Naïve Bayesian Classifier with R

Camille Hendry, Shulav Neupane, Huaying Qiu

Theory

Bayesian Rule

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

Bayesian Rule

Posterior
Probability
of A

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayesian Rule

Posterior Probability of A

Prior Probability of A

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

Bayesian Rule

$$\underbrace{P(A \mid B)}_{\text{Posterior Probability of A}} = \frac{\underbrace{P(A)}_{\text{Prior Probability of A}}\underbrace{P(B \mid A)}_{\text{Probability of B given A}}}{P(B)}$$

Bayesian Rule

Posterior Probability of A

Prior Probability of A

Probability of B given A

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Probability of B

Settings

The events we are interested in are mutually exclusive and they exhaust the probability space.

For all $A_i \in A$,

$$A_i \cap A_j = \varnothing, \forall i \neq j$$

And,

$$\sum_{i=1}^{k} P(A_i) = 1$$

$$P(A_i \mid B) \leq 1 \quad ?$$

By the law of total probability,

$$P(B) = \sum_{i=1}^{k} P(A_j)P(B \mid A_j)$$

Thus,

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{i=1}^{k} P(A_i)P(B \mid A_i)}$$

Normalizing constant

What are A_i and B ?

A_i represent some class, and B represents some attributes that may affect the probability of A_i

---

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{\sum_{i=1}^{k} P(A_i)P(B | A_i)}$$

Class-Probability Inference

Inference approach based on calculating conditional class probability of the following form

$$P(\underbrace{c = d}_{\text{some } class} \mid a_1 = a_1(x), \underbrace{a_2 = a_2(x)}_{\text{some attributes of } x}, \ldots, a_n = a_n(x))$$

For example, *developing country* can be a class; *GDP, poverty rate, etc.* can be some attributes; *x* can be any country in the world.

Prior Class Probability

We have some training data T.

$$P(c = d) = P_T(c = d) = \frac{|T^d|}{|T|}$$

How to count?

Consider

$$P(a_1 = a_1(x), \ldots, a_n = a_n(x) \,|\, c = d)$$

This is not very useful for classification

Ex. Assume we only have 100 words in our language and we are trying to filter spam emails.

Independence Assumption

Assume

$$P(a_1 = a_1(x), \ldots, a_n = a_n(x) \mid c = d) = \prod_{i=1}^{n} P(a_i = a_i(x) \mid c = d)$$

$$P(a_i = a_i(x) \mid c = d) = \frac{|T^d_{a_i = a_i(x)}|}{|T^d|}$$

Model Construction

The Recipe

$$P(c = d)$$

for each class

$$P(a_i = v_i \mid c = d)$$

for each class, each attribute,
each possible value

Zero Probabilities

If

$$P(a_i = v_i | c = d) = 0$$

Zero Probabilities

If

$$P(a_i = v_i \,|\, c = d) = 0$$

$$P(c = d) \prod_{i=1}^{n} P(a_i = a_i(x) \,|\, c = d) = 0$$

Zero Probabilities

If

$$P(a_i = v_i \,|\, c = d) = 0$$

$$P(c = d)\prod_{i=1}^{n} P(a_i = a_i(x) \,|\, c = d) = 0$$

$$P(c = d \,|\, x) = 0$$

Zero Probabilities

What if

$$P(a_{j_1} = v_{j_1} \,|\, c = d_1) = 0$$

$$P(a_{j_2} = v_{j_2} \,|\, c = d_2) = 0$$

$$\vdots$$

i.e. for every class, there is some value of some attribute that never occurs

Zero Probabilities

The Abyss

$$P(d\,|\,x) = 0$$

for all classes. Therefore, it is impossible to classify x

Two ways around

The ε-method

Define

$$P(a_i = v_i \,|\, c = d) = \begin{cases} \dfrac{|T^d_{a_i=v_i}|}{|T_d|}, \text{ if } T^d_{a_i=v_i} \neq \varnothing \\ \varepsilon, \text{ Otherwise} \end{cases}$$

ε should be considerably less than

$$\frac{1}{|T_d|}$$

m-estimation

Define

$$P(a_i = v_i \,|\, c = d) = \frac{|T^d_{a_i=v_i}| + mp}{|T^d| + m}$$

and let $p = \dfrac{1}{|A_i|}, m = n \in \mathbb{N} \backslash \{0\}$

Missing Attribute

Missing something?
Take it out!

$$P(a_i = v_i \mid c = d) = \frac{|\,T^d_{a_i=v_i}\,|}{|\,T^d\,| - |\,T^d_{a_i=?}\,|}$$

Example

Email Filtering

We have a set of emails

“Bill for your mortgage”	→	SPAM
“Payment for your bill”	→	NOT
“Bill your mortgage”	→	NOT
“Bill for”	→	SPAM
“Payment your mortgage”	→	SPAM
“Payment for your ticket”	→	SPAM

Model

$$\underline{P(c = d) \quad \text{and} \quad P(a_i = v_i \,|\, c = d)}$$

$$P(\text{SPAM}) = \frac{2}{3} \qquad P(\text{NOT}) = \frac{1}{3}$$

	SPAM	NOT
Mortgage	2/4	1/2
Bill	1/4	2/2
Payment	3/4	1/2
For	3/4	1/2
Your	3/4	1/2
Ticket	1/4	0/2

Testing

New email

“Bill for food”

SPAM? NOT