

# **Naïve Bayesian**

**Graduate Program in Software  
SEIS 763: Machine Learning  
Dr. Chih Lai**

# References

## ■ Matlab

- <https://www.mathworks.com/help/stats/classification-naive-bayes.html>
- <https://www.mathworks.com/help/stats/compactclassificationnaivebayes.predict.html>


```
load fisheriris;  X = meas;  Y = species;  
  
Mdl = fitcnb(X,Y)  
[label, Posterior, Cost] = predict(Mdl, X);
```

## ■ sklearn

- [http://scikit-learn.org/stable/modules/generated/sklearn.naive\\_bayes.GaussianNB.html](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html)
- [http://scikit-learn.org/stable/modules/naive\\_bayes.html](http://scikit-learn.org/stable/modules/naive_bayes.html)
- [http://scikit-learn.org/stable/modules/generated/sklearn.naive\\_bayes.MultinomialNB.html](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html)

```
clf = GaussianNB()  
clf.fit(X, Y)  
clf.predict(X)  
clf.predict_log_proba(X)  
clf.predict_proba(X)
```

# Outline

- Classification– Examples and Issues
- Associative and Instance-Based Classification
  - Lazy and Eager Classifications
- Decision Tree and Entropy
- Issues Regarding Decision Trees
  - Overfitting, Missing Values, Noisy Data, Numeric Data, Pruning
- Sequential Covering Algorithms, Bayes Theorem 
  - Precision, Confusion Matrix
- Other Methods
  - Support Vector Machine (SVM), Neural Networks

# Naïve Bayes Classification, Simple Function Calls

## ■ Matlab

- <https://www.mathworks.com/help/stats/classification-naive-bayes.html>
- <https://www.mathworks.com/help/stats/compactclassificationnaivebayes.predict.html>

```
load fisheriris; X = meas; Y = species;
```

```
Mdl = fitcnb(X,Y)
```

```
[label, Posterior, Cost] = predict(Mdl, X);
```

## ■ sklearn

- [http://scikit-learn.org/stable/modules/generated/sklearn.naive\\_bayes.GaussianNB.html](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html)
- [http://scikit-learn.org/stable/modules/naive\\_bayes.html](http://scikit-learn.org/stable/modules/naive_bayes.html)
- [http://scikit-learn.org/stable/modules/generated/sklearn.naive\\_bayes.MultinomialNB.html](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html)

```
clf = GaussianNB()  
clf.fit(X, Y)  
clf.predict(X)  
clf.predict_log_proba(X)  
clf.predict_proba(X)
```

## ■ Problems of just using tools??

- We know the prediction results. *Lazy method.*
- But, we don't know which predictors are more important.
- Can we obtain this information by ourselves?
- Yes, very easy... as long as you know SQL.
- Importance is NOT fixed like  $\theta$ s in other ML methods.
- Importance under NB is **dynamic** → **varies on different data**.

-1.2 0.2 SVM  
0.6 -0.5 0.8 -0.5

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

(S, C, **H↑**, T)

(S, **W↑**, H, T)

(S, W, H, **F↓**)

# Naïve Bayes Classification



- A very easy classifier
  - NO need to select attrs. like in DT or SC algorithm.
  - NO need to do divide-and-conquer.
  - Easy SQL (or easiest SQL).
  - Lazy evaluation, generating NO rules (**explicitly**).
  
- Two assumptions on data attributes
  - Equally important & statistically independent.
    - Knowledge of a one attribute has nothing to do with another attribute.
  - Study shows NB still works well in practice even assumptions do not hold.
  
- One of few potential problems:
  - **Zero**-frequency.

# Bayesian Theorem

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

- If we have a training set that have a set of attributes
  - Given  $d = (S, C, H, W)$  as a new instance.
  - Probability that the new instance  $d$  belong to class *Yes* & *No*.

$$P(\text{Yes} | d) = \frac{P(d | \text{Yes}) \times P(\text{Yes})}{P(d)}$$

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

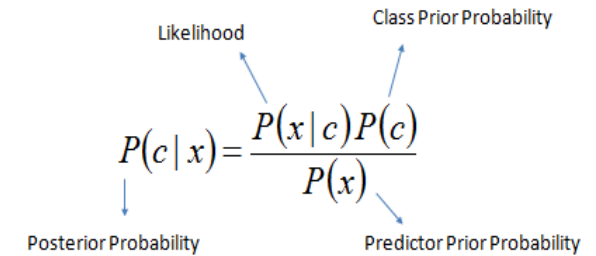
Likelihood:  $P(x | c)$   
 Class Prior Probability:  $P(c)$   
 Posterior Probability:  $P(c | x)$   
 Predictor Prior Probability:  $P(x)$

- 1) Posterior probability**  $d$  belongs to class *Yes*, if  $d$  has attributes (sunny, cool, high humidity, windy)?
- 2) Likelihood** our trainings have attributes like  $d$  (sunny, cool, high humidity, windy) & belong to class *Yes*?
- 3) Prior probability** of trainings belong to class *Yes* in our training set.

$$P(\text{Yes} | d) = \frac{P(\text{Sunny} | \text{Yes}) \times P(\text{Cool} | \text{Yes}) \times P(\text{HighHumidity} | \text{Yes}) \times P(\text{Windy} | \text{Yes}) \times P(\text{Yes})}{P(d)}$$

# Naïve Bayes Theorem for Classification

- Assume attributes of instances are independent.
  - Is Saturday  $d = (S, C, H, T)$  OK for playing golf?



$$P(\text{Yes} | d) = \frac{P(\text{Sunny}|\text{Yes}) \times P(\text{Cool}|\text{Yes}) \times P(\text{HighHumidity}|\text{Yes}) \times P(\text{Windy}|\text{Yes}) \times \mathbf{P(\text{Yes})}}{P(d)}$$

Normalize so sum = 1

$$\begin{aligned}
 P(\mathbf{\text{Yes}} | d) &\rightarrow \text{If } d \text{ has attributes..., what's the probability that } d \text{ is Y?} \\
 &= \frac{P(S|\text{Yes}) \times P(C|\text{Yes}) \times P(\text{High}|\text{Yes}) \times P(T|\text{Yes}) \times \mathbf{P(\text{Yes})}}{P(d)} \\
 &= \frac{\left(\frac{2}{9}\right) \times \left(\frac{3}{9}\right) \times \left(\frac{3}{9}\right) \times \left(\frac{3}{9}\right) \times \mathbf{\left(\frac{9}{14}\right)}}{P(d)} = \frac{0.0053}{P(d)} \\
 P(\mathbf{\text{No}} | d) &\rightarrow \text{If } d \text{ has attributes..., what's the probability that } d \text{ is N?}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{P(S|\text{No}) \times P(C|\text{No}) \times P(\text{High}|\text{No}) \times P(T|\text{No}) \times \mathbf{P(\text{No})}}{P(d)} \\
 &= \frac{\left(\frac{3}{5}\right) \times \left(\frac{1}{5}\right) \times \left(\frac{4}{5}\right) \times \left(\frac{3}{5}\right) \times \mathbf{\left(\frac{5}{14}\right)}}{P(d)} = \frac{0.0206}{P(d)}
 \end{aligned}$$

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

## Normalizing Naïve Bayes Theorem

- Is Saturday morning  $d = (S, C, H, T)$  OK for playing golf?

