Naive Bayes

Updating the state of knowledge

step by step

with new information

What time is it? Is it 9 AM?

What time is it? Is it 9 AM?

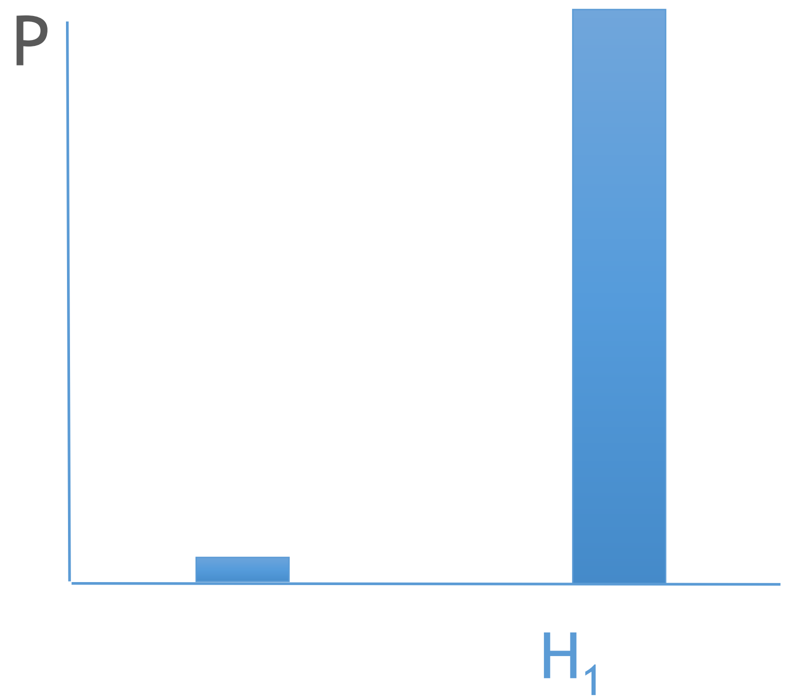
Decide between the two hypotheses (9AM/Not 9AM)
Using all the information we have

4.16% Right Time

What time is it? Is it 9 AM?

H_1 : It is 9 AM (9 to 10)

H_2 : It is not 9 AM



H_2

prior

dist.

$P(H_1) = P($ 9AM $) = 4.16\%$

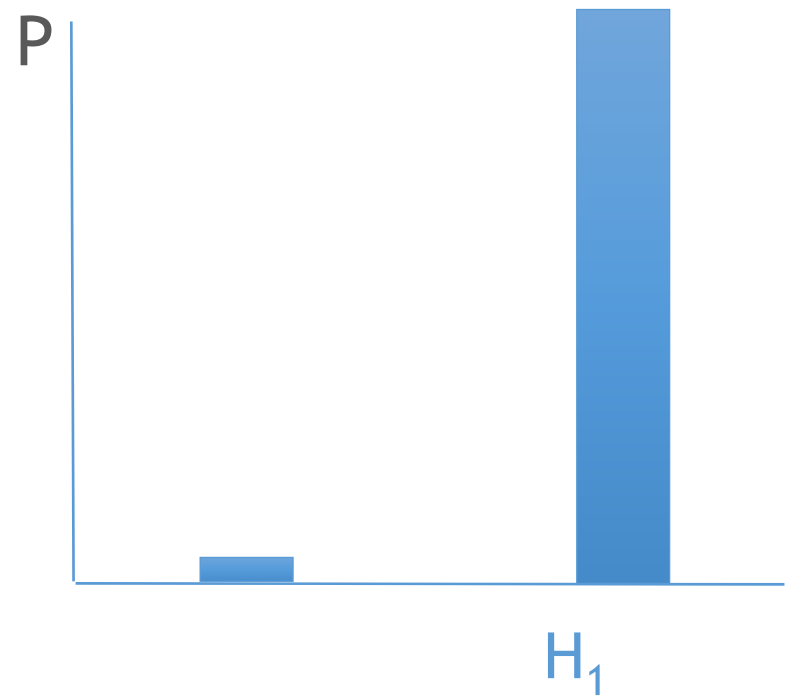
$P(H_2) = P(\text{not } 9AM) = 95.84\%$

4.16%
Right
Time

What time is it? Is it 9 AM?

H_1 : It is 9 AM (9 to 10)

H_2 : It is **not** 9 AM



New information:
5 train is full

H_2

prior

dist.

$P(H_1) = P(\text{9AM}) = 4.16\%$

$P(H_2) = P(\text{not 9AM}) = 95.84\%$

4.16% Right Time

What time is it? Is it 9 AM?

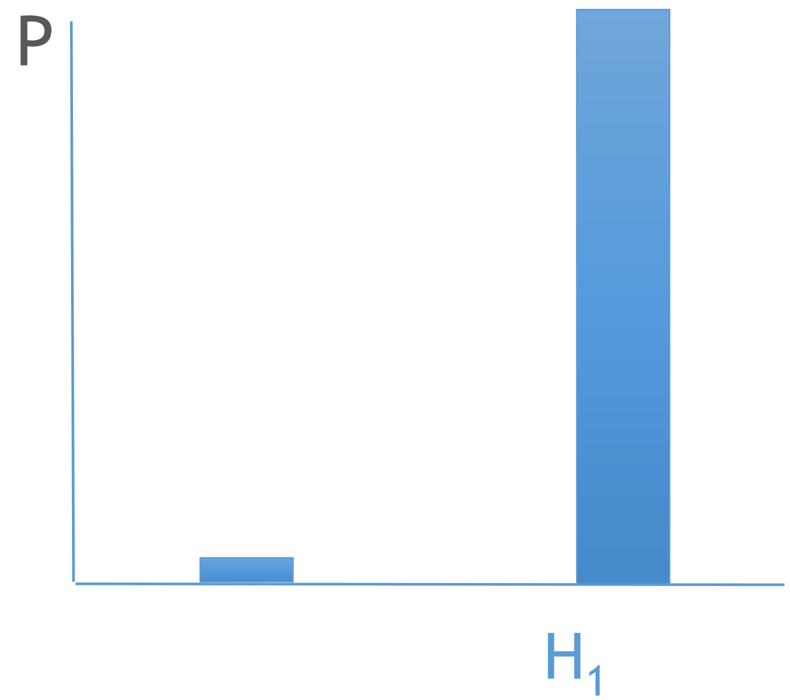
H_1 : It is 9 AM (9 to 10)

H_2 : It is not 9 AM

New information:

5 train is full

$$P(9AM|5tf) = \frac{P(5tf|9AM)\,P(9AM)}{P(5tf)}$$



$P(H_1) = P(\quad 9AM \quad) = 4.16\%$

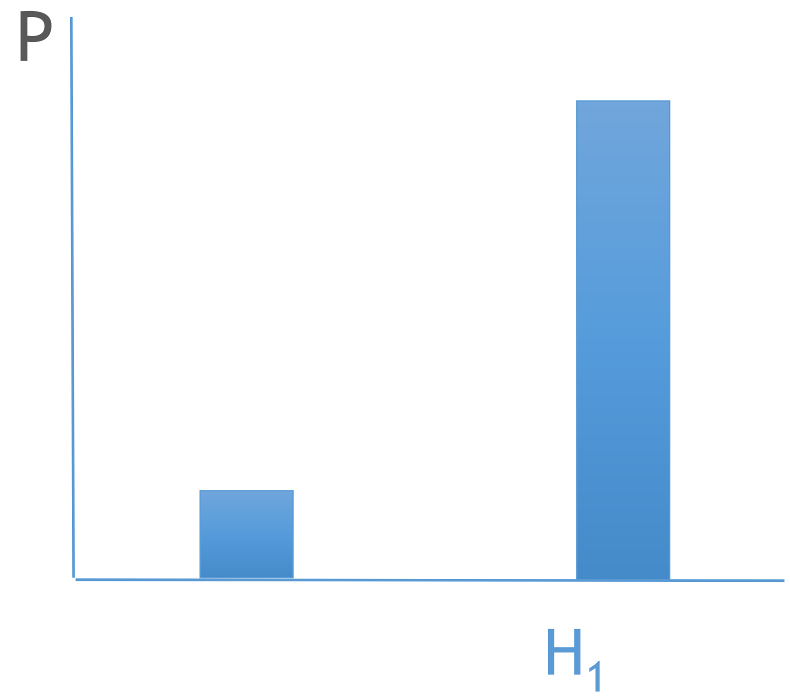
$P(H_2) = P(\text{not } 9AM) = 95.84\%$

13.19%
Right
Time

What time is it? Is it 9 AM?

H_1 : It is 9 AM (9 to 10)

H_2 : It is **not** 9 AM



H_2

posterior

dist.

$P(H_1) = P(\text{ 9AM } | \text{5tf}) = 13.19\%$

$P(H_2) = P(\text{not 9AM} | \text{5tf}) = 86.81\%$

What is classification?

