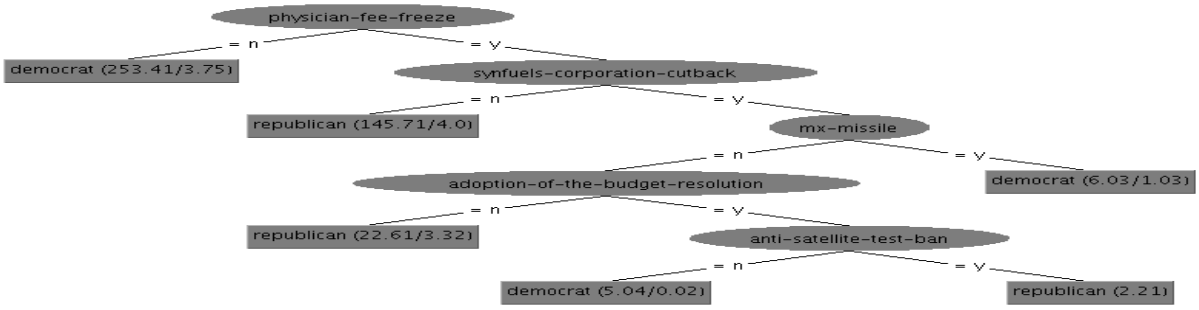
Mid Term machine learning 2016 (B)
GIVEN NAME: **FAMILY NAME:**

Question 1 (6 points): In the following dataset, 'physician-fee-freeze' is the 4th attribute and 'synfuels-corporation-cutback' is the 11th attribute. Which subset of the data **support** the rule:
IF 'physician-fee-freeze'=y **AND** 'synfuels-corporation-cutback'=n ➔ CLASS=Republicans ? What is the **confidence** of this rule (% of instances fitting the left side hand of the rule that also fit the right hand side).

1. 'n','n','y','n','n','y','n','y','y','y','n','n','n','y','n','y','democrat'
2. 'n','n','n','y','y','y','n','n','n','y','n','y','y','y','n',?,'republican'
3. 'n','n','n','y','y','y','n','n','n','y','n','y','y','y','n','y','republican'
4. 'n','n','n','y','y','y','n','n','n','n','n','y','y','y','n',?,'republican'
5. 'n','n','n','y','y','y','n','n','n','y','n','y','y','y','n','n','republican'
6. 'n','y','n','y','y','y','n','n','n','y','y','y','y','n','n','y','republican'
7. 'n',?,'y','n','n','y','y','y','y','y','n','n','n','y','y','y','democrat'
8. 'n','n','y','n','n','y','y','y','y','y','n','n','n','y','n','y','democrat'
9. 'y','n','y','n','n','y','y','y','y','n','n','n','n','n','y','y','democrat'
10. 'n','n','n','y','n','n','y','y','y','y','n','n','y','y','n','y','republican'
11. 'n','n','n','y','y','y','y','y','y','y','n','y','y','y',?,'y','republican'
12. 'n','n','n','y','y','y','y','y','y','y','n','y','y','y','n','y','republican'
13. ?,'y','n','n','n','n','y','y','y','y','y','n','n','y','y','y','democrat'
14. 'n',?,'n','y','n','y','y','y','y','y','n','n','n','y','n',?,'democrat'
15. 'n','n','y','n','n','y','y','y','y','y','n','n','n','y',?,'y','democrat'
16. 'n','y','n','y','y','y','n','n','n','n','y','y','y','y','n','y','republican'



Question 2: Version Space (8 points)
Write the pseudo code of VERSION SPACE algorithm, and then explain the algorithm in your words.
If your writing is real bad, please write in block letters.

Question 3: Evaluation (9 points)
The previous Decision Tree, when tested on a sample S of 435 instances, has an accuracy of 96.32%. Compute the confidence N% that $-0.55\sigma \leq |p - error(h)_S| \leq 0.55\sigma$. Also compute precisely the interval (i.e. compute σ).

Question 4 (7 points)
Given the 16 instances of Question 1, compute the Information Gain of testing on attribute 'physician-fee-freeze'.