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

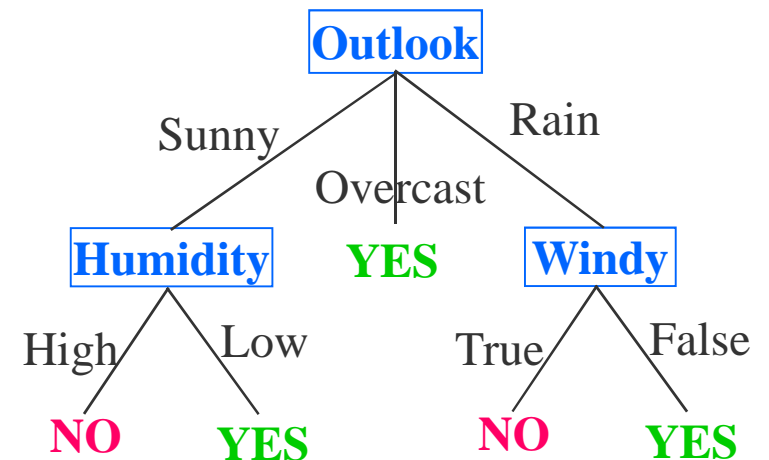
$$P(Yes | d) = \frac{0.0053}{P(d)} \quad P(No | d) = \frac{0.0206}{P(d)}$$

$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0.0259}{P(d)}$$

### Normalization

$$P(Yes | d) = \frac{0.0053}{P(d)} \times \frac{1}{SUM} = \frac{0.0053}{0.0259} = 0.205$$

$$P(No | d) = \frac{0.0206}{P(d)} \times \frac{1}{SUM} = \frac{0.0206}{0.0259} = 0.795$$





# Comparing Naïve Bayes Theorem with DT

- Is Saturday morning  $d = (S, C, H, T)$  OK for playing golf?

- DT prediction?

- $(2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0.0053$  **0.205**

- $(3/5) * (1/5) * (4/5) * (3/5) * (5/14) = 0.0206$  **0.795**

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

- $d = (S, W, H, T)$ ?

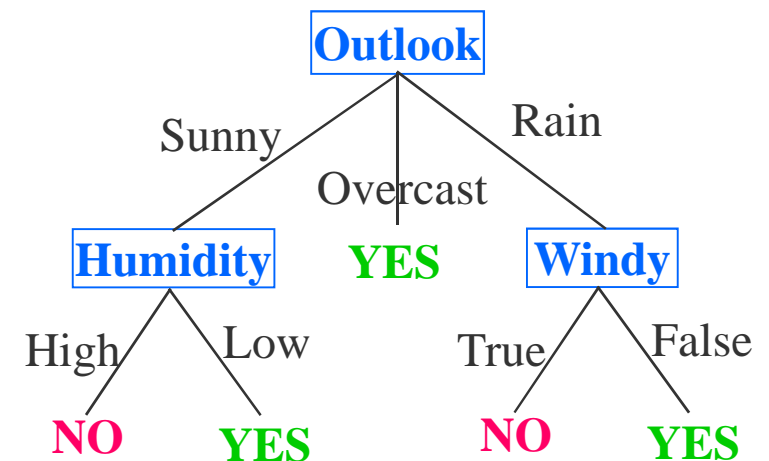
- DT prediction?

- $(2/9) * (2/9) * (3/9) * (3/9) * (9/14) = 0.0035$  **0.0785**

- $(3/5) * (2/5) * (4/5) * (3/5) * (5/14) = 0.0411$  **0.9215**

- S + H already a bad day to play.

- S + H + W make it even worse!!



## Another Example

- Is Saturday morning  $d = (S, \text{W}, H, \underline{F})$  OK for playing golf?

$$P(\text{Yes} \mid d) = \frac{\left(\frac{2}{9}\right) \times \left(\frac{2}{9}\right) \times \left(\frac{3}{9}\right) \times \left(\frac{6}{9}\right) \times \left(\frac{9}{14}\right)}{P(d)} = \frac{0.0007}{P(d)}$$

$$P(\text{No} \mid d) = \frac{\left(\frac{3}{5}\right) \times \left(\frac{2}{5}\right) \times \left(\frac{4}{5}\right) \times \left(\frac{2}{5}\right) \times \left(\frac{5}{14}\right)}{P(d)} = \frac{0.028}{P(d)}$$

$$SUM = \Sigma(P(\text{Yes} \mid d), P(\text{No} \mid d)) = \frac{0.035}{P(d)}$$

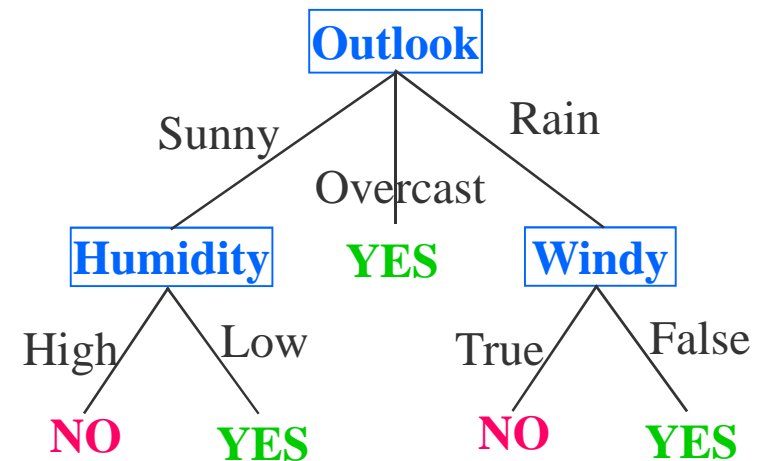
Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

### Normalization

$$P(\text{Yes} \mid d) = \frac{0.007}{P(d)} \times \frac{1}{SUM} = \frac{0.007}{0.035} = 0.2$$

$$P(\text{No} \mid d) = \frac{0.028}{P(d)} \times \frac{1}{SUM} = \frac{0.028}{0.035} = 0.8$$

- $S + H + \underline{W} + \underline{F}$  improve a bit!!



## Different Prediction than DT

- Is Saturday morning  $d = (S, W, L, T)$  OK for playing golf?
  - Comparing the result with DT...

$$P(Yes | d) = \frac{\frac{2}{9} \times \frac{2}{9} \times \frac{6}{9} \times \frac{3}{9} \times \frac{9}{14}}{P(d)} = \frac{0.0007}{P(d)}$$

$$P(No | d) = \frac{\frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{5}{14}}{P(d)} = \frac{0.01}{P(d)}$$

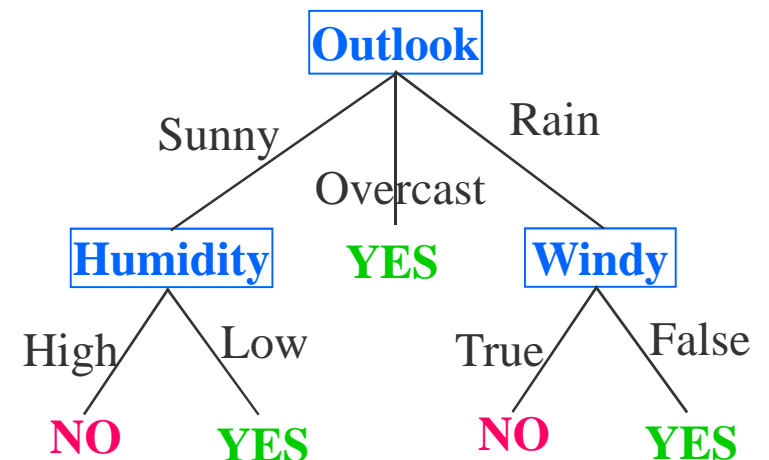
$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0.017}{P(d)}$$

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

### Normalization

$$P(Yes | d) = \frac{0.007}{P(d)} \times \frac{1}{SUM} = \frac{0.007}{0.017} = 0.412$$

$$P(No | d) = \frac{0.01}{P(d)} \times \frac{1}{SUM} = \frac{0.01}{0.017} = 0.588$$



# Building An NB Model??

Outlook			Temp.			Humid			Windy			Play		
	Y	N		Y	N		Y	N		Y	N		Y	N
S	2	3	W	2	2	H	3	4	T	3	3		9	5
O	4	0	M	4	2	L	6	1	F	6	2			
R	3	2	C	3	1									