Deciding among hypotheses (labels),
using information we have (features)
for each example

3 Features: Votes on 3 Bills

2 Labels: Democrat / Republican

Prediction:

I know your votes, I'm trying to guess your party

3 Features: Votes on 3 Bills

2 Labels: Democrat / Republican

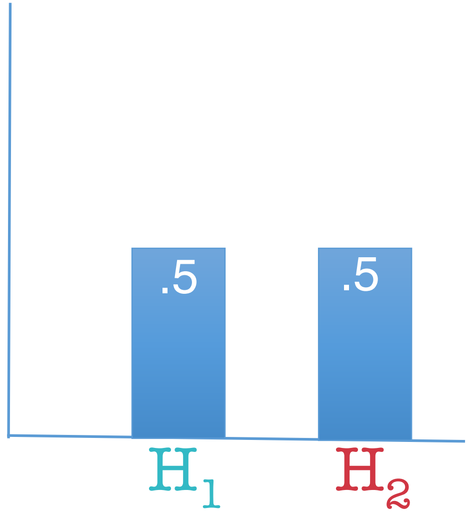
Prediction:

I know your votes, I'm trying to guess your party

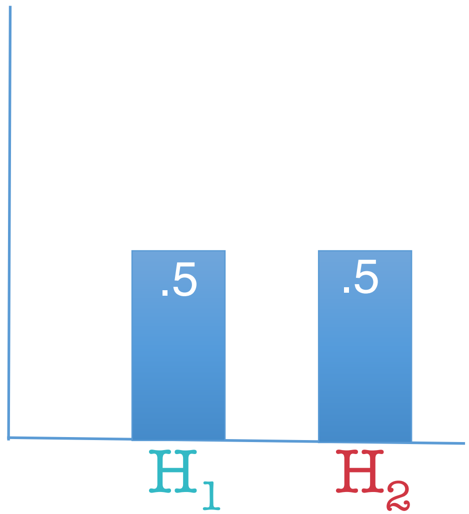
2 Labels

H_1: Democrat

H_2: Republican



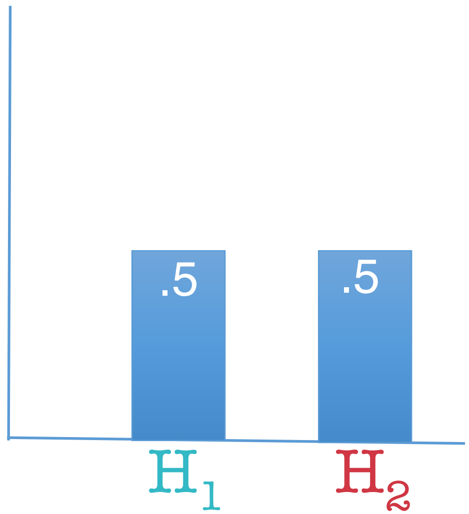
Prior: Initial belief



Prior: Initial belief

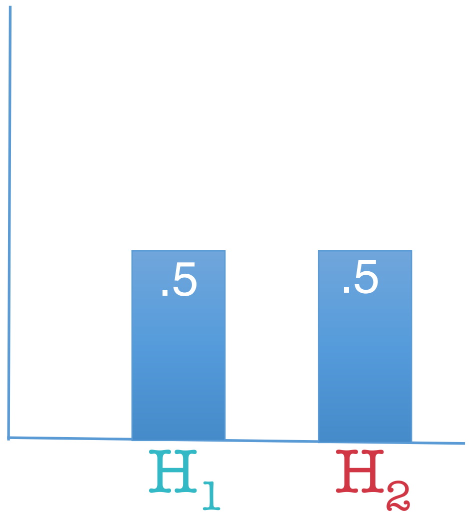
50 - 50 ?   P(Democrat) = 0.5 ?

(Uninformative prior)



Prior: Initial belief

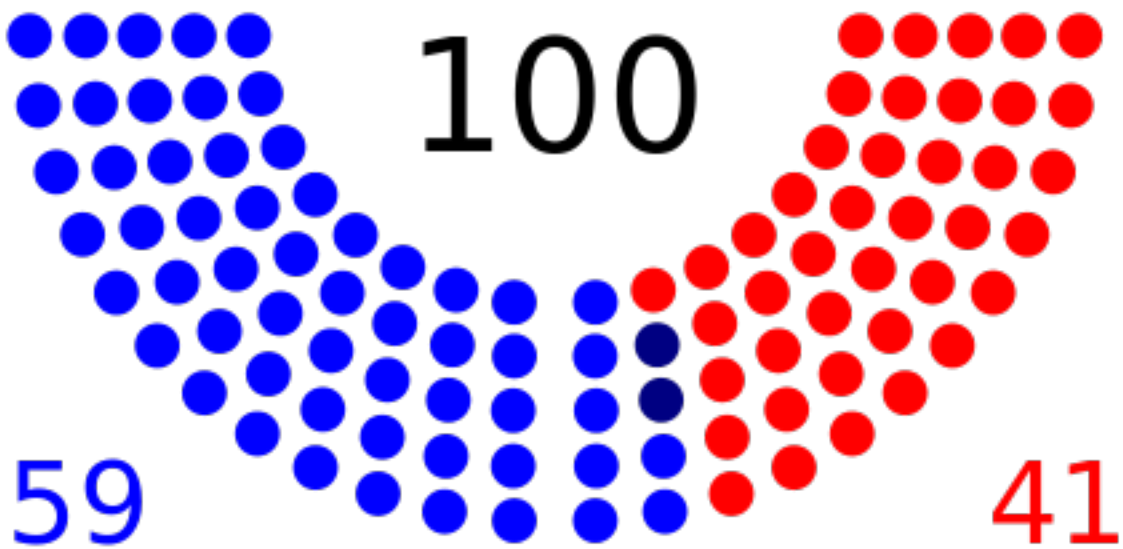
What's my best guess without any vote info?

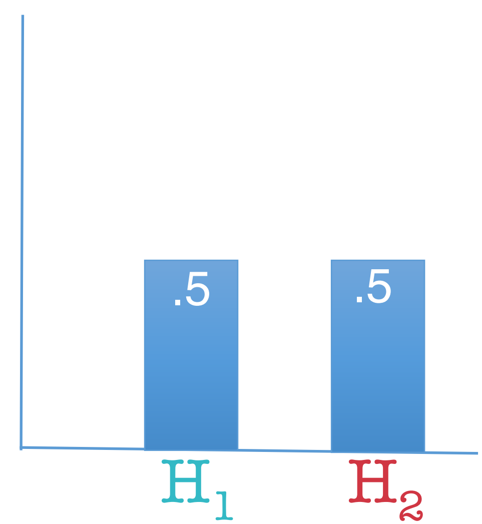


Prior: Initial belief

What's my best guess without any vote info?

I'd guess democrat since there are more of them.



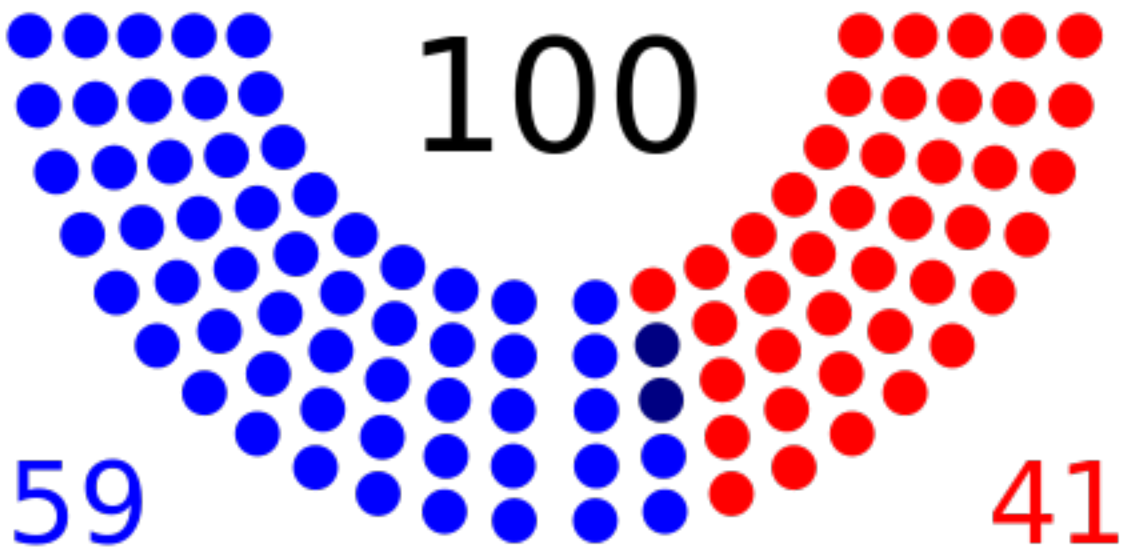


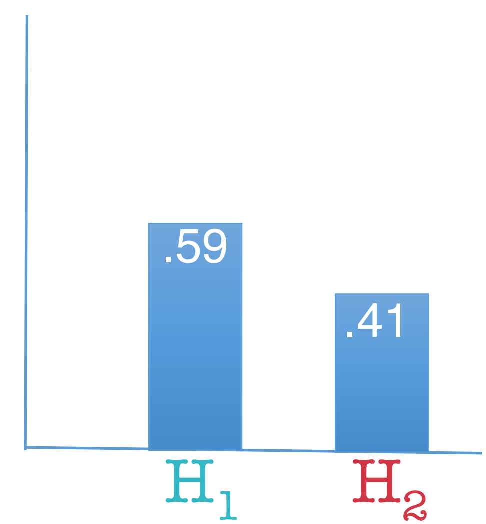
Prior: Initial belief

What's my best guess without any vote info?

