{0.8113}$$

$$H(E_3) = (-\frac{1}{2}\log_2\frac{1}{2}) + (-\frac{1}{2}\log_2\frac{1}{2}) = \mathbf{1}$$

$$I_B = 1 - \left(\left(\frac{4}{10} * 0.8113\right) + \left(\frac{4}{10} * 0.8113\right) + \left(\frac{2}{10} * 1\right)\right)$$

$$= 1 - (0.32452 + 0.32452 + 0.2)$$

$$\boldsymbol{I_B = 0.15}$$

Calculating Information Gain for attribute 'C'

$$H(E_1) = (-\frac{1}{5}\log_2\frac{1}{5}) + (-\frac{4}{5}\log_2\frac{4}{5}) = 0.4644 + 0.2575 = \mathbf{0.7219}$$

$$H(E_2) = (-\frac{3}{4}\log_2\frac{3}{4}) + (-\frac{1}{4}\log_2\frac{1}{4}) = 0.3113 + 0.5 = \mathbf{0.8113}$$

$$H(E_3) = (-\frac{1}{1}\log_2\frac{1}{1}) + (-\frac{0}{1}\log_2\frac{0}{1}) = \mathbf{0}$$

$$I_C = 1 - \left(\left(\frac{5}{10} * 0.7219\right) + \left(\frac{4}{10} * 0.8113\right) + \left(\frac{1}{10} * 0\right)\right) = 1 - (0.36095 + 0.32452 + 0)$$

$$\boldsymbol{I_C = 0.32}$$

Result: Attribute 'A' achieves the highest information gain.

Problem 3) Suppose that, at a node N of a decision tree, we have 1000 training examples. There are four possible class labels (A, B, C, D) for each of these training examples.

A. ***What is the highest possible and lowest possible entropy value at node N?***

The highest possible value for entropy would be $\log_2(4) = 2$ and the least possible value could be 0.

B. *Suppose that, at node N, we choose an attribute K. What is the highest possible and lowest possible information gain for that attribute?*

The highest possible value for information gain would be $\log_2(4) = 2$ and the least possible value could be 0.

Problem 4) Your boss at a software company gives you a binary classifier (i.e., a classifier with only two possible output values) that predicts, for any basketball game, whether the home team will win or not. This classifier has a 28% accuracy, and your boss assigns you the task of improving that classifier, so that you get an accuracy that is better than 60%. How do you achieve that task? Can you guarantee achieving better than 60% accuracy?

We can negate the output. This will give us an accuracy of 72%. Thus we will be able to achieve accuracy of more than 60% too.

Problem 5) Consider the Training set for a Pattern Classification problem given below

Attribute 1	Attribute 2	Class
15	28	A
20	10	B
18	32	A
32	15	B
25	15	B

Assuming we want to build a Pseudo-Bayes classifier for this problem using one dimensional gaussians (with naive-bayes assumption) to approximate the required probabilities. Calculate the probability density functions required.

Formulae used:

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

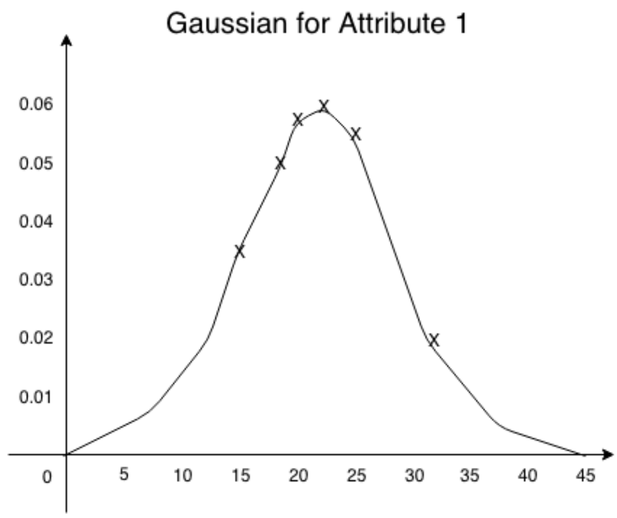
$$\sigma = \sqrt{\frac{1}{N-1} * \sum_{i=1}^{n} (x_i - \mu)^2}$$

