Decision Tree & Naive Bayes Questions

A Decision Tree Question

Here is a table which records some data about whether a student will go out to play. Use decision tree to analysis the following questions:

(1) Which attribute you will choose as root node among outlook, temperature, humidity and windy?

(2) Write your analysis process **in question (1)**.

(3) Draw the decision tree.

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Entropy Review

- Entropy is a measure of the uncertainty of a random variable; acquisition of information corresponds to a reduction in entropy.

$$H(V) = -\sum_k P(v_k) log_2 P(v_k)$$

- the entropy of a fair coin flip :

$$H(Fair) = -0.5 * log_2^{0.5} \quad - 0.5 * log_2^{0.5} = 1$$

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

No:5

Yes:9

$$H(Play) = -\frac{9}{14} * log_2^{\frac{9}{14}} - \frac{5}{14} * log_2^{\frac{5}{14}} = 0.940$$

Information Gain

- $H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x)$
- $\mathrm{IG(Y|X)} = H(Y) - H(Y|X)$

	YES	No	total
sunny	2	3	5
overcast	4	0	4
rain	3	2	5
total	9	5	14

outlook=sunny

outlook=overcast

outlook=rain

$$H(play|outlook) = \boxed{\frac{5}{14} * \left(-\frac{2}{5} * log_2^{\frac{2}{5}} - \frac{3}{5} * log_2^{\frac{3}{5}}\right)} + \boxed{\frac{4}{14} * \left(-\frac{4}{4} * log_2^{\frac{4}{4}} - 0 * log_2^{0}\right)} + \boxed{\frac{5}{14} * \left(-\frac{3}{5} * log_2^{\frac{3}{5}} - \frac{2}{5} * log_2^{\frac{2}{5}}\right)}$$

A Navie Bayes Question

Consider a dataset shown below, the task is to predict whether a person is ill. There are four boolean features ‘running nose’, ‘coughing’, ‘reddened skin’ and ‘fever’.

(1) Determine all the (estimated) probabilities required by the naive Bayes classifier for pre- dicting whether a person is ill or not. (assuming the prior is uniform)

(2) Verify whether the naive Bayes classifier classifies training examples $x^{(2)}, x^{(4)}, x^{(6)}$ correctly. Please show your calculation.

(3) Apply your naive Bayes classifier to new examples $x^{(7)} = \langle \bar{N}, C, \bar{R}, F \rangle, x^{(8)} = \langle N, \bar{C}, \bar{R}, F \rangle$ *and* $x^{(9)} = \langle N, \bar{C}, R, \bar{F} \rangle$. Notation $\bar{C}$ means F (not coughing).

Training example	N (running nose)	C (coughing)	R (reddened skin)	F (fever)	Ill
$\boldsymbol{x}^{(1)}$	T	T	T	F	T
$\boldsymbol{x}^{(2)}$	T	T	F	F	T
$\boldsymbol{x}^{(3)}$	F	F	T	T	T
$\boldsymbol{x}^{(4)}$	T	F	F	F	F
$\boldsymbol{x}^{(5)}$	F	F	F	F	F
$\boldsymbol{x}^{(6)}$	F	T	T	F	F

NBC Question Answer(1)

- Here the class label y ∈ {ill,healthy}.
- According to Bayes' theorem and the attribute conditional independence assumption of naive Bayes classifier, we have

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)} = \frac{P(y)}{P(x)}\prod_{i=1}^{d} P(x_i|y)$$

where d = 4 is the dimension of x.