I'd guess democrat since there are more of them.

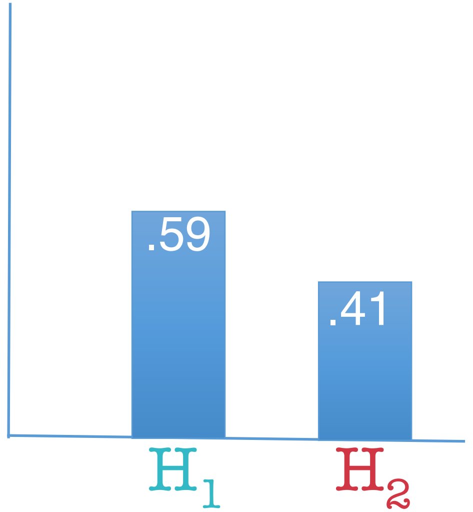
P(Democrat) = 0.59





Prior: Initial belief

P(Democrat) = 0.59

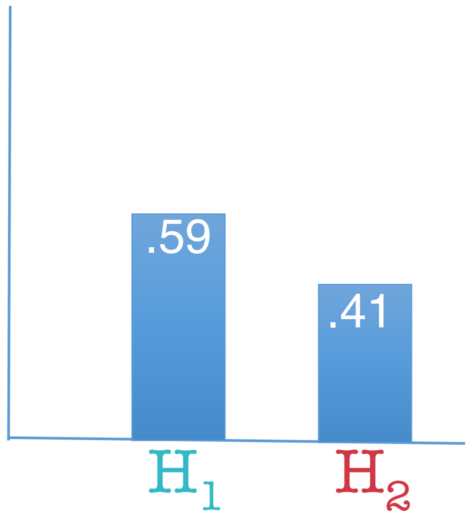


Prior: Initial belief

P(Democrat) = 0.59

New information (feature 1):

Voted YES on Net Neutrality



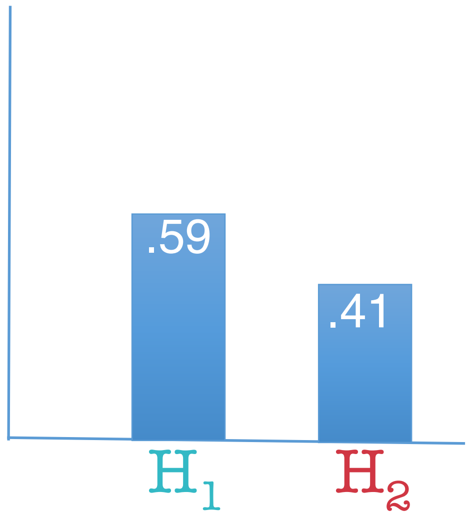
Prior: Initial belief

P(Democrat) = 0.59

New information (feature 1):

Voted YES on Net Neutrality

$$P(Dem|Y_{NN}) = \frac{P(Y_{NN}|Dem)\ P(Dem)}{P(Y_{NN})}$$



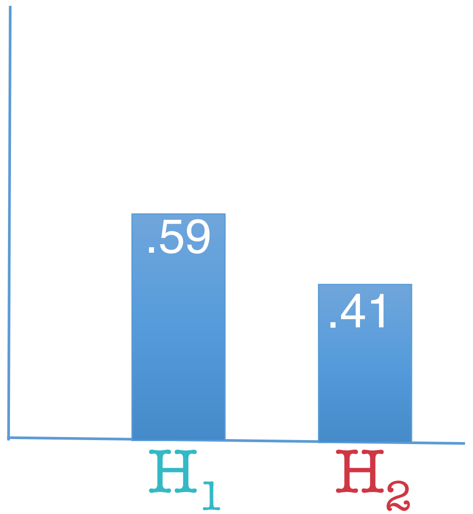
Prior: Initial belief

P(Democrat) = 0.59

New information (feature 1):

Voted YES on Net Neutrality

$$\underbrace{P(Dem|Y_{NN})}_{\text{posterior}} = \frac{\overbrace{P(Y_{NN}|Dem)}^{\text{likelihood}}\ \overbrace{P(Dem)}^{\text{prior}}}{\underbrace{P(Y_{NN})}_{\text{evidence (normalization factor)}}}$$



Prior: Initial belief

P(Democrat) = 0.59

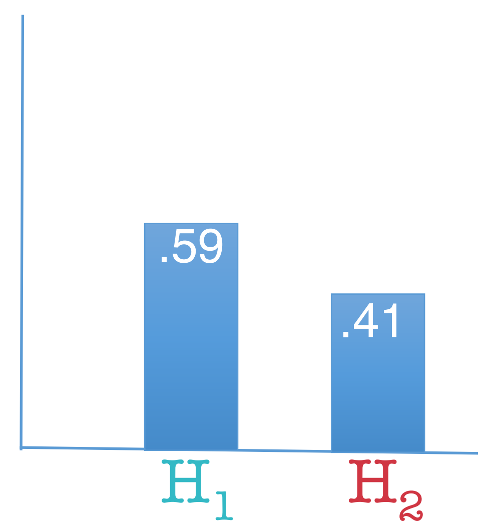
New information (feature 1):

Voted YES on Net Neutrality

$P(Y_{NN}|Dem)$

Prob. of voting yes on net neutrality if you're democrat

$$\underbrace{P(Dem|Y_{NN})}_{\text{posterior}} = \frac{\overbrace{P(Y_{NN}|Dem)}^{\text{likelihood}}\ \overbrace{P(Dem)}^{\text{prior}}}{\underbrace{P(Y_{NN})}_{\substack{\text{evidence} \\ \text{(normalization factor)}}}}$$



Prior: Initial belief

P(Democrat) = 0.59

New information (feature 1):

Voted YES on Net Neutrality

$P(Y_{NN}|Dem)$

Prob. of voting yes on net neutrality if you're democrat

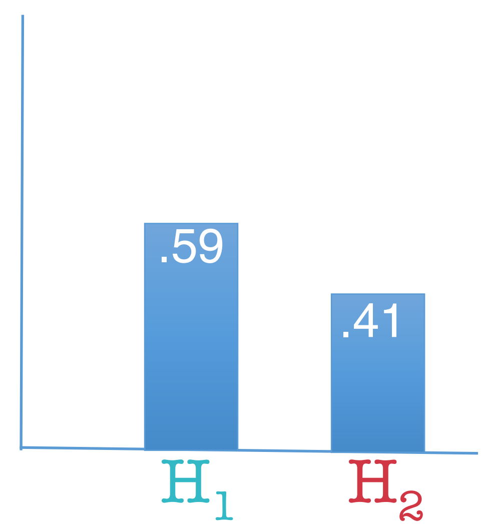
likelihood prior

posterior

$$P(Dem|Y_{NN}) = \frac{P(Y_{NN}|Dem)\,P(Dem)}{P(Y_{NN})}$$

evidence (normalization factor)

$$P(Rep|Y_{NN}) = \frac{P(Y_{NN}|Rep)\,P(Rep)}{P(Y_{NN})}$$



Prior: Initial belief

P(Democrat) = 0.59

New information (feature 1):

Voted YES on Net Neutrality

$P(Y_{NN}|Dem)$

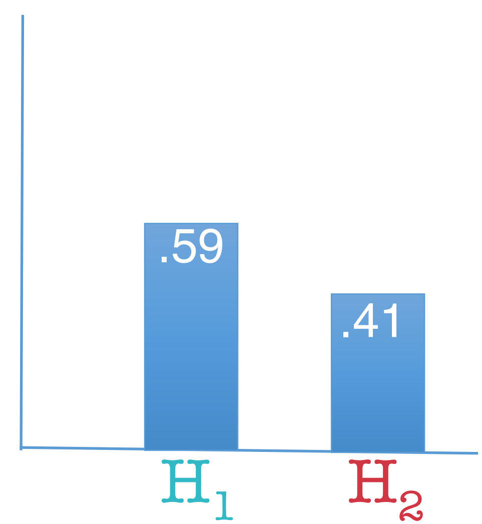
Prob. of voting yes on net neutrality if you're democrat

$$P(Dem|Y_{NN}) = \frac{P(Y_{NN}|Dem)\,P(Dem)}{P(Y_{NN})}$$

posterior, likelihood, prior, evidence (normalization factor)