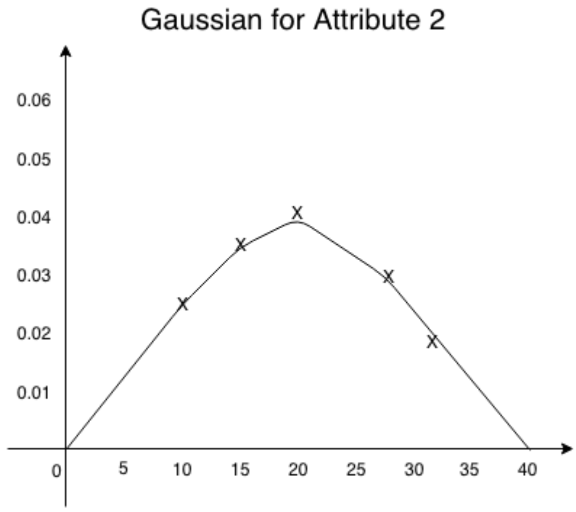
$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Naman Jain Vimal Kumar
1001670153

Class	**A1**	**A2**	$(A1-\mu_1)^2$	$(A2-\mu_2)^2$	**N(A1)**	**N(A2)**
A	15	28	49	64	0.03448451	0.02949308
B	20	10	4	100	0.05717566	0.02411992
A	18	32	16	144	0.04996371	0.01886349
B	32	15	100	25	0.0194428	0.0366727
B	25	15	9	25	0.05405211	0.0366727
Sum	110	100	178	358		
	22	**20**	**6.67083203**	**9.46044396**	0.05980398	0.04216951
	μ_1	μ_2	σ_1	σ_2	$N(\mu_1)$	$N(\mu_2)$
	Mean(μ)		**Variance(σ)**		**Normalized Values**	

Note: Excel is attached for these values.





From the above table:

Class	SUM(Attribute 1)	SUM(Attribute 2)	Total
A	0.08444822	0.04835657	0.13280479
B	0.13067057	0.09746532	0.22813589
Total	0.21511879	0.14582189	0.36094068

To find: P(A / A1, A2) and P(B / A1, A2)

P(A) = 0.13280479 / 0.36094068 = 0.3679
P(B) = 0.22813589 / 0.36094068 = 0.6321

Using Naïve – Bayes Assumption that is ‘A’ and ‘B’ are independent
P(A,B / C) = P(A / C) * P(B / C)

Therefore,
P(A1, A2 / A) = P(A1 / A) * P(A2 / A)

$$= \frac{0.08444822}{0.13280479} * \frac{0.04835657}{0.13280479} = 0.6359 * 0.3641 = 0.2315$$

P(A1, A2 / B) = P(A1 / B) * P(A2 / B)

$$= \frac{0.13067057}{0.22813589} * \frac{0.09746532}{0.22813589} = 0.5728 * 0.4272 = 0.2447$$

P(A1, A2) $= \frac{0.21511879}{0.36094068} * \frac{0.14582189}{0.36094068} = 0.5960 * 0.4040 = 0.2408$

Finding the required probabilities (using Bayes Rule):

P(A / A1, A2) = (P(A) * P(A1, A2 / A)) / P(A1, A2)

$$= \frac{\mathbf{0.3679 * 0.2315}}{\mathbf{0.2408}} = \mathbf{0.3537}$$

P(B / A1, A2) = (P(B) * P(A1, A2 / B)) / P(A1, A2)

$$= \frac{\mathbf{0.6321 * 0.2447}}{\mathbf{0.2408}} = \mathbf{0.6463}$$

Problem 6) Can you represent the following function as a single neuron. If so design that neuron. If you cannot use a single neuron can you use a network of neurons. If so design that network.

The function must return 1 if 3x - 4y = 15, 0 otherwise

Assume the following transfer function

The transfer function returns 1 if the given value is GREATER THAN OR EQUAL TO 0. 0 otherwise

Assume a bias input of +1.