Solutions

Question 1. There are 7 instance supporting the rule (both left and right hand) and 1 rule supporting the left side but not the righ side (instance n. 14) The support of the rule is 7/16 (or simply 7) and the confidence 7/8

1. 'n','n','y','n','n','y','n','y','y','y','n','n','n','y','n','y','democrat'
2. 'n','n','n','**y**','y','y','n','n','n','y','**n**','y','y','y','n',?,'republican' √
3. 'n','n','n','**y**','y','y','n','n','n','y','**n**','y','y','y','n','y','republican'√
4. 'n','n','n','**y**','y','y','n','n','n','n','**n**','y','y','y','n',?,'republican'√
5. 'n','n','n','**y**','y','y','n','n','n','y','**n**','y','y','y','n','n','republican'√
6. 'n','y','n','y','y','y','n','n','n','y','y','y','y','n','n','y','republican'
7. 'n',?,'y','n','n','y','y','y','y','y','n','n','n','y','y','y','democrat'
8. 'n','n','y','n','n','y','y','y','y','y','n','n','n','y','n','y','democrat'
9. 'y','n','y','n','n','y','y','y','y','n','n','n','n','n','y','y','democrat'
10. 'n','n','n','**y**','n','n','y','y','y','y','**n**','n','y','y','n','y','republican'√
11. 'n','n','n','**y**','y','y','y','y','y','y','**n**','y','y','y',?,'y','republican'√
12. 'n','n','n','**y**','y','y','y','y','y','y','**n**','y','y','y','n','y','republican'√
13. ?,'y','n','n','n','n','y','y','y','y','y','n','n','n','y','y','y','democrat'
14. 'n',?,'n','**y**','n','y','y','y','y','y','**n**','n','n','y','n',?,'**democrat**' X
15. 'n','n','y','n','n','y','y','y','y','y','n','n','n','n','y',?,'y','democrat'
16. 'n','y','n','y','y','y','n','n','n','n','y','y','y','y','y','n','y','republican'

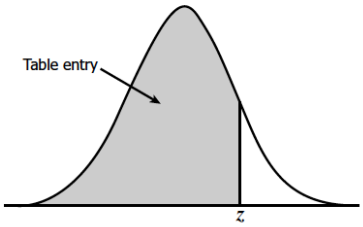
Question 2: Refer to the course slides

Question 3:
The above Decision Tree, when tested on a sample S of 435 instances, has an accuracy of 96.32%. Compute the confidence N% that $-0.55\sigma \leq |p - error(h)_S| \leq 0.55\sigma$. Also compute precisely the interval (i.e. compute σ).

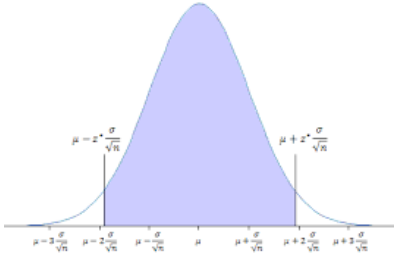
The error rate is 1-0.9632=0.368
We then have:

$$\sigma_S \cong \sqrt{\frac{0.0368(1-0.0368)}{435}} = \sqrt{0.000081} = 0.009$$

Looking on the z-table, we see that the area to the left of z=0.55 is 0.7088 (70,88%) However, this is the entire probability mass to the left of z (the gray area), we need to remove the tail.



Remember that an N% confidence interval is the interval around the mean WITHIN which the area of the Gaussian is N% (the blue area)



Because the gaussian is symmetric, the area to remove is equal to the area to the right of z, which is $(1 - 0.7088) = 0.2912$. Thefore N=0.7088-0.2912=0.4176

We have a confidence 41,76% that the error in estimating the real error rate of the Decision Tree does not exceed $\pm 0.55 \times 0.009 = \pm 0.00495$ or, in other terms , that the real error p lies in the interval $[0.0368 - 0.00495, 0.0368 + 0.00495]$

Note that, since the interval is small, also the confidence is small (only about 40%).

Question 4

In the sample we have 9 republicans and 7 democrats, therefore the initial entropy is

$$E(S) = -\frac{9}{16} log \frac{9}{16} - \frac{7}{16} log \frac{7}{16}$$

If we consider the attribute 'physician-fee-freeze' (PFF), there are 10 instances with PFF=yes and 6 with PFF=no. Let's denote the first population with $S_{PFF=Y}$ and the second with $S_{PFF=N}$. In $S_{PFF=Y}$ 9 instances are classified republicans and 1 is democrat. In $S_{PFF=N}$, all instances are classified democrats.

We then have that the entropy of $S_{PFF=Y}$ is $E(S_{PFF=Y}) = -\frac{9}{10} log \frac{9}{10} - \frac{1}{10} log \frac{1}{10}$ and $E(S_{PFF=N}) = 0$

These two values must be respectively multiplied by the fraction of PFF=yes (10/16) and the fraction of PFF=no (6/16).

The Gain is hence:

$$E(S) - (\frac{10}{16} E(S_{PFF=Y}) - \frac{6}{16} E(S_{PFF=N}) = E(S) - \frac{10}{16} E(S_{PFF=Y})$$