$P(Y_{NN}|Rep)$

$$P(Rep|Y_{NN}) = \frac{P(Y_{NN}|Rep)\,P(Rep)}{P(Y_{NN})}$$

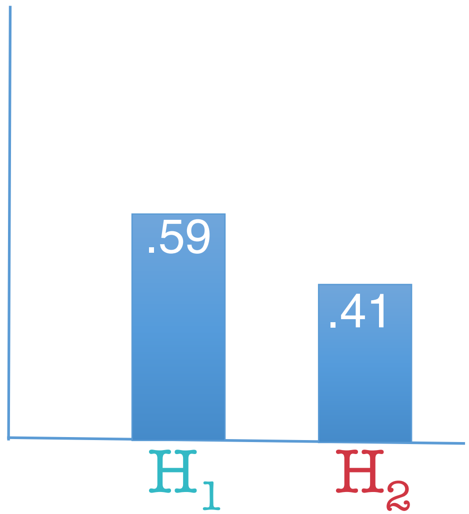


$P(Y_{NN}|Dem)$

Prob. of voting yes
on net neutrality
if you're democrat

$P(Y_{NN}|Rep)$

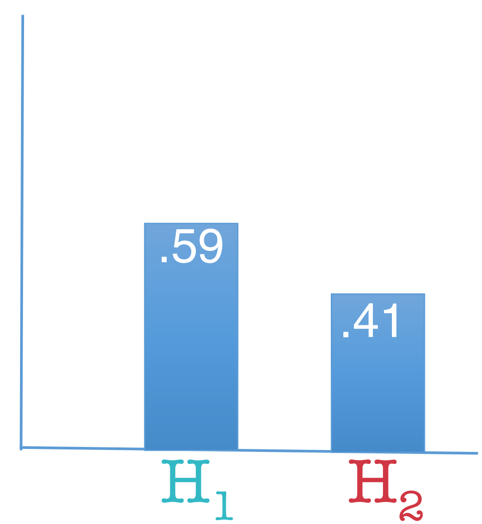
Prob. of voting
yes
on net neutrality
if you're
republican



Training set has the answers!
We know Dem/Rep for each person, we know their votes!

$P(Y_{NN} | Dem)$
Prob. of voting yes on net neutrality if you're democrat

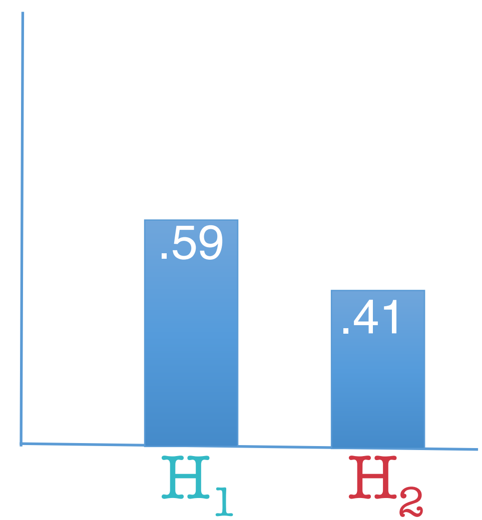
$P(Y_{NN} | Rep)$
Prob. of voting yes on net neutrality if you're republican



Training set has the answers!
We know Dem/Rep for each person, we know their votes!

$$P(Y_{NN} \mid Dem) \approx \frac{\text{\# democrats that } Y_{NN}}{\text{\# all democrats}}$$

$$P(Y_{NN} \mid Rep) \approx \frac{\text{\# republicans that } Y_{NN}}{\text{\# all republicans}}$$

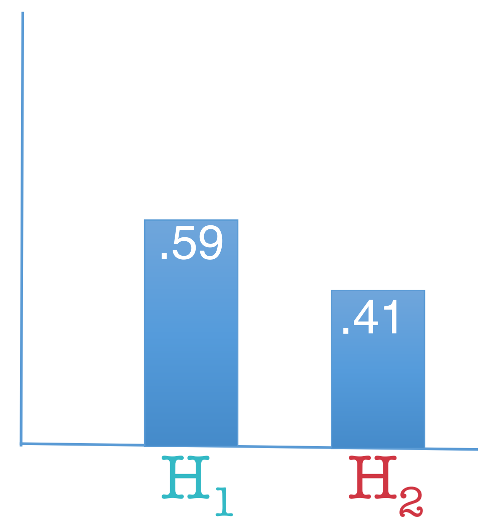


Training set has the answers!
We know Dem/Rep for each person, we know their votes!

$$P(Y_{NN}|Dem) \approx \frac{\#\text{ democrats that } Y_{NN}}{\#\text{ all democrats}}$$

$$P(Y_{NN}|Rep) \approx \frac{\#\text{ republicans that } Y_{NN}}{\#\text{ all republicans}}$$

For likelihoods of discrete data,
training/fitting means counting!
(and estimating likelihoods by dividing counts)

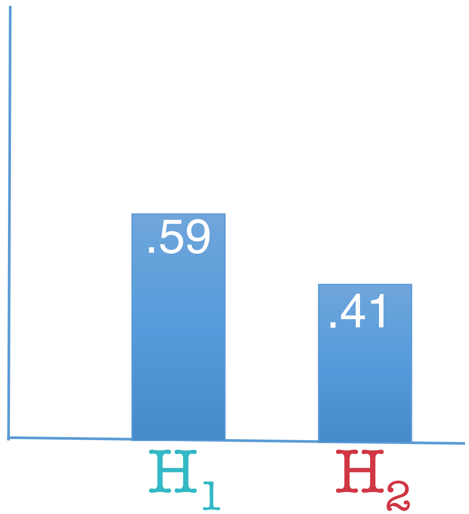


Training set has the answers!
We know Dem/Rep for each person, we know their votes!

$$P(Y_{NN} | Dem) \approx \frac{56}{59} = 0.949$$

$$P(Y_{NN} | Rep) \approx \frac{34}{41} = 0.829$$

For likelihoods of discrete data,
training/fitting means counting!
(and estimating likelihoods by dividing counts)



Prior: Initial belief

P(Democrat) = 0.59

New information (feature 1):

Voted YES on Net Neutrality

likelihood prior

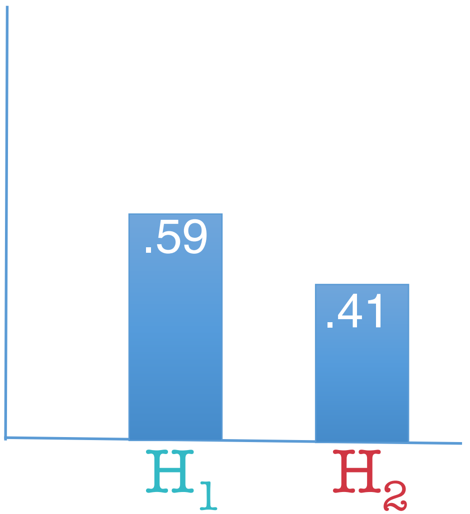
posterior

$$P(Dem|Y_{NN}) = \frac{P(Y_{NN}|Dem)\,P(Dem)}{P(Y_{NN})}$$

evidence

(normalization factor)

$$P(Rep|Y_{NN}) = \frac{P(Y_{NN}|Rep)\,P(Rep)}{P(Y_{NN})}$$



Prior: Initial belief

$P(\text{Democrat}) = 0.59$

New information (feature 1):

Voted YES on Net Neutrality

likelihood prior

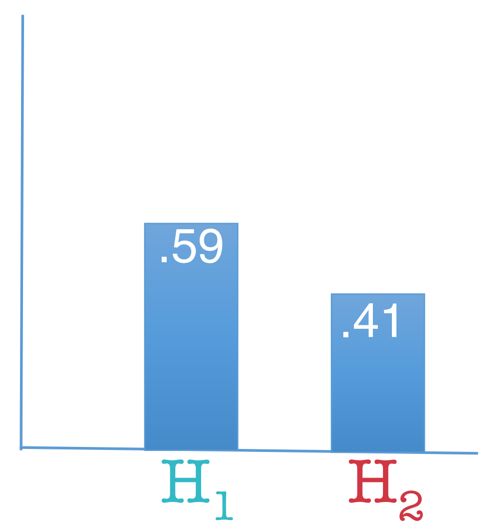
posterior

$$P(Dem|Y_{NN}) = \frac{0.949 * 0.59}{P(Y_{NN})}$$

evidence

(normalization factor)

$$P(Rep|Y_{NN}) = \frac{0.829 * 0.41}{P(Y_{NN})}$$



Current belief

$P(\text{Democrat} | Y_{NN}) = 0.62$

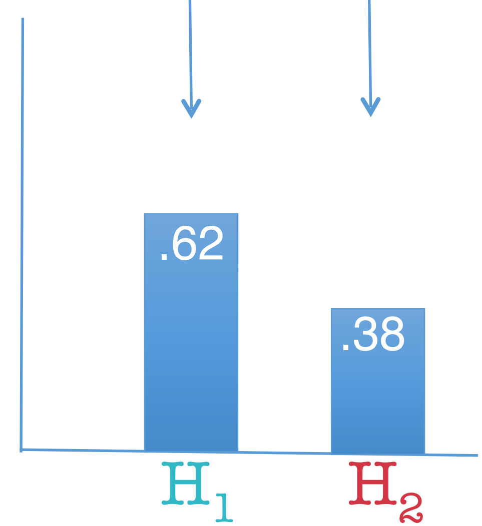
New information (feature 1):

