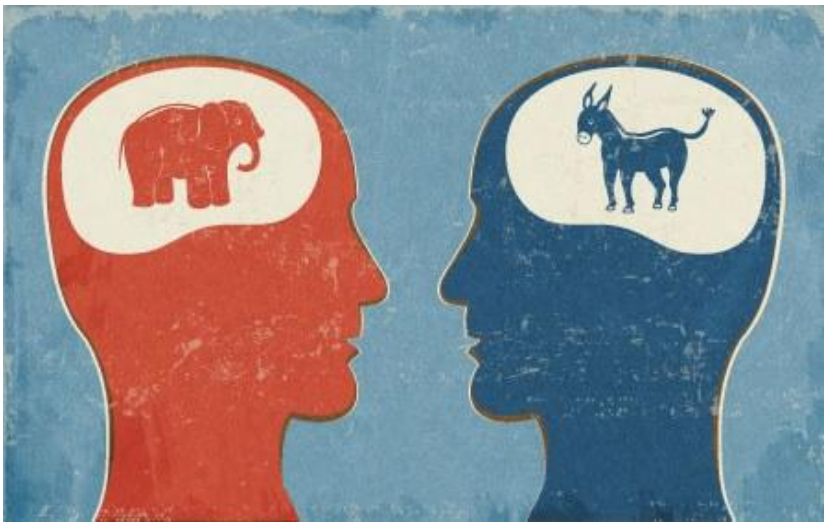
# Project 04: Decision Tree, Naive Bayes, K-means. (Due Thur Dec. 4 11:59PM)



Republican or Democrat?

Image source: www.rantchic.com

## What to submit:

You will implement question1_solver.py to question3_solver.py during the assignment. You should submit these three files along with a report. Your code should be well-documented. All the project submissions must be made through Blackboard.

**Report (10 points)** This should include the description of your algorithm. Specifically, report the output of

```
python learning.py -q1 -q2 -q3.1 -q3.2
```

**Academic Honesty:** We will be checking your code against other submissions in the class for logical redundancy. We trust you all to submit your own work only - own means you and your team member - ; If we find that you copied you will get a F for the course.

**Getting Help:** Feel free to use the Blackboard discussion board to discuss or get clarifications on homework-related issues. You are also encouraged to answer each other's questions on blackboard. Please meet the TA in his office hours if you have questions.

## Introduction

We are trying to classify a congressmen as republican or democrat, using his/her votes as features. The train.data contains all the training data and validation.data contains all the validation data. The data format is shown below:

| Field | Brief |
|---|---|
| 1 | Class Name: 2 (democrat, republican) |
| 2 | handicapped-infants: 2 (y,n) |
| 3 | water-project-cost-sharing: 2 (y,n) |
| 4 | adoption-of-the-budget-resolution: 2 (y,n) |
| 5 | physician-fee-freeze: 2 (y,n) |

6 el-salvador-aid: 2 (y,n)
7 religious-groups-in-schools: 2 (y,n)
8 anti-satellite-test-ban: 2 (y,n)
9 aid-to-nicaraguan-contras: 2 (y,n)
10 mx-missile: 2 (y,n)
11 immigration: 2 (y,n)
12 synfuels-corporation-cutback: 2 (y,n)
13 education-spending: 2 (y,n)
14 superfund-right-to-sue: 2 (y,n)
15 crime: 2 (y,n)
16 duty-free-exports: 2 (y,n)
17 export-administration-act-south-africa: 2 (y,n)

For detailed information of this dataset, please visit: UCI dataset. Note that some attributes (votes, in this case) are missing, which means the values are unkown. Missing attributes are shown as "?" in the data file. You can just treat the "?" as a different value. Therefore, all attributes can take "?", "n", and "?".

## Decision Tree (35 points)

You need to implement the decision tree (ID3) in question1_solver.py. Detailed information are in the comments of the source code. To avoid overfitting, you need to define a threshold, which is the minimum number of instances in a non-leaf node. If the number of instances is below this threshold, stop growing this branch (set the node to be a leaf node) immediately.

You can use the following command to show the accuracy of your algorithm on the validation set:

```
python learning.py -q1
```

## Naive Bayes (35 points)

You need to implement the Naive Bayes classifier in question2_solver.py. Detailed information are in the comments of the source code. Note that you need to smooth $Pr(Xi|Y)$, the probability of an attribute $Xi$ given the class $Y$. For more information, please check the course slides or Here.

You can use the following command to show the accuracy of your algorithm on the validation set:

```
python learning.py -q2
```

## K-means (30 points)

K-means is an unsupervised learning algorithm. You need to implement the K-means algorithm in question3_solver.py.

There are two sets of points to be clustered. You can use the following command to visualize the result:

```
python learning.py -q3.1 -q3.2
```