no question of testing with ex2.

Version Space

Version Space $VS_{H,D}$ is the subset of the hypothesis from H consistent with training example in D.

In simple terms we can say that version space consists of all those hypothesis that are consistent with training examples.

$$VS_{H,D} \equiv \{ h \in H \mid \text{Consistent}(h, D) \}$$

List-Then Eliminate Algorithm

Used to find set of hypothesis ~~from~~ consistent with training example ~~that are consistent~~

1. Version Space ← A list consisting of every hypothesis in H.
2. For each training example, $\langle x, c(x) \rangle$, remove from version space any hypothesis h for which $h(x) \neq c(x)$

Consistent Hypothesis, Version Space
List Then Eliminate Algorithm

A hypothesis h is said to be consistent with the given set of training example iff $h(x) = c(x)$ for each examples in D.

$$\boxed{Consistent(h, D) \equiv (\forall \langle x, c(x)\rangle \in D)\ h(x) = c(x)}$$

Example	Citation	Size	inLibrary	Price	Edition	Buy
1	Some	Small	No	Affordable	One	No
2	Many	Big	No	Expensive	Many	Yes

Target variable

Let us take one hypothesis

$h1 = (?, ?\ No, ?, Many)$ — Consistent

Test whether $h1$ is consistent with all training example.

$h1$ classifies ex1 as negative & ex2 as positive so, $h1$ is consistent.

$h2 = (?, ?, No, ?, ?)$ — Not Consistent

Classifies ex1 as +ve but we except -ve, since its inconsistent with first training example

Candidate Elimination Algorithm

Example	Sky	Airtemp	Humidity	Wind	Water	Forecast	Enjoy Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

6 Attribute

↑ Target variable

S_0 <∅, ∅, ∅, ∅, ∅, ∅>

S_1 : Sunny, Warm, Normal, Strong, Warm, Same

S_2 : S3: Sunny, warm, ~~High~~ ?, Strong, Warm, Same

S_3 : <Sunny, Warm, ?, Strong, Warm, Same>

S_4 : <Sunny, Warm, ? Strong, ?, ?>

G_4 : <Sunny, ????> <?, Warm, ????>

G_3 : <Sunny, ?????> <?, Warm, ????> <?? Normal???> <????Cool?> <????? Same>

G_0 : G_1 : G_2 <?, ?, ?, ?, ?, ?>

S <Sunny, Warm, ? Strong, ?, ?>

<Sunny, Warm, ? ? ? ?>

<Sunny, ?, ?, Strong, ? ?>

<? Warm, ? Strong ? ?>

G <Sunny, ?, ?, ?, ? ?>

<?, Warm, ? ? ? ?>

1. First we set most generic boundary & most specific boundary.

If the example is a true example, we go to generic boundary & then check whether hypothesis at generic boundary is consistent or not, if not we write next general hypothesis.

We go to specific boundary to check if it is consistent, then, we write more general hypothesis

Consider the given dataset, Apply Naive Bayes Algorithm and predict fruit has the following properties which type of fruit it is

Fruit = { Yellow, Sweet, Long }

Frequency Table

Fruit	Yellow	Sweet	Long	Total
Mango	350	450	0	650
Banana	400	300	350	400
Others	50	100	50	150
Total	800	850	400	1200

Soln:-

Naivee Bayes Formula:

$$P(A|B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

P(A) → Probability of A being true

P(A|B) → Probability of 'A' being true given 'B' is true

P(B/A) → Probability of B being true given A is true.

P(B) → Probability of B being true

(1) Assume 'x' is a mango

$P(x \mid mango) = ?$

$$P(yellow \mid mango) = \frac{P(mango \mid yellow) \cdot P(yell}{P(mango)}$$

$$= \frac{\frac{350}{800} * \frac{800}{1200}}{650/1200}$$

$$= \underline{\underline{0.53}}$$

$$P(sweet \mid mango) = \frac{P(mango \mid sweet) \cdot P(sweet)}{P(mango)}$$

$$= \frac{450/850 * 850/1200}{650/1200}$$

$$= 0.69$$

$$P(long \mid mango) = \frac{P(mango \mid long) \cdot P(long)}{P(mango)}$$

$$= \frac{(0/400) * (400/1200)}{(650/1200)}$$

$$= 0$$

So, $P(x \mid mango) = 0.53 * 0.69 * 0 = \underline{\underline{0}}$

(2) Similarly, Assume 'x' is banana)