Voted YES on Net Neutrality

$$\underbrace{P(\text{Dem}|Y_{NN})}_{\text{posterior}} = \frac{\overbrace{0.949}^{\text{likelihood}} * \overbrace{0.59}^{\text{prior}}}{\underbrace{P(Y_{NN})}_{\text{evidence (normalization factor)}}}$$

$$P(\text{Rep}|Y_{NN}) = \frac{0.829 * 0.41}{P(Y_{NN})}$$

New posterior distribution

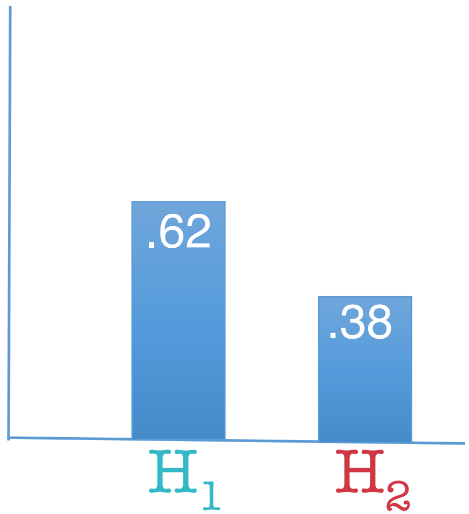


Current belief

$P(\text{Democrat} \mid Y_{NN}) = 0.62$

New information (feature 2):

Voted YES on Tax Cuts



Current belief

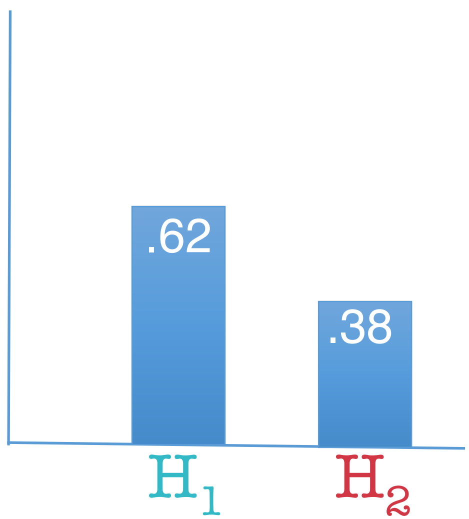
$P(\text{Democrat} | Y_{NN}) = 0.62$

New information (feature 2):

Voted YES on Tax Cuts

$$P(\text{Dem}|Y_{NN}, Y_{TC}) = \frac{P(Y_{TC}|\text{Dem})\, P(\text{Dem}|Y_{NN})}{P(Y_{TC})}$$

$$P(\text{Rep}|Y_{NN}, Y_{TC}) = \frac{P(Y_{TC}|\text{Rep})\, P(\text{Rep}|Y_{NN})}{P(Y_{TC})}$$



Current belief

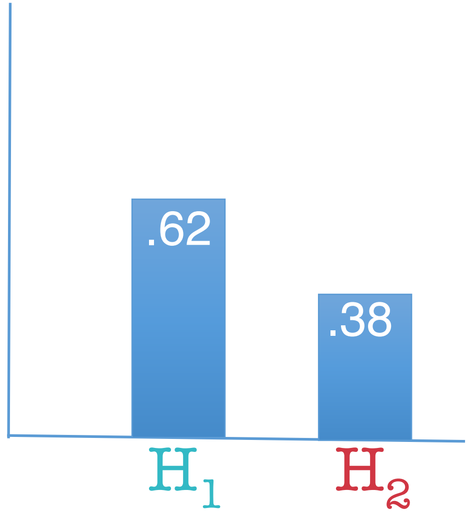
$P(\text{Democrat} \mid Y_{NN}) = 0.62$

$P(Y_{TC} \mid \text{Dem})$

$$\approx \frac{10}{59} = 0.169$$

$P(Y_{TC} \mid \text{Rep})$

$$\approx \frac{35}{41} = 0.854$$



Current belief

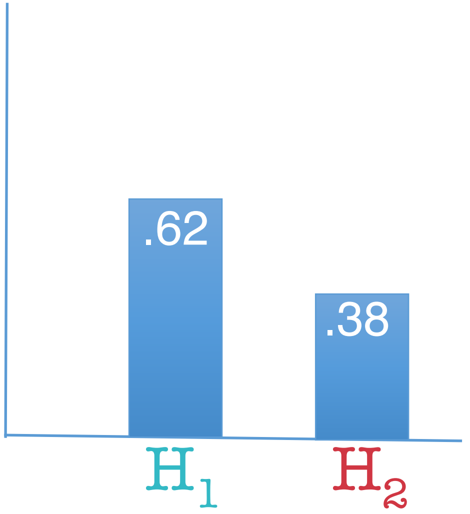
$P(\text{Democrat} \mid Y_{NN}) = 0.62$

New information (feature 2):

Voted YES on Tax Cuts

$$P(\text{Dem} \mid Y_{NN}, Y_{TC}) = \frac{P(Y_{TC} \mid \text{Dem})\, P(\text{Dem} \mid Y_{NN})}{P(Y_{TC})}$$

$$P(\text{Rep} \mid Y_{NN}, Y_{TC}) = \frac{P(Y_{TC} \mid \text{Rep})\, P(\text{Rep} \mid Y_{NN})}{P(Y_{TC})}$$



Current belief

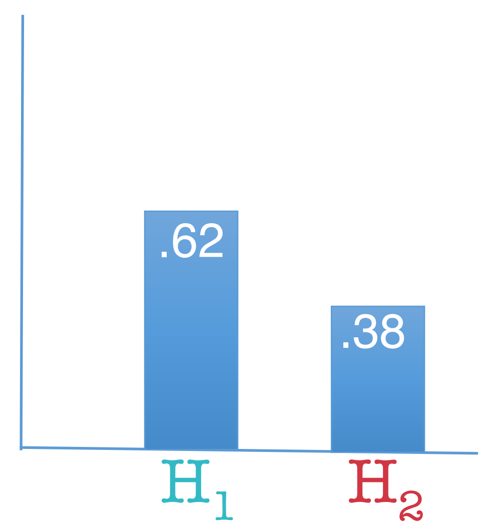
$P(\text{Democrat} \mid Y_{NN}) = 0.62$

New information (feature 2):

Voted YES on Tax Cuts

$$P(\text{Dem} \mid Y_{NN}, Y_{TC}) = \frac{0.169 \quad * \quad 0.62}{P(Y_{TC})}$$

$$P(\text{Rep} \mid Y_{NN}, Y_{TC}) = \frac{0.854 \quad * \quad 0.38}{P(Y_{TC})}$$



Current belief

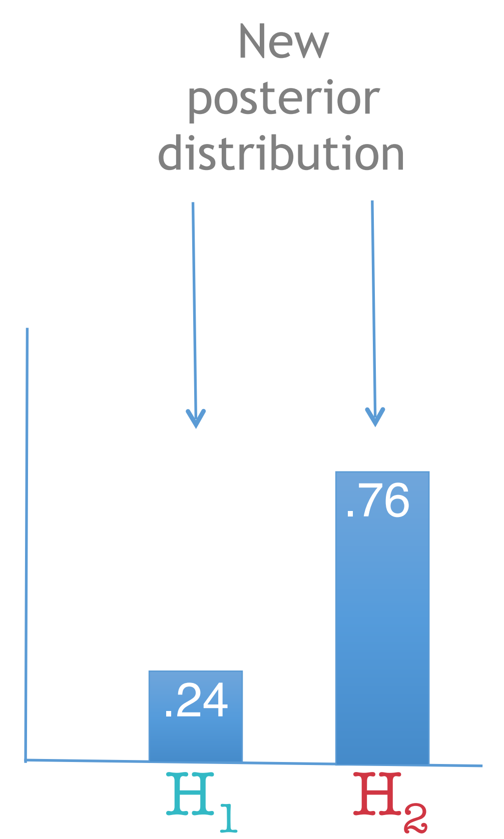
$P(\text{Democrat} \mid Y_{NN}, Y_{TC}) = 0.24$

New information (feature 2):

Voted YES on Tax Cuts

$$P(\text{Dem} \mid Y_{NN}, Y_{TC}) = \frac{0.169 \quad * \quad 0.62}{P(Y_{TC})}$$

$$P(\text{Rep} \mid Y_{NN}, Y_{TC}) = \frac{0.854 \quad * \quad 0.38}{P(Y_{TC})}$$

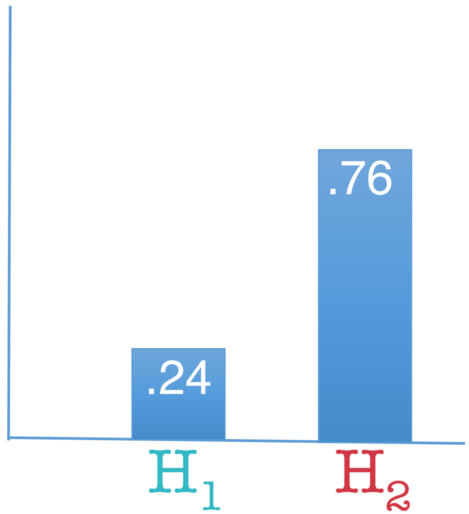


Current belief

P(Democrat | Y_{NN}, Y_{TC}) = 0.24

New information (feature 3):