## Eager or Lazy?? Distribution Table

Attribute	Parameter	Class 0	Class 1
GENDER	value=Female	0.545671429	0.577463614
GENDER	value=Male	0.454328571	0.422536386
BLACK_RACE_IND	value=0	0.889912857	0.902520574
BLACK_RACE_IND	value=1	0.110087143	0.097479426
OTHER_RACE_IND	value=0	0.954807143	0.96471772
OTHER_RACE_IND	value=1	0.045192857	0.03528228
HISPANIC_RACE_IND	value=0	0.975504286	0.979322241
HISPANIC_RACE_IND	value=1	0.024495714	0.020677759
DIMENTIA_IND	value=0	0.832718571	0.62376972
DIMENTIA_IND	value=1	0.167281429	0.37623028
KIDNEY_IND	value=0	0.858205714	0.640800661
KIDNEY_IND	value=1	0.141794286	0.359199339
CANCER_IND	value=1	0.056975714	0.136483272
CANCER_IND	value=0	0.943024286	0.863516728
COPD_IND	value=1	0.105595714	0.277692888
COPD_IND	value=0	0.894404286	0.722307112
DEPRESSION_IND	value=0	0.81729	0.605798685
DEPRESSION_IND	value=1	0.18271	0.394201315
ISCHEMIC_IND	value=1	0.361364286	0.748454379
ISCHEMIC_IND	value=0	0.638635714	0.251545621

## Troubling Example, Where is the Trouble??

- Is Saturday morning  $d = (O, C, L, F)$  OK for playing golf?

$$P(Yes | d) = \frac{\frac{4}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}}{P(d)} = \frac{0.0423}{P(d)}$$

$$P(No | d) = \frac{\frac{0}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14}}{P(d)} = \frac{0.0}{P(d)}$$

$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0.0423}{P(d)}$$

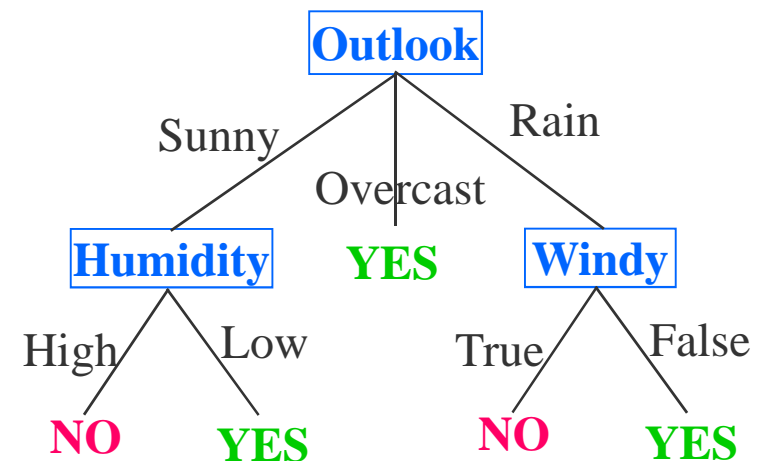
Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

Outlook	Play
O	Y
O	Y
O	Y
O	Y
R	N
R	N
R	Y
R	Y
R	Y
R	Y
S	N
S	N
S	N
S	Y
S	Y

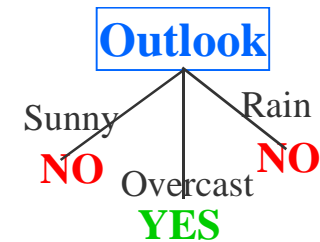
### Normalization

$$P(Yes | d) = \frac{0.0423}{P(d)} \times \frac{1}{SUM} = \frac{0.0423}{0.0423} = 1.0$$

$$P(No | d) = \frac{0}{P(d)} \times \frac{1}{SUM} = \frac{0}{0.0423} = 0.0$$



# The Zero-Frequency Problem<sup>1</sup>



- Is Saturday morning  $d = (S, M, L, F)$  OK for playing golf?

$$P(Yes | d) = \frac{\frac{0}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{14}}{P(d)} = \frac{0}{P(d)}$$

$$P(No | d) = \frac{\frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{5}{14}}{P(d)} = \frac{0.0006}{P(d)}$$

$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0.0006}{P(d)}$$

## Normalization

$$P(Yes | d) = \frac{0}{P(d)} \times \frac{1}{SUM} = \frac{0}{0.0006} = 0.0$$

$$P(No | d) = \frac{0.0006}{P(d)} \times \frac{1}{SUM} = \frac{0.028}{0.0006} = 1.0$$

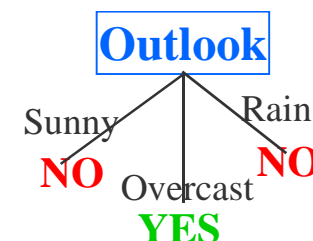
Note, Table (1)!!!

Outlook	Temp.	Humid	Windy	Play
S	M	L	F	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y

- 0 probability veto over others, no matter how likely the other values are!
- Remedy: add 1 to the count for every attribute value- class combination
  - It does not influence the probability P(yes) or P(no) (Weka p85)
  - Result: probabilities will never be zero!

## The Zero-Frequency Problem– Even Worse!

- Is Saturday morning  $d = (S, M, L, F)$  OK for playing golf?



$$P(Yes | d) = \frac{\frac{0}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{14}}{P(d)} = \frac{0}{P(d)}$$

$$P(No | d) = \frac{\frac{1}{5} \times \frac{0}{5} \times \frac{0}{5} \times \frac{0}{5} \times \frac{5}{14}}{P(d)} = \frac{0}{P(d)}$$

$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0}{P(d)}$$

### Normalization

$$P(Yes | d) = \frac{0}{P(d)} \times \frac{1}{SUM} = \frac{0}{0} = 0.0$$

$$P(No | d) = \frac{0.0006}{P(d)} \times \frac{1}{SUM} = \frac{0}{0} = 0.0$$

Note, Table (2)!!!

Outlook	Temp.	Humid	Windy	Play
S	W	H	T	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y

- 0 probability veto** over others, no matter how likely the other values are!
- Remedy: add 1 to the count for every attribute value- class combination
  - It does not influence the probability P(yes) or P(no) (Weka p85)
  - Result: probabilities will never be zero!

# Probability Estimation

- Original:  $P(A_i | C) = \frac{N_{ic}}{N_c}$

- Laplace:  $P(A_i | C) = \frac{N_{ic}+1}{N_c+c}$        $\frac{N_{ic}+1}{N_c+a}$

- $\mu$ -estimate:  $P(A_i | C) = \frac{N_{ic}+\mu p}{N_c+\mu}$

$c$ : # of classes (or attribute values)

$p$ : prior probability

$\mu$ : parameter

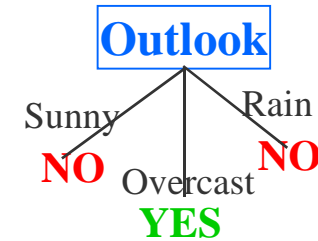
Outlook	Temp.	Humid	Windy	Play
S	M	L	F	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y

Outlook			Temp.			Humid			Windy			Play		
	Y	N		Y	N		Y	N		Y	N		Y	N
S	0→1	1→2	W	0→1	4→5	H	0→1	4→5	T	0→1	4→5		9	5
O	9→10	0→1	M	9→10	1→2	L	9→10	1→2	F	9→10	1→2			
R	0→1	4→5	C	0→1	0→1									



# Laplace Estimator for The Zero-Frequency Problem<sup>1</sup>

- Is Saturday morning  $d = (S, M, L, F)$  OK for playing golf?



$$P(Yes | d) = \frac{\frac{1}{12} \times \frac{10}{12} \times \frac{10}{11} \times \frac{10}{11} \times \frac{9}{14}}{P(d)} = \frac{0.0369}{P(d)}$$

$$P(Yes | d) = \frac{\frac{0}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{14}}{P(d)}$$

$$P(No | d) = \frac{\frac{2}{8} \times \frac{2}{8} \times \frac{2}{7} \times \frac{2}{7} \times \frac{5}{14}}{P(d)} = \frac{0.0018}{P(d)}$$

$$P(No | d) = \frac{\frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{5}{14}}{P(d)}$$