The naive Bayes classifier is:

$$h(x) = \underset{y}{\operatorname{argmax}}\, P(y)\prod_{i=1}^{d} P(x_i|y)$$

Training example	N (running nose)	C (coughing)	R (reddened skin)	F (fever)	Ill	
$\boldsymbol{x}^{(1)}$	T	T	T	F	T	
$\boldsymbol{x}^{(2)}$	T $\frac{2}{3}$	T $\frac{2}{3}$	F $\frac{2}{3}$	F $\frac{1}{3}$	T	ill
$\boldsymbol{x}^{(3)}$	F	F	T	T	T	
$\boldsymbol{x}^{(4)}$	T	F	F	F	F	
$\boldsymbol{x}^{(5)}$	F $\frac{1}{3}$	F $\frac{1}{3}$	F $\frac{1}{3}$	F $\frac{0}{3}$	F	healthy
$\boldsymbol{x}^{(6)}$	F	T	T	F	F	

NBC Question Answer(1)

- In this question the class prior is assumed to be uniform, $P\,(ill) \;=\; P\,(healthy) \;=\; \frac{1}{2}$

$$P(N\,|\,ill) \;= \frac{2}{3}, P(N\,|\,healthy) \;= \frac{1}{3}$$
$$P(C\,|\,ill) \;= \frac{2}{3}, P(C\,|\,healthy) \;= \frac{1}{3}$$
$$P(R\,|\,ill) \;= \frac{2}{3}, P(R\,|\,healthy) \;= \frac{1}{3}$$
$$P(F\,|\,ill) \;\; = \frac{1}{3}, P(F\;|\;\; healthy) \;= \frac{0}{3} = 0$$

- Note that P(F | healthy) = 0 because of the very limited data samples, we'd better do Laplacian correction to avoid 0 term in the multiplication.

$$P(N\,|\,ill) \;= \frac{2+1}{3+2*1} = \frac{3}{5}, P(N\,|\,healthy) \;= \frac{1+1}{3+2*1} = \frac{2}{5}$$
$$P(C\,|\,ill) \;= \frac{3}{5}, P(C\,|\,healthy) \;= \frac{2}{5}$$
$$P(R\,|\,ill) \;= \frac{3}{5}, P(R\,|\,healthy) \;= \frac{2}{5}$$
$$P(F\,|\,ill) \;= \frac{2}{5}, P(F\,|\,healthy) \;= \frac{1}{5}$$

Laplacian correction (or Laplacian smoothing)

Laplace correction is a smoothing technique that handles the problem of zero probability in Naïve Bayes.

Using Laplace correction, we can represent P(w'|positive) as

$$P(w'|positive) = \frac{\text{number of reviews with w' and y = positive} + \alpha}{\text{N} + \alpha * K}$$

Here,

alpha represents the smoothing parameter,

K represents the number of values in the data, and

N represents the number of reviews with y=positive

If we choose a value of alpha!=0 (not equal to 0), the probability will no longer be zero even if a word is not present in the training dataset.

Most of the time, alpha = 1 is being used to remove the problem of zero probability.

NBC Question Answer(2)

- Because the class prior $P(y)$ is uniform, we only need to check $P(x \mid y)$
- $x^{(2)} = \langle N, C, \bar{R}, \bar{F} \rangle$

$$P(N, C, \bar{R}, \bar{F} \mid ill) = P(N \mid ill) * P(C \mid ill) * (1 - P(R \mid ill)) * (1 - P(F \mid ill))$$
$$= \frac{3 \cdot 3 \cdot 2 \cdot 3}{5^4} = \frac{54}{5^4}$$

$$P(N, C, \bar{R}, \bar{F} \mid healthy) = P(N \mid healthy) * P(C \mid healthy) * (1 - P(R \mid healthy)) * (1 - P(F \mid healthy))$$
$$= \frac{2 \cdot 2 \cdot 3 \cdot 4}{5^4} = \frac{48}{5^4}$$

$$P(N \mid ill) = \frac{3}{5}, P(N \mid healthy) = \frac{2}{5}$$
$$P(C \mid ill) = \frac{3}{5}, P(C \mid healthy) = \frac{2}{5}$$
$$P(R \mid ill) = \frac{3}{5}, P(R \mid healthy) = \frac{2}{5}$$
$$P(F \mid ill) = \frac{2}{5}, P(F \mid healthy) = \frac{1}{5}$$

So, $h(x^{(2)}) = ill$, correct.