Voted NO on License-free Guns



Current belief

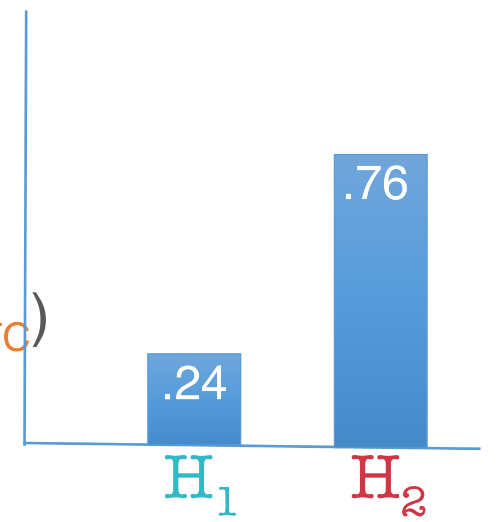
$P(\text{Democrat} \mid Y_{NN}, Y_{TC}) = 0.24$

New information (feature 3):

Voted NO on License-free Guns

$$P(\text{Dem} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{P(N_{LG} \mid \text{Dem})\, P(\text{Dem} \mid Y_{NN}, Y_{TC})}{P(N_{LG})}$$

$$P(\text{Rep} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{P(N_{LG} \mid \text{Rep})\, P(\text{Rep} \mid Y_{NN}, Y_{TC})}{P(N_{LG})}$$



Current belief

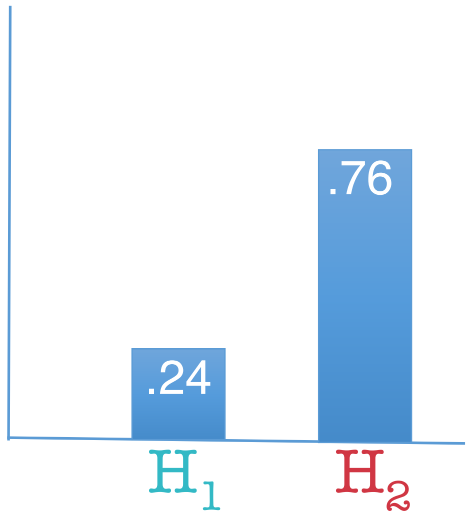
$P(\text{Democrat} \mid Y_{NN}, Y_{TC}) = 0.24$

$P(N_{LG} \mid \text{Dem})$

$$\approx \frac{53}{59} = 0.898$$

$P(N_{LG} \mid \text{Rep})$

$$\approx \frac{23}{41} = 0.561$$



Current belief

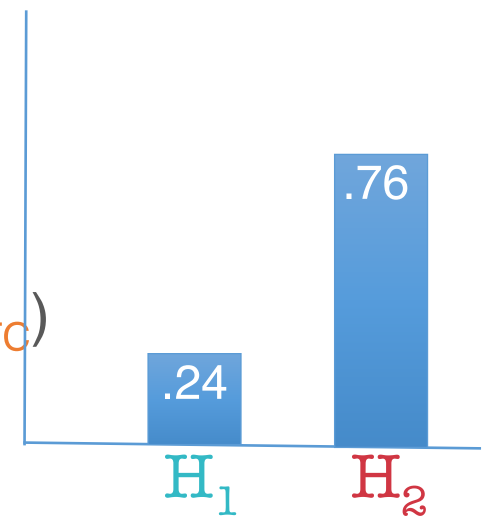
$P(\text{Democrat} \mid Y_{NN}, Y_{TC}) = 0.24$

New information (feature 3):

Voted NO on License-free Guns

$$P(\text{Dem} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{P(N_{LG} \mid \text{Dem})\, P(\text{Dem} \mid Y_{NN}, Y_{TC})}{P(N_{LG})}$$

$$P(\text{Rep} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{P(N_{LG} \mid \text{Rep})\, P(\text{Rep} \mid Y_{NN}, Y_{TC})}{P(N_{LG})}$$



Current belief

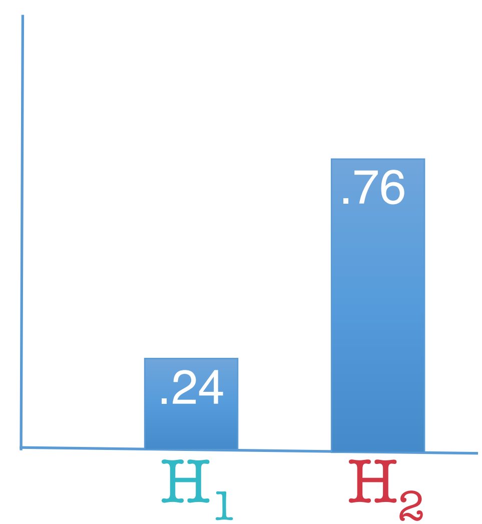
$$P(\text{Democrat} \mid Y_{NN}, Y_{TC}) = 0.24$$

New information (feature 3):

Voted NO on License-free Guns

$$P(\text{Dem} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{0.898 \quad * \quad 0.24}{P(N_{LG})}$$

$$P(\text{Rep} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{0.561 \quad * \quad 0.76}{P(N_{LG})}$$



Current belief

$P(\text{Democrat} \mid Y_{NN}, Y_{TC}, N_{LG}) = 0.34$

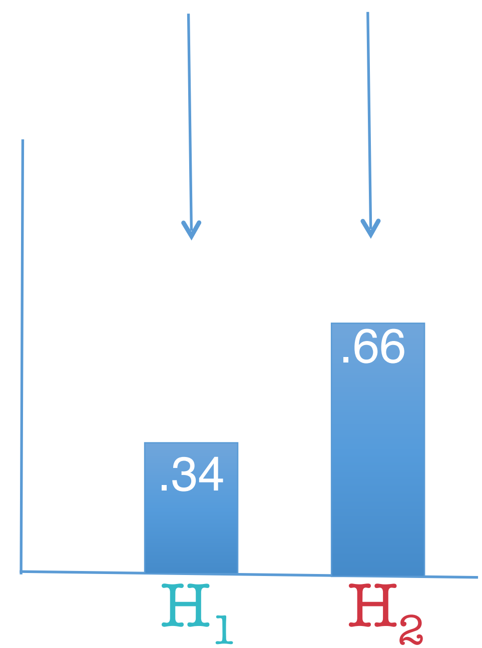
New information (feature 3):

Voted NO on License-free Guns

$$P(\text{Dem} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{0.898 \quad * \quad 0.24}{P(N_{LG})}$$

$$P(\text{Rep} \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{0.561 \quad * \quad 0.76}{P(N_{LG})}$$

New posterior distribution



Current belief

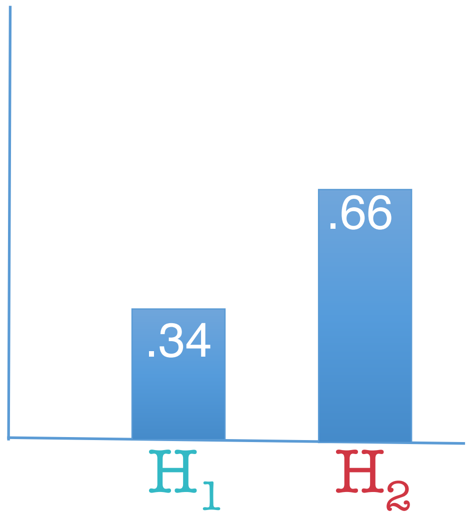
$P(\text{Democrat} | Y_{NN}, Y_{TC}, N_{LG}) = 0.34$

Classify this person that voted

Yes on Net Neutrality (Y_{NN}),

Yes on Tax Cuts (Y_{TC}),

No on License-free Guns (N_{LG})

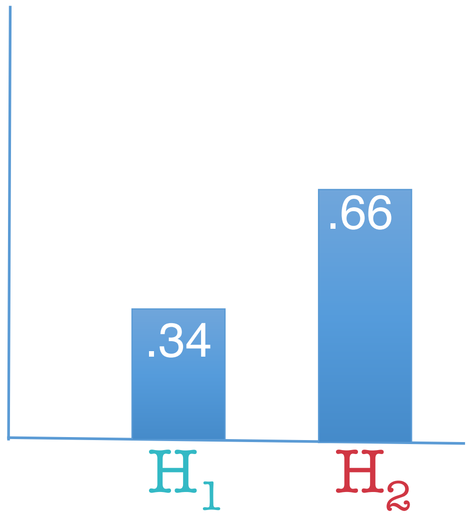


Current belief

$P(\text{Democrat} \mid Y_{NN}, Y_{TC}, N_{LG}) = 0.34$

Classify this person that voted
Yes on Net Neutrality (Y_{NN}),
Yes on Tax Cuts (Y_{TC}),
No on License-free Guns (N_{LG})

My strongest belief is in H_2,
I classify this person with the label Republican.