$$P(yellow \mid banana) = \frac{P(banana \mid yellow) \cdot P(yellow)}{P(banana)}$$

$$= \frac{400/800 \times 800/1200}{(400/1200)}$$

$$= 1$$

Apply

P(

P(

~~Apply Naive~~

$P(\text{sweet} \mid \text{banana}) = 0.75$

$P(\text{long} \mid \text{banana}) = 0.875$

$P(x \mid \text{banana}) = 1 \times 0.75 \times 0.875$

$= \underline{\underline{0.65}}$

(3) Assume 'x' is others

$P(\text{yellow} \mid \text{others}) = 0.33$

$P(\text{sweet} \mid \text{others}) = 0.66$

$P(\text{long} \mid \text{others}) = 0.33$

$P(x \mid \text{others}) = 0.33 \times 0.66 \times 0.33 = \underline{\underline{0.072}}$.

Among, all the three cases, $P(x \mid \text{banana}) = 0.65$ is large, so we can conclude that, x belongs to

Banana

Find-S Algorithm

1. Initialize H to more specific Hypothesis in H.
2. For each true instance in x.
   For each attribute Constraint a_i.
   - If Constraint is a_i in h is satisfied by x do nothing.
     else replace a_i by h by next more general constraint satisfied by x.
3. Output hypothesis h.

Example	Color	Toughness	Fungus	Appearance	Poison
1.	Green	Hard	No	Wrinkled	Yes
2.	Green	Hard	Yes	Smooth	No
3.	Brown	Soft	No	Wrinkled	No
4.	Orange	Hard	No	Wrinkled	Yes
5.	Green	Soft	Yes	Smooth	Yes
6.	Green	Hard	Yes	Wrinkled	Yes
7.	Orange	Hard	No	Wrinkled	Yes

$h = \{ \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \}$

$h_1 = \{$ Green, Hard, No, Wrinkled $\}$

$h_2 = \{$ Green, Hard, No, Wrinkled $\}$ $h_2 = h_1$

$h_3 = \{$ Green, Hard, No, Wrinkled $\}$ $h_3 = h_2$

h_4 = { ?, Hard, No, Wrinkled }

h_5 = { ?, ?, ?, ? }

h_6 = h_7 = h_5

Final hypothesis : h = { ?, ?, ?, ? }

Decision Tree

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Head Ache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	Yes	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

$$Entropy(S) = -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S) = \underset{(3)}{Strep\ Throat} + \underset{(3)}{Allergy} + \underset{(4)}{Cold}$$

$$= -\left[\frac{3}{10}\log_2\frac{3}{10} + \frac{3}{10}\log_2\frac{3}{10} + \frac{4}{10}\log_2\frac{4}{10}\right]$$

$\times 2$

$$Gain(S) = 0.6(1.73) + 0.4(1.318) = \underline{\underline{1.562}}$$

Finding the best splitting attribute

Sorethroat → Yes (5), No (5)

$$Info(Yes) = -\left[\frac{2}{5}\log_2\frac{2}{5} + \frac{1}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5}\right]$$

$$= 1.52$$

$$Info(No) = -\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5}\right]$$

$$= \underline{1.52}$$

$$Entropy(Sorethroat) = P(yes) * 1.52 + P(No) * 1.52$$

$$= \frac{5}{10} * 1.52 + \frac{5}{10} * 1.52$$

$$= 1.52$$

S_3 = { Green, Hard, No, Wrink..., ...

Gain (Sorethroat) = Info (S) - Entropy (Sorethroat)

= 1.562 - 1.52 = 0.05

Gain (Fever) → Yes, No)

Info (Yes) → $-\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{2}{5}\right]$ = ?

Info (No) → $-\left[-\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right]$ = ?

= 0.72

Similarly Gain (Swollen glands) = 0.88

Gain (Congestion) = 0.45

Gain (Headache = 0.05