$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0.0387}{P(d)}$$

## Normalization

$$P(Yes | d) = \frac{0.0369}{P(d)} \times \frac{1}{SUM} = \frac{0.0369}{0.0387} = 0.9535$$

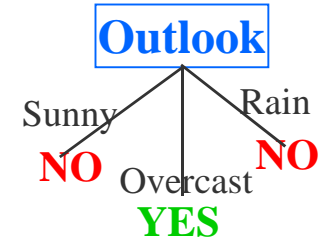
$$P(No | d) = \frac{0.0018}{P(d)} \times \frac{1}{SUM} = \frac{0.0018}{0.0387} = 0.0465$$

Outlook	Temp.	Humid	Windy	Play
S	M	L	F	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y

Outlook			Temp.			Humid			Windy			Play	
	Y	N		Y	N		Y	N		Y	N		
S	0→1	1→2	W	0→1	4→5	H	0→1	4→5	T	0→1	4→5	9	5
O	9→10	0→1	M	9→10	1→2	L	9→10	1→2	F	9→10	1→2		
R	0→1	4→5	C	0→1	0→1								

# Laplace Estimator for The Zero-Frequency Problem<sup>2</sup>

- Is Saturday morning  $d = (S, M, L, F)$  OK for playing golf?



$$P(Yes | d) = \frac{\frac{1}{12} \times \frac{10}{12} \times \frac{10}{11} \times \frac{10}{11} \times \frac{9}{14}}{P(d)} = \frac{0.0369}{P(d)}$$

$$P(No | d) = \frac{\frac{2}{8} \times \frac{1}{8} \times \frac{1}{7} \times \frac{1}{7} \times \frac{5}{14}}{P(d)} = \frac{0.0002}{P(d)}$$

$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0.0371}{P(d)}$$

## Normalization

$$P(Yes | d) = \frac{0.0369}{P(d)} \times \frac{1}{SUM} = \frac{0.0369}{0.0371} = 0.9946$$

$$P(No | d) = \frac{0.0002}{P(d)} \times \frac{1}{SUM} = \frac{0.0002}{0.0371} = 0.0054$$

$$P(Yes | d) = \frac{\frac{0}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{9} \times \frac{9}{14}}{P(d)} = \frac{0}{P(d)}$$

$$P(No | d) = \frac{\frac{1}{5} \times \frac{0}{5} \times \frac{0}{5} \times \frac{0}{5} \times \frac{5}{14}}{P(d)} = \frac{0}{P(d)}$$

Outlook	Temp.	Humid	Windy	Play
S	W	H	T	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y

## Troubling Example, Where is the Trouble??



- Is Saturday morning  $d = (O, C, L, F)$  OK for playing golf?

$$P(Yes | d) = \frac{\frac{5}{12} \times \frac{4}{12} \times \frac{7}{11} \times \frac{7}{11} \times \frac{10}{16}}{P(d)} = \frac{0.0352}{P(d)}$$

$$P(No | d) = \frac{\frac{1}{8} \times \frac{2}{8} \times \frac{2}{7} \times \frac{3}{7} \times \frac{6}{16}}{P(d)} = \frac{0.0014}{P(d)}$$

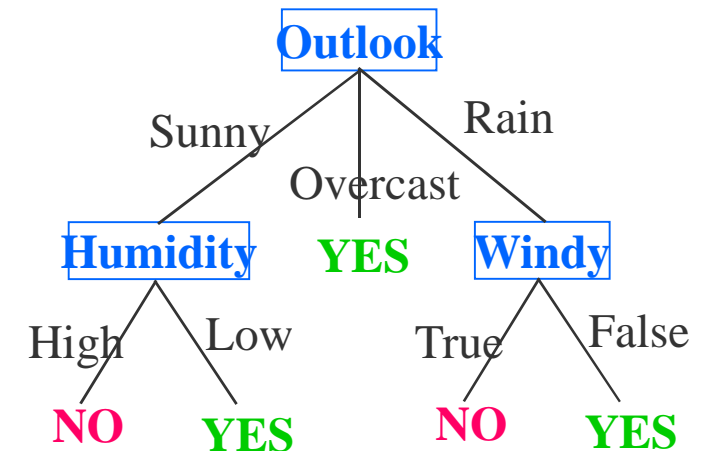
$$SUM = \Sigma(P(Yes | d), P(No | d)) = \frac{0.0366}{P(d)}$$

### Normalization

$$P(Yes | d) = \frac{0.0352}{P(d)} \times \frac{1}{SUM} = \frac{0.0352}{0.0366} = 0.9608$$

$$P(No | d) = \frac{0.0014}{P(d)} \times \frac{1}{SUM} = \frac{0.0014}{0.0366} = 0.0392$$

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y



$$P(Yes | d) = \frac{\frac{4}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}}{P(d)} = \frac{0.0423}{P(d)}$$

$$P(No | d) = \frac{\frac{0}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14}}{P(d)} = \frac{0.0}{P(d)}$$

## More On Laplace Estimator

- Another remedy to 0-veto problem is Laplace estimator  $\mu$ ,  $\mu$  can be any value
- ➡ Large  $\mu \rightarrow$  prior probabilities of trainings are very important.
  - In practice add 1 to denominator & assume each attribute value is equally probable

$$\frac{N_{ic}+1}{N_c+a}$$

$$\frac{N_{ic}+\mu p}{N_c+\mu}$$

- Add any  $\mu$  (i.e.  $\mu = 1.7$ ) to the denominator, and
- Add  $\mu p_i$  to the numerator
  - $p_i$  = prior probability of the  $i$  values of the attribute, and  $\sum p_i = 1$
  - For “Humidity”,  $p_{high} = 5/14$ ,  $p_{low} = 9/14$

$$P(High | Yes) = \frac{N_{ic} + \mu p}{N_c + \mu} = \frac{0 + \mu p_i}{9 + \mu} = \frac{0 + 1.7 \times 5/14}{9 + 1.7}$$

Outlook	Temp.	Humid	Windy	Play
S	W	H	T	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
R	W	H	T	N
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y
O	M	L	F	Y

# Animal Prediction

Original:  $P(A_i | C) = \frac{N_{ic}}{N_c}$

Laplace:  $P(A_i | C) = \frac{N_{ic}+1}{N_c+c}$        $\frac{N_{ic}+1}{N_c+a}$

$\mu$ -estimate:  $P(A_i | C) = \frac{N_{ic}+\mu p}{N_c+\mu}$

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
turtle	cold	no	no	sometimes	?

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
penguin	warm	no	no	sometimes	

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
dogfish shark	cold	yes	no	yes	

## More On Laplace Estimator

- Add a constant  $U$  to each denominator, and  $U \cdot p_i$  to each numerator
  - ➡ Larger  $U \rightarrow$  the prior probabilities of trainings are more important
  - ➡ Smaller  $U \rightarrow$  less important w.r.t. test samples
    - In practice add 1 to denominator & assume each attribute value is equally probable
    - i.e. all  $p_i$  are the same,  $p_1 = p_2 = \dots$

$$\text{Original: } P(A_i | C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace: } P(A_i | C) = \frac{N_{ic} + 1}{N_c + c} \quad \frac{N_{ic} + 1}{N_c + a}$$

$$\mu\text{-estimate: } P(A_i | C) = \frac{N_{ic} + \mu p}{N_c + \mu}$$

U = 0

'dogfish'	'fishes'	[ 1]	'amphibians'	[ 0]
'turtle'	'amphibians'	[0.6667]	'reptiles'	[0.3333]
'pengiun'	'birds'	[ 1]	'amphibians'	[ 0]
'bear'	'mammals'	[ 1]	'amphibians'	[ 0]
'hawk'	'birds'	[0.9363]	'mammals'	[0.0637]