Naïve Bayes

Training:

Count and calculate the likelihood of each feature value for each class:

$P(Y_{NN}|Dem) \quad = 1 - P(N_{NN}|Dem)$

$P(Y_{NN}|Rep) \quad = 1 - P(N_{NN}|Rep)$

$P(Y_{TC}|Dem) \quad = 1 - P(N_{TC}|Dem)$

$P(Y_{TC}|Rep) \quad = 1 - P(N_{TC}|Rep)$

$P(Y_{LG}|Dem) \quad = 1 - P(N_{LG}|Dem)$

$P(Y_{LG}|Rep) \quad = 1 - P(N_{LG}|Rep)$

Prediction:

Use Bayes to update priors with the likelihoods,
Pick label with the highest posterior probability.

What was the naïve part?

We assumed each feature is independent

We assumed each feature is independent

Easier to see in a single update rather than sequential

We assumed each feature is independent

Easier to see in a single update rather than sequential

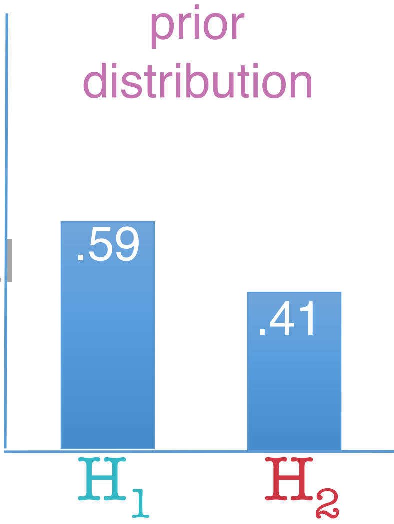
Prob. of this example having label Democrat,
given the values Yes, Yes and No
on the features NN, TC and LG

$$P(Dem | Y_{NN}, Y_{TC}, N_{LG}) = \frac{P(Y_{NN}, Y_{TC}, N_{LG} | Dem) \quad P(Dem)}{P(Y_{NN}, Y_{TC}, N_{LG})}$$

We assumed each feature is independent

Easier to see in a single update rather than sequential

Prob. of this example having label Democrat,
given the values Yes, Yes and No
on the features NN, TC and LG

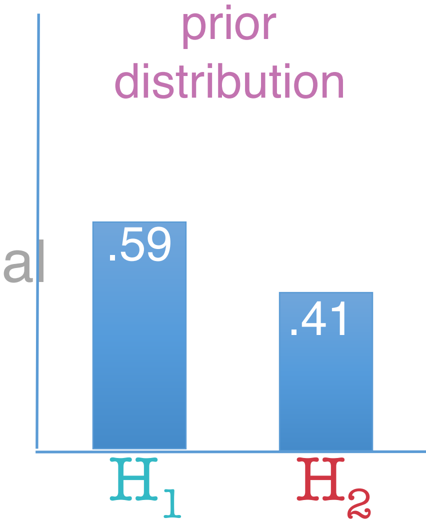


likelihood prior

posterior

$$P(Dem|Y_{NN},Y_{TC},N_{LG}) = \frac{P(Y_{NN},Y_{TC},N_{LG}|Dem)\ \ P(Dem)}{P(Y_{NN},Y_{TC},N_{LG})}$$

We assumed each feature is independent



Easier to see in a single update rather than sequential

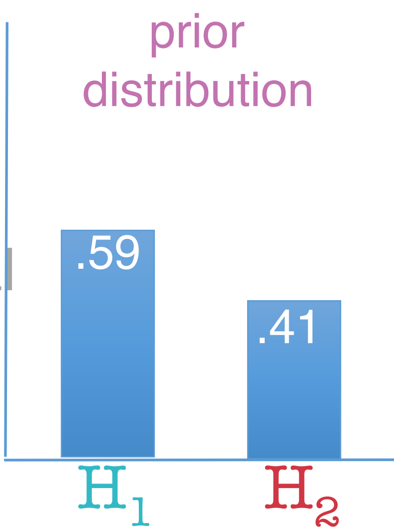
Prob. of this example having label Democrat,
given the values Yes, Yes and No
on the features NN, TC and LG

$$\overset{\text{posterior}}{P(Dem \mid Y_{NN}, Y_{TC}, N_{LG})} = \frac{\overset{\text{likelihood}}{P(Y_{NN}, Y_{TC}, N_{LG} \mid Dem)} \quad \overset{\text{prior}}{P(Dem)}}{P(Y_{NN}, Y_{TC}, N_{LG})}$$

Independence Assumption:

$P(Y_{NN}, Y_{TC}, N_{LG} \mid Dem) = P(Y_{NN} \mid Dem)\, P(Y_{TC} \mid Dem)\, P(N_{LG} \mid Dem)$

We assumed each feature is independent



Easier to see in a single update rather than sequential

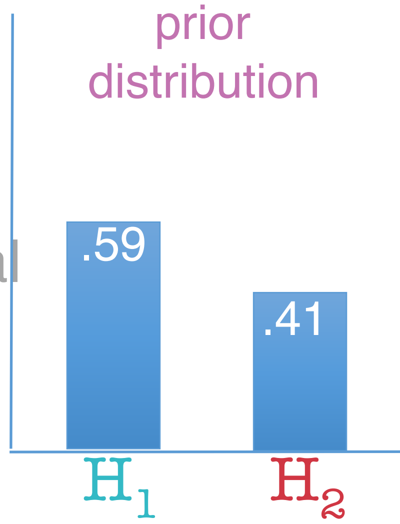
Prob. of this example having label Democrat,
given the values Yes, Yes and No
on the features NN, TC and LG

$$\underbrace{P(Dem|Y_{NN},Y_{TC},N_{LG})}_{\text{posterior}} = \frac{\overbrace{P(Y_{NN},Y_{TC},N_{LG}|Dem)}^{\text{likelihood}}\ \overbrace{P(Dem)}^{\text{prior}}}{P(Y_{NN},Y_{TC},N_{LG})}$$

Independence Assumption:
$P(Y_{NN},Y_{TC},N_{LG}|Dem) = P(Y_{NN}|Dem)\ P(Y_{TC}|Dem)\ P(N_{LG}|Dem)$

Not even close in most cases!
Naïve Bayes still works well.

We assumed each feature is independent



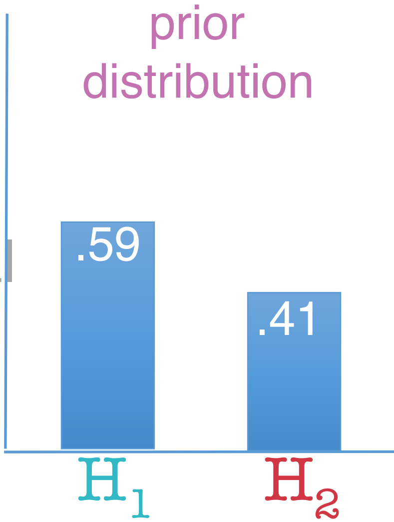
Easier to see in a single update rather than sequential

Prob. of this example having label Democrat,
given the values Yes, Yes and No
on the features NN, TC and LG

posterior / likelihood / prior

$$P(Dem \mid Y_{NN}, Y_{TC}, N_{LG}) = \frac{P(Y_{NN} \mid Dem)\, P(Y_{TC} \mid Dem)\, P(N_{LG} \mid Dem)\, P(Dem)}{P(Y_{NN}, Y_{TC}, N_{LG})}$$

We assumed each feature is independent



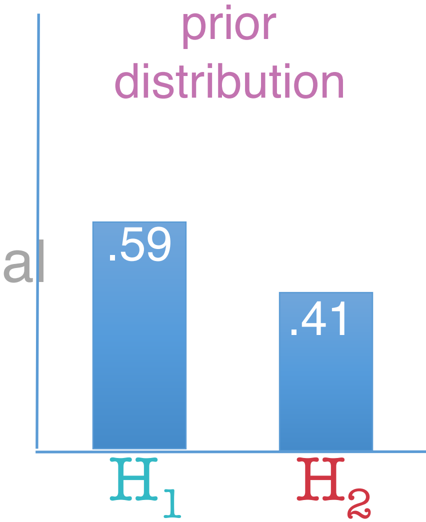
Easier to see in a single update rather than sequential

Prob. of this example having label Democrat, given the values Yes, Yes and No on the features NN, TC and LG

likelihood prior

posterior

$$P(Dem|Y_{NN},Y_{TC},N_{LG}) = \frac{0.949 * 0.169 * 0.898 \quad * \quad 0.59}{P(Y_{NN},Y_{TC},N_{LG})}$$



We assumed each feature is independent

Easier to see in a single update rather than sequential

Prob. of this example having label Democrat,
given the values Yes, Yes and No
on the features NN, TC and LG

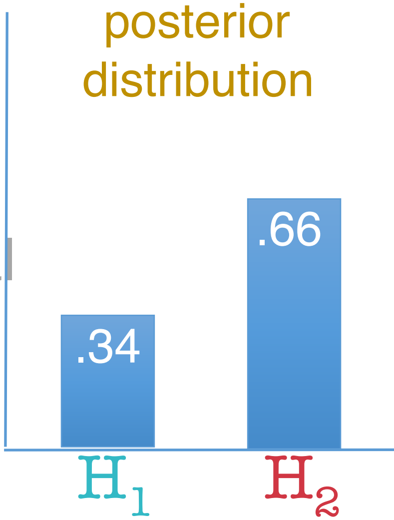
likelihood prior

posterior

$$P(Dem|Y_{NN},Y_{TC},N_{LG}) = \frac{0.949 * 0.169 * 0.898 \quad * \quad 0.59}{P(Y_{NN},Y_{TC},N_{LG})}$$

$$P(Rep|Y_{NN},Y_{TC},N_{LG}) = \frac{0.829 * 0.854 * 0.561 \quad * \quad 0.41}{P(Y_{NN},Y_{TC},N_{LG})}$$

We assumed each feature is independent



Easier to see in a single update rather than sequential

Prob. of this example having label Democrat, given the values Yes, Yes and No on the features NN, TC and LG

$P(Dem|Y_{NN},Y_{TC},N_{LG}) = 0.34$

$P(Rep|Y_{NN},Y_{TC},N_{LG}) = 0.66$ ← predict!

What about multiple classes?












Straightforward!

Update each hypothesis,
given the values Yes, Yes and No
on the features NN, TC and LG

$P(Dem|Y_{NN},Y_{TC},N_{LG})$
$P(Rep|Y_{NN},Y_{TC},N_{LG})$
$P(Indep|Y_{NN},Y_{TC},N_{LG})$

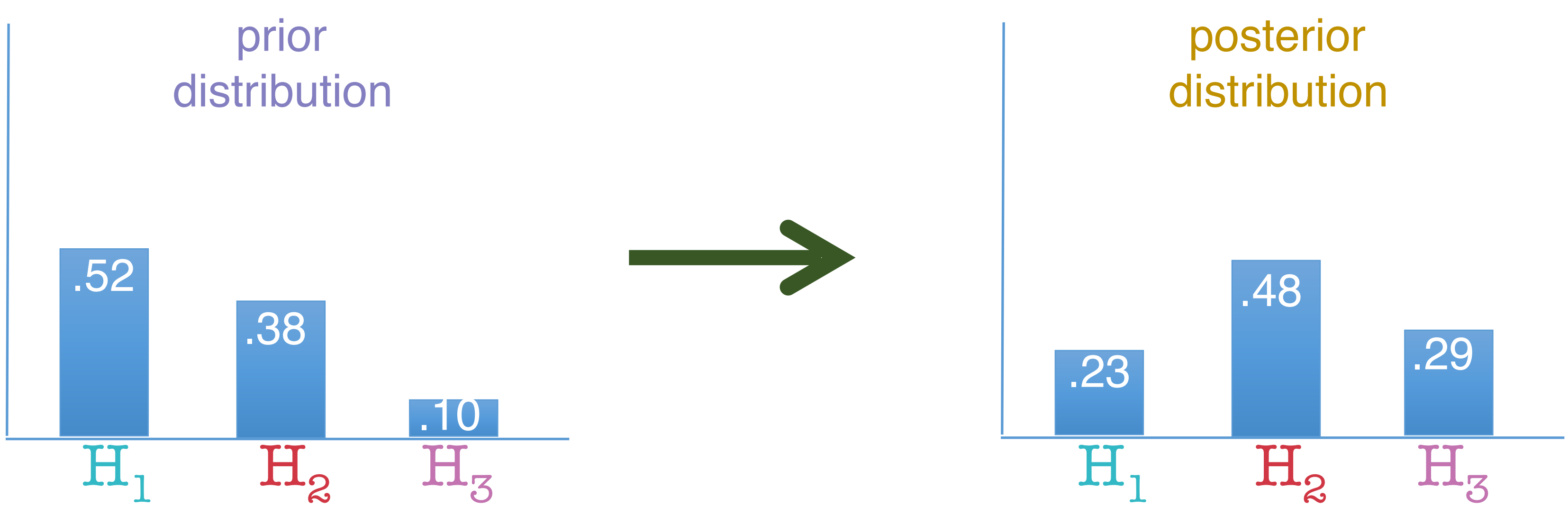
Straightforward!

Update each hypothesis,
given the values Yes, Yes and No
on the features NN, TC and LG

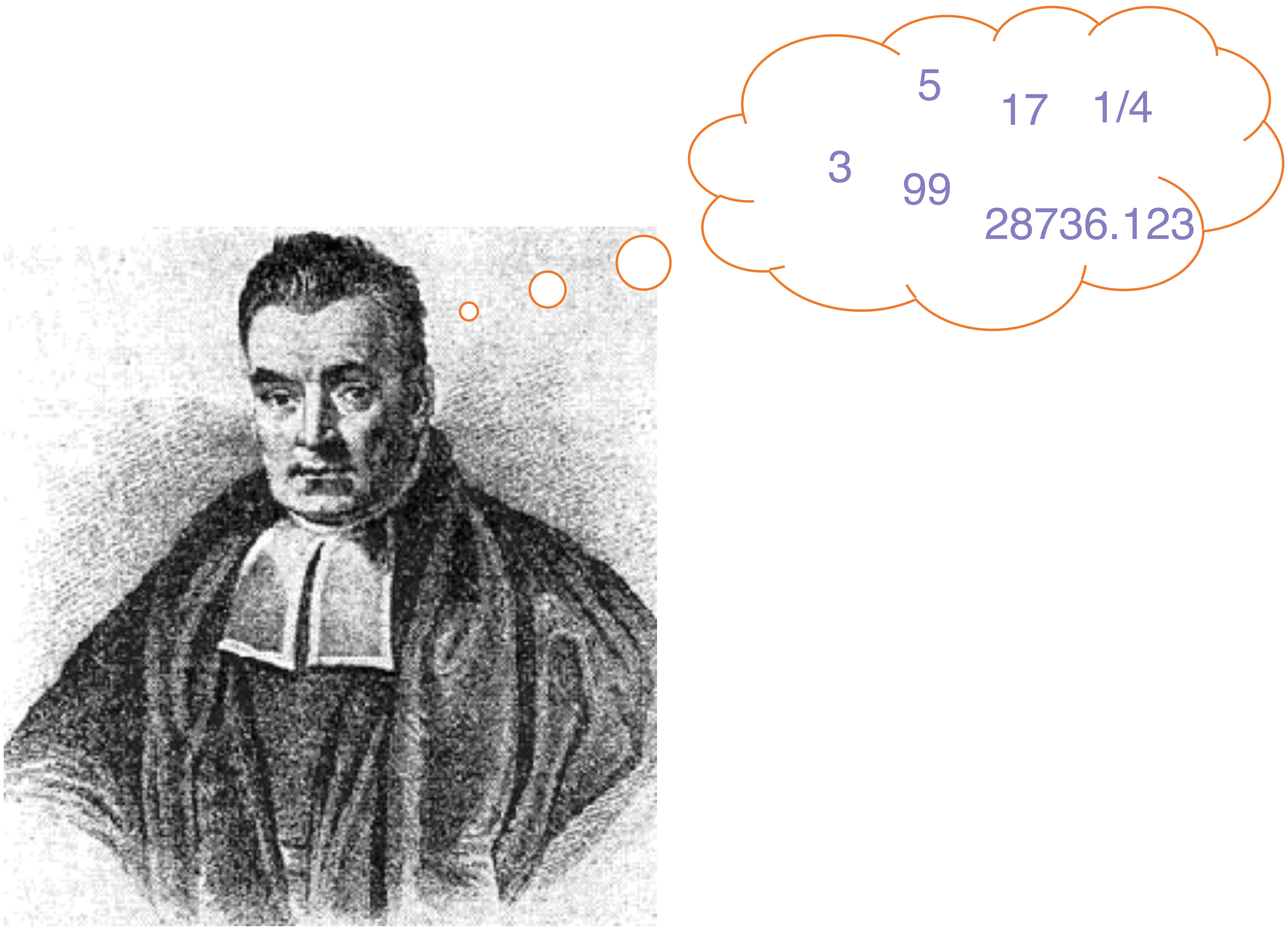
$P(Dem|Y_{NN},Y_{TC},N_{LG})$

$P(Rep|Y_{NN},Y_{TC},N_{LG})$

$P(Indep|Y_{NN},Y_{TC},N_{LG})$



How about numeric features?



Fit a likelihood function!

Instead of $P(Y_{NN}|Dem)$ or $P(N_{NN}|Dem)$,
you get a function $P(age{=}x \mid Dem)$

Fit a likelihood function!

Instead of $P(Y_{NN}|Dem)$ or $P(N_{NN}|Dem)$,
you get a function $P(age=x \mid Dem)$

Look at the training set, fit a Gaussian distribution to $P(age=x|Dem)$
Use this Gaussian likelihood when predicting.

Flavors of Bayes in `sklearn`:

Numeric Features:
Gaussian Naïve Bayes

Features that are 0 or 1 (and both matter):
Bernoulli Naïve Bayes

Features that are count-like (and only non-zero matters):
Multinomial Naïve Bayes

Flavors of Bayes in `sklearn`:

Numeric Features:
Gaussian Naïve Bayes

Features that are 0 or 1 (and both matter):
Bernoulli Naïve Bayes

Features that are count-like (and only non-zero matters):
Multinomial Naïve Bayes