U = 0.5

'dogfish'	'fishes'	[0.9277]	'mammals'	[0.0520]
'turtle'	'amphibians'	[0.5630]	'reptiles'	[0.3893]
'pengiun'	'birds'	[0.6789]	'amphibians'	[0.2061]
'bear'	'mammals'	[0.9739]	'birds'	[0.0222]
'hawk'	'birds'	[0.9001]	'mammals'	[0.0928]

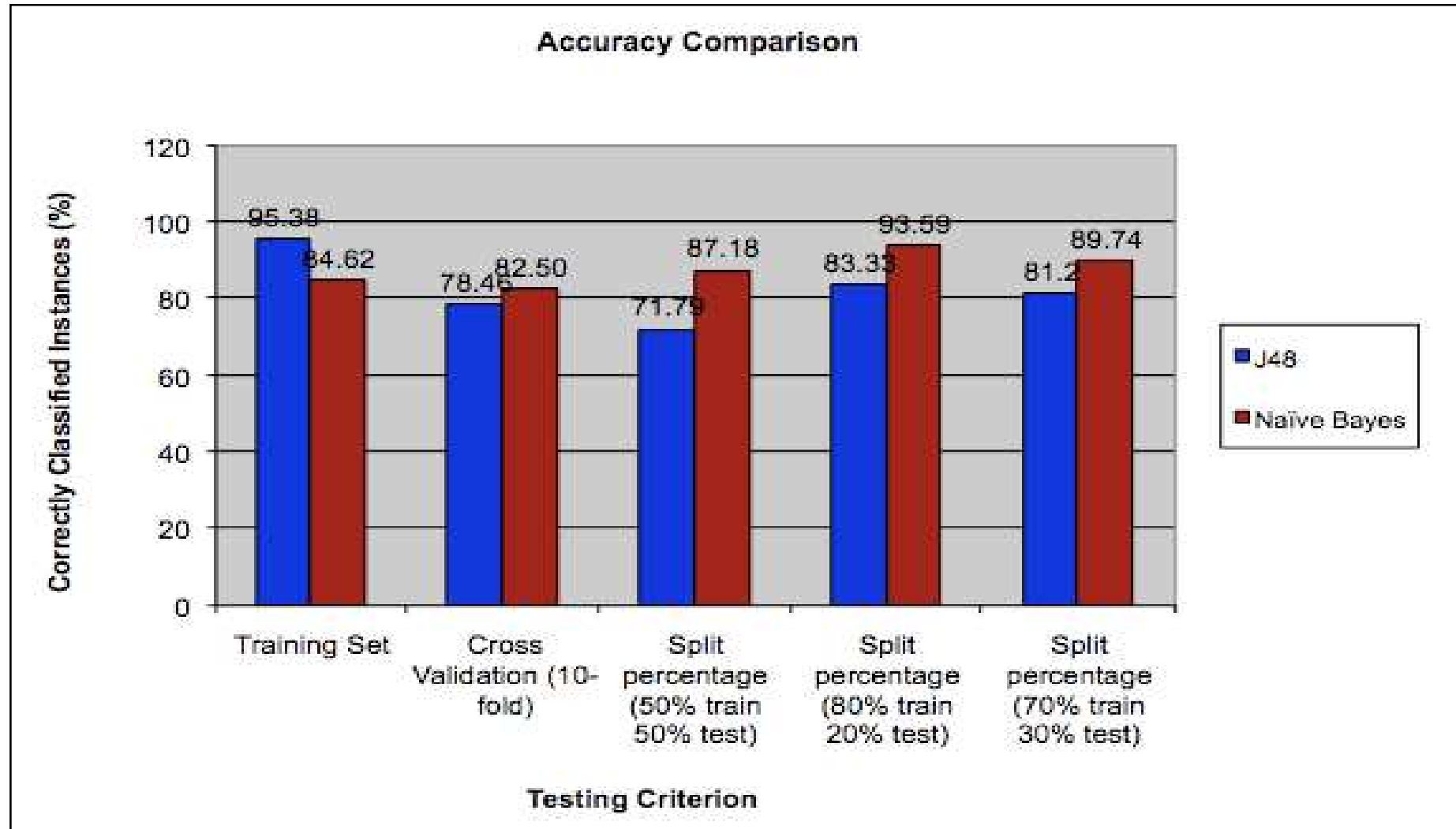
U = 2

'dogfish'	'fishes'	[0.6978]	'mammals'	[0.1727]
'turtle'	'reptiles'	[0.4446]	'amphibians'	[0.3889]
'pengiun'	'birds'	[0.4705]	'amphibians'	[0.2287]
'bear'	'mammals'	[0.8986]	'birds'	[0.0644]
'hawk'	'birds'	[0.7551]	'mammals'	[0.1721]

## Simplifying NB Method??

- Use Lasso results to select predictors before building an NB model.
  - To remove less useful predictors from NB, hopefully reduce noise contributions.
- Why simplifying models?
  - Simplified models have less chance of overfitting.
  - Simplified models MAY perform better on minority class(es), when unbalanced.

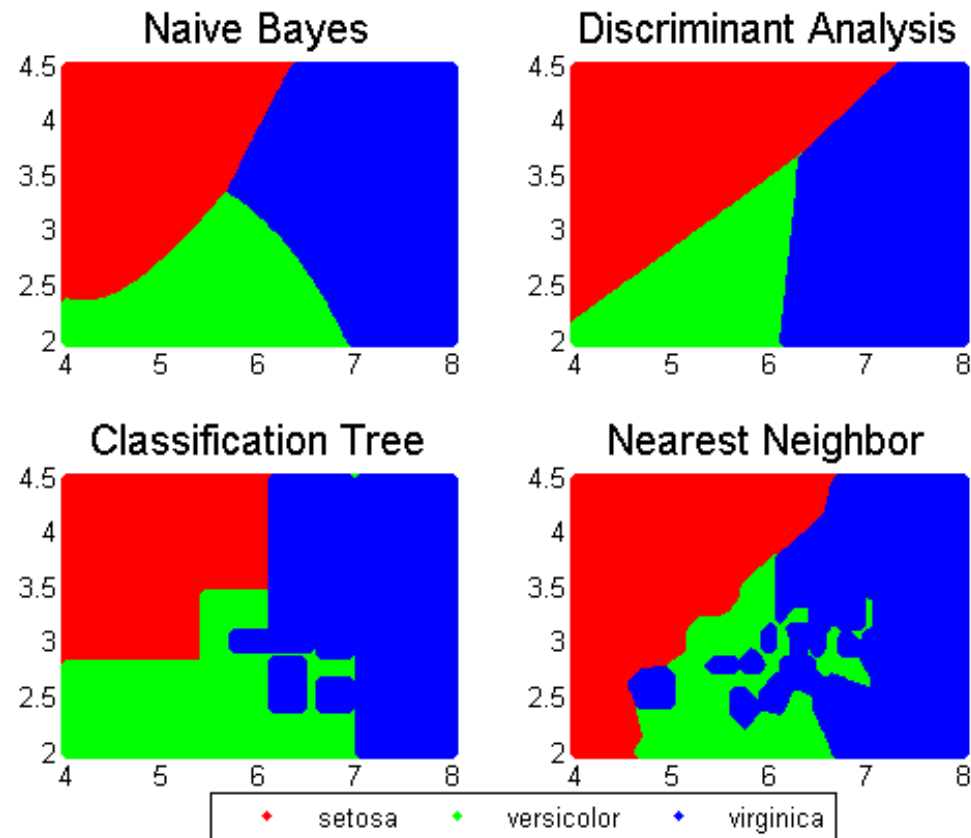
# Algorithm Comparison





# Comparing Attribute Boundary in Predictions

- [http://html?file=/products/demos/machine-learning/decision\\_surface/decision\\_surface.html](http://html?file=/products/demos/machine-learning/decision_surface/decision_surface.html)



## Naïve Bayes Advantages

- Simple lazy evaluation classifier.
  - NO need to select attributes like in DT / SC algorithm, NO need of **divide-&-conquer**.
  - Calculate  $P$  for each attribute is very fast → Good for real-time prediction.
  - Parallelize probability calculation on multiple predictors.
- Easy handling of missing data
  - A missing predictor value can be ignored while computing class probability.
- Required less data than you think (no need to have huge data).      **think about N.N.**
- Robust to *isolated* noise points.      **Mutually exclusive and exhaustive.**
- Robust to irrelevant predictors.      How about highly correlated predictors???

# Naïve Bayes Disadvantages

- Zero frequency problem.
- Attributes must be discrete.
  - *Usually*, numeric data must be clustered / grouped / binning...
- NB ignores interactions between attributes; hence need less data than...
  - Other algorithms that find predictor interactions, such as logistic regression.
- Better remove correlated (i.e. redundant) predictors
  - NB degrades if too many highly correlated predictors since they vote multiple times.
  - Inflating their importance.
- **Bayesian Network** for dependent attributes.....
  - Use other techniques such as Bayesian Belief Networks (BBN)

## NB– Predicting Medical Keywords Example

- Can you write a SQL to perform NB from the modified table? Can you increase quality?

Training example

PM	CD
423,851	3,559
423,851	16,482
423,851	53,232

Predicted Code

PM	CD	DUI	DC_RATE	NOT_DC_RATE	NORM_DC	NORM_NOT_DC
423,851	-11,861	-2,539	0.00000000459538	0.00000000010204	0.978	0.021
423,851	-7,884	-11,861	0.00000000489396	0.000000000022	0.956	0.043
423,851	-5,005	-6,462	0.00000000028137	0.00000000001967	0.934	0.065
423,851	-2,539	-772	0.00000000004309	0.00000000003886	0.525	0.474
423,851	-772	-5,005	0.0000000001142	0.0000000001041	0.523	0.476
		-8,526	0.00000620724036	0.00000617928881	0.501	0.498
		-14,717	0.00000018530627	0.00000023784869	0.437	0.562
		-7,884	0.00000000010307	0.00000000013785	0.427	0.572
		-9,801	0.00000026640444	0.00000041618732	0.39	0.609
		-1,403	0.00000454082739	0.0000071196741	0.389	0.61
		-564	0.00000371470724	0.00000748692979	0.331	0.668
		-10,278	0.00000348050866	0.00000801893789	0.302	0.697
		-12,728	0.00000014648152	0.00000036866277	0.284	0.715
		-12,419	0.00000312277743	0.00000808105849	0.278	0.721
		-17,647	0.00000295591372	0.00000822707533	0.264	0.735

	Top 5 predictions	Mis-predictions	
Real DUI (5)	4	1	80%
Others (26,592)	1	XXX	
	80%		<b>F1=0.8</b>

	Top 3 predictions	Mis-predictions	
Real DUI (5)	2	3	40%
Others (26,592)	1	XXX	
	66%		<b>F1=0.498</b>

# How To Apply NB to Datasets with Numeric Predictors?

- Binning / discretizing numerical predictors before counting *frequency*.

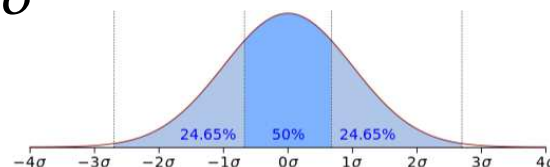
Outlook			Temp.			Humid			Windy			Play		
	Y	N		Y	N		Y	N		Y	N		Y	N
S	0→1	1→2	W	0→1	4→5	H	0→1	4→5	T	0→1	4→5		9	5
O	9→10	0→1	M	9→10	1→2	L	9→10	1→2	F	9→10	1→2			
R	0→1	4→5	C	0→1	0→1									

- Alternative: estimate frequency from the *distribution* of the numerical predictor.
  - One common practice is to assume *normal distribution* for numerical predictors.

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\sigma = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$$

$$f(w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(w-\mu)^2}{\sigma^2}}$$



	Temp.	$\mu$	$\sigma$
<b>Yes</b>	86, 96, 80, 65, 70, 80, 70, 90, 75	79.1	10.2
<b>No</b>	85, 90, 70, 95, 91	86.2	9.7

Play

Play

$$P(\text{Temp.} = 73 \mid \text{Play} = \text{Yes}) = 0.0345 \rightarrow 0.65$$

$$P(\text{Temp.} = 73 \mid \text{Play} = \text{No}) = 0.0187 \rightarrow 0.35$$

## Numeric Predictors, Another View

$$\begin{array}{l}
 \text{Pre-Norm Prob.} \\
 = \text{PNP}_0 \\
 P(No \mid d) = \frac{\overbrace{\left(\frac{2}{9}\right) \times \left(\frac{2}{9}\right) \times \left(\frac{3}{9}\right) \times \left(\frac{6}{9}\right)}^{\text{PDF}_{00} \times \text{PDF}_{10} \times \dots} \times \overbrace{\left(\frac{9}{14}\right)}^{\text{Priori Pr}_0}}{P(d)} = \frac{\overbrace{0.0007}^{\text{PNP}_0}}{P(d)}
 \end{array}$$

$$\begin{array}{l}
 \text{Pre-Norm Prob.} \\
 = \text{PNP}_1 \\
 P(Yes \mid d) = \frac{\overbrace{\left(\frac{3}{5}\right) \times \left(\frac{2}{5}\right) \times \left(\frac{4}{5}\right) \times \left(\frac{2}{5}\right)}^{\text{PDF}_{01} \times \text{PDF}_{11} \times \dots} \times \overbrace{\left(\frac{5}{14}\right)}^{\text{Priori Pr}_1}}{P(d)} = \frac{\overbrace{0.028}^{\text{PNP}_1}}{P(d)}
 \end{array}$$

$$SUM = \Sigma(P(Yes \mid d), P(No \mid d)) = \frac{0.035}{P(d)}$$

### Normalization

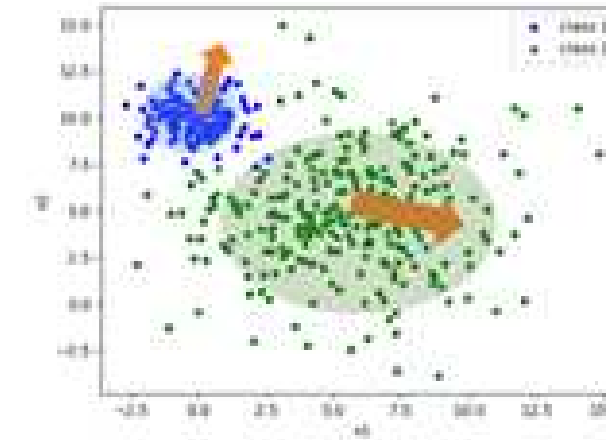
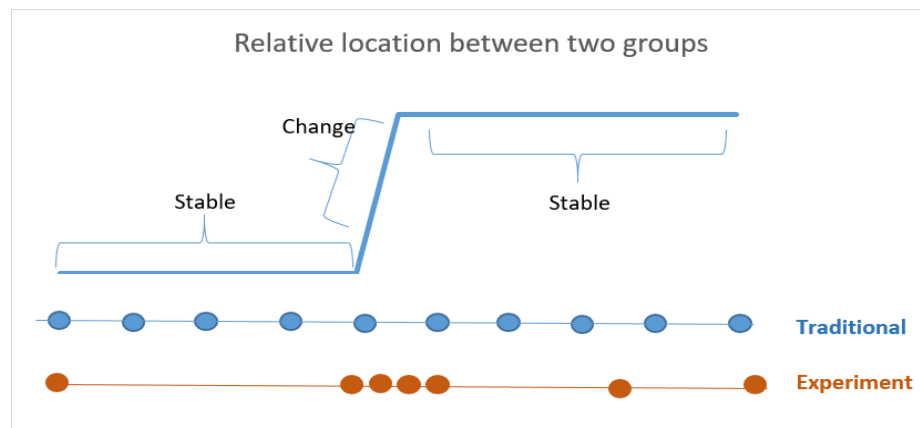
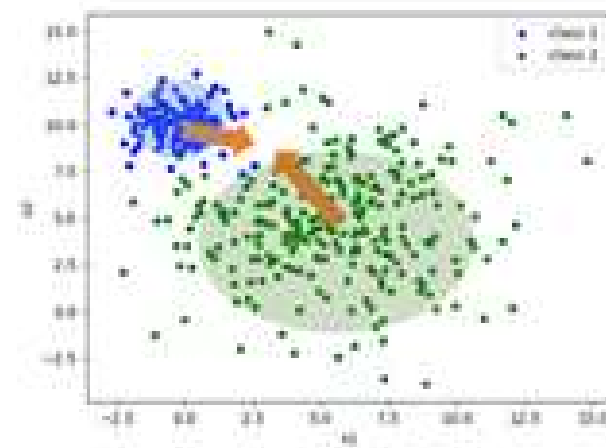
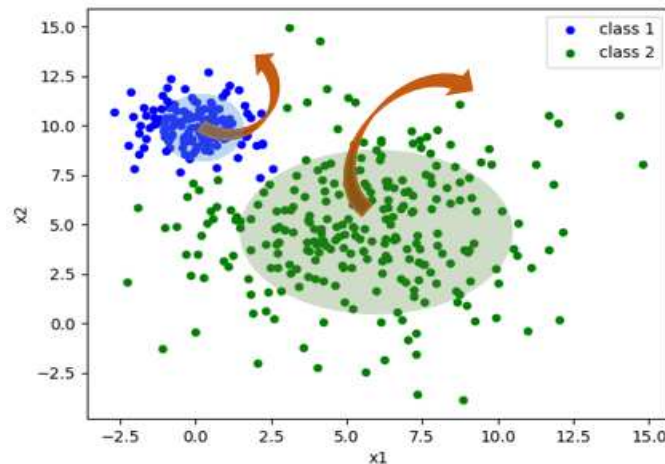
$$\overbrace{P(No \mid d)}^{P_0} = \frac{0.007}{P(d)} \times \frac{1}{SUM} = \frac{0.007}{0.035} = 0.2$$

$$\overbrace{P(Yes \mid d)}^{P_1} = \frac{0.0206}{P(d)} \times \frac{1}{SUM} = \frac{0.028}{0.035} = 0.8$$

Nerve	B.S.	B.P.	Smoke	Care
P.	H	H	F	Y
P.	H	H	T	Y
M.	L	L	T	Y
P.	M	H	F	Y
M.	M	H	T	Y
A.	H	H	F	N
M.	M	H	F	N
M.	L	L	F	N
A.	L	L	T	N
P.	L	L	F	N
M.	M	L	F	N
P.	M	L	T	N
A.	M	H	T	N
A.	H	L	F	N

# Concept Drifting (that causes misclassifications)

- Predict important concept drifting events/time (that can cause misclassifications).
  - Storm prediction, customer preference prediction, market share prediction.



# Moving Data, Moving **PDF**

$$\text{PDF}(\mathbf{x} | \mathbf{C}) = \text{PDF}(\mathbf{x}, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- For a moving data ( $\mathbf{x}_0, \mathbf{x}_1, \dots$ ) of  $n$  predictors with velocity ( $\mathbf{v}_0, \mathbf{v}_1, \dots$ ), the **PDF for class 0**

- $\text{PDF}_{00} = \text{PDF}(\mathbf{x}_0, \mathbf{v}_0, \mu_{00}, \sigma_{00}) = \Pi_{00} \times \exp((\mathbf{x}_0 + \mathbf{t} \times \mathbf{v}_0 - \mathbf{u}_{00})^2 / (-2\sigma_{00}^2)) = \Pi_{00} \times \exp((\mathbf{t} \times \mathbf{v}_0 + \Delta_{00})^2 / (-2\sigma_{00}^2))$ 
  - where  $\Delta_{00} = \mathbf{x}_0 - \mathbf{u}_{00}$ ,  $\Pi_{00} = \frac{1}{\sqrt{2\pi\sigma_{00}^2}}$
  - $\text{PDF}_{00} = \text{PDF}$  for class 0,  $\mu_{00} = \mu$  of predictor 0 under class 0,  $\sigma_{00} = \sigma$  of predictor 0 under class 0
- $\text{PDF}_{10} = \text{PDF}(\mathbf{x}_1, \mathbf{v}_1, \mu_{10}, \sigma_{10}) = \Pi_{10} \times \exp((\mathbf{x}_1 + \mathbf{t} \times \mathbf{v}_1 - \mathbf{u}_{10})^2 / (-2\sigma_{10}^2)) = \Pi_{10} \times \exp((\mathbf{t} \times \mathbf{v}_1 + \Delta_{10})^2 / (-2\sigma_{10}^2))$ 
  - where  $\Delta_{10} = \mathbf{x}_1 - \mathbf{u}_{10}$ ,  $\Pi_{10} = \frac{1}{\sqrt{2\pi\sigma_{10}^2}}$
- $\text{PDF}_{20} = \text{PDF}(\mathbf{x}_2, \mathbf{v}_2, \mu_{20}, \sigma_{20}) = \dots$

$$\sigma_{00} = \frac{1}{N} \sqrt{(N \times SS - LS^2)}$$

- Pre-Norm Probability,  **$\text{PNP}_0 = \text{Pr}_0 \times (\text{PDF}_{00} \times \text{PDF}_{10} \times \dots)$**

$$\text{PNP}_0 = \text{Pr}_0 \times \left( \prod_{i=0}^n \frac{1}{\sqrt{2\sigma_{i,0}^2}} \right) \times \exp( ((\mathbf{t} \times \mathbf{v}_0 + \Delta_{00})^2 / (-2\sigma_{00}^2)) + ((\mathbf{t} \times \mathbf{v}_1 + \Delta_{10})^2 / (-2\sigma_{10}^2)) + \dots )$$

$$= \text{Pr}_0 \times (\Pi_0) \times \exp(\mathbf{A}_0 \times \mathbf{t}^2 + \mathbf{B}_0 \times \mathbf{t} + \mathbf{C}_0) = \text{Pr}_0 \times \prod_{i=0}^n \Pi_{i,0} \times \exp(\mathbf{A}_0 \times \mathbf{t}^2 + \mathbf{B}_0 \times \mathbf{t} + \mathbf{C}_0)$$

$$\bullet ((\mathbf{t} \times \mathbf{v}_0 + \Delta_{00})^2 / (-2\sigma_{00}^2)) + ((\mathbf{t} \times \mathbf{v}_1 + \Delta_{10})^2 / (-2\sigma_{10}^2)) + \dots =$$

$$((\mathbf{v}_0^2 / (-2\sigma_{00}^2)) + (\mathbf{v}_1^2 / (-2\sigma_{10}^2)) + \dots) \times \mathbf{t}^2 + ((\mathbf{v}_0\Delta_{00} / (-2\sigma_{00}^2)) + (\mathbf{v}_1\Delta_{10} / (-2\sigma_{10}^2)) + \dots) \times 2\mathbf{t} + ((\Delta_{00}^2 / (-2\sigma_{00}^2)) + (\Delta_{10}^2 / (-2\sigma_{10}^2)) + \dots)$$

$$= \mathbf{A}_0 \times \mathbf{t}^2 + \mathbf{B}_0 \times 2\mathbf{t} + \mathbf{C}_0$$

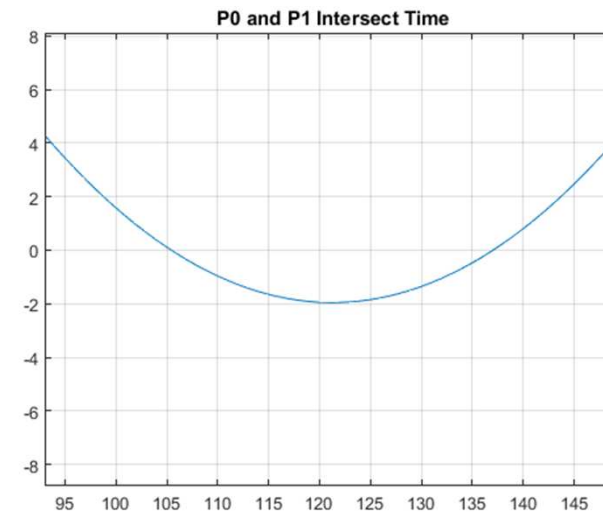
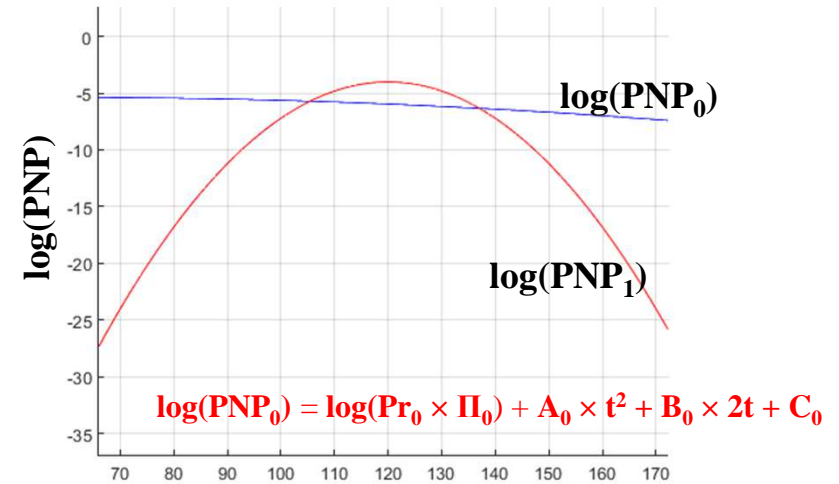
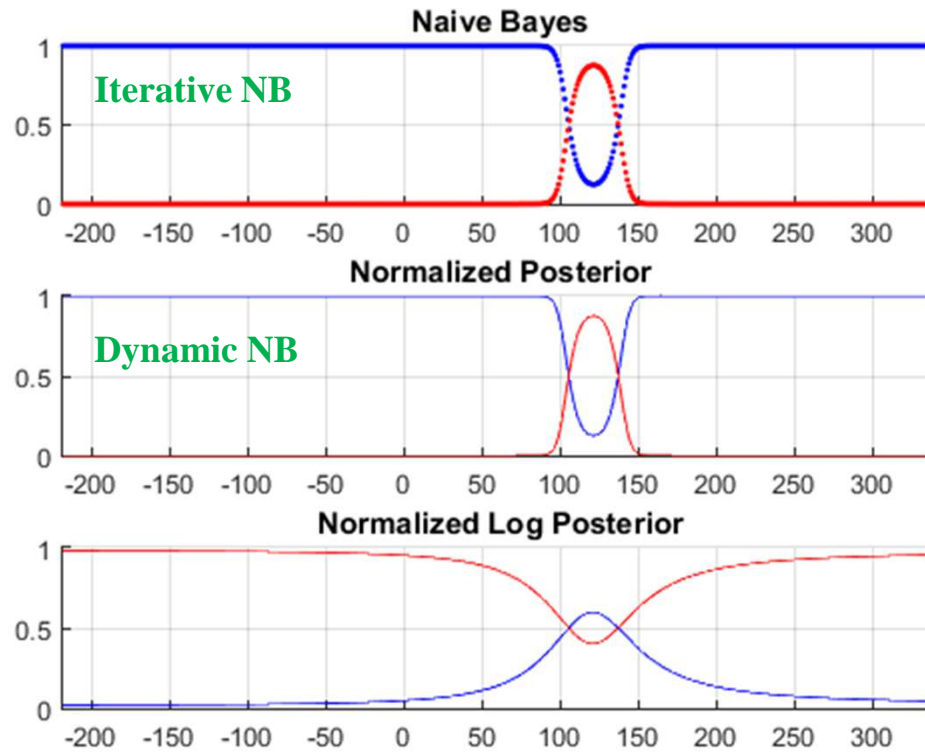


## Normalization

- $\text{PNP}_0 = \text{Pr}_0 \times \Pi_0 \times \exp(\mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0)$
- $\text{PNP}_1 = \text{Pr}_1 \times \Pi_1 \times \exp(\mathbf{A}_1 \times t^2 + \mathbf{B}_1 \times t + \mathbf{C}_1)$
- $\mathbf{P}_0 = \frac{\text{PNP}_0}{\text{PNP}_0 + \text{PNP}_1} = \frac{\text{Pr}_0 \times \Pi_0 \times \exp(\mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0)}{\text{PNP}_0 + \text{PNP}_1}$
- $\mathbf{P}_1 = \frac{\text{PNP}_1}{\text{PNP}_0 + \text{PNP}_1} = \frac{\text{Pr}_1 \times \Pi_1 \times \exp(\mathbf{A}_1 \times t^2 + \mathbf{B}_1 \times t + \mathbf{C}_1)}{\text{PNP}_0 + \text{PNP}_1}$
- We want to know when will  $\mathbf{P}_0 = \mathbf{P}_1 = 0.5$ , so want to know when  $\text{PNP}_0 = \text{PNP}_1$ ?
  - $\text{Pr}_0 \times \Pi_0 \times \exp(\mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0) = \text{Pr}_1 \times \Pi_1 \times \exp(\mathbf{A}_1 \times t^2 + \mathbf{B}_1 \times t + \mathbf{C}_1)$
  - $\log(\text{Pr}_0 \times \Pi_0 \times \exp(\mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0)) = \log(\text{Pr}_1 \times \Pi_1 \times \exp(\mathbf{A}_1 \times t^2 + \mathbf{B}_1 \times t + \mathbf{C}_1))$ 
    - Real PNP of each class =  $\exp(\mathbf{G})$ ,
    - where  $\mathbf{G} = \log(\text{Pr}_0 \times \Pi_0 \times \exp(\mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0)) = \log(\text{Pr}_0 \times \Pi_0) + \mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0$
  - $\log(\text{Pr}_0 \times \Pi_0) + \log(\exp(\mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0)) = \log(\text{Pr}_1 \times \Pi_1) + \log(\exp(\mathbf{A}_1 \times t^2 + \mathbf{B}_1 \times t + \mathbf{C}_1))$
  - $\log(\text{Pr}_0 \times \Pi_0) + \mathbf{A}_0 \times t^2 + \mathbf{B}_0 \times t + \mathbf{C}_0 = \log(\text{Pr}_1 \times \Pi_1) + \mathbf{A}_1 \times t^2 + \mathbf{B}_1 \times t + \mathbf{C}_1$
  - $(\mathbf{A}_0 - \mathbf{A}_1) \times t^2 + (\mathbf{B}_0 - \mathbf{B}_1) \times t + (\mathbf{C}_0 + \log(\text{Pr}_0 \times \Pi_0) - \mathbf{C}_1 - \log(\text{Pr}_1 \times \Pi_1)) = 0$
  - Still a parabola  $\rightarrow \mathbf{A} \times t^2 + \mathbf{B} \times t + \mathbf{C}$
  - $\mathbf{P}_0 = \mathbf{P}_1 = 0.5$  **at time**  $[-\mathbf{B} \pm \text{SQRT}(\mathbf{B}^2 - 4 \times \mathbf{A} \times \mathbf{C})] / 2\mathbf{A}$  when  $\mathbf{B}^2 - 4 \times \mathbf{A} \times \mathbf{C} \geq 0$

# Comparing Results

- Dynamic NB predicted intersect time = [105.436, 137.038]
- $\approx$  Iterative NB time = [105.437, 137.038]



# Dynamic NB for Concept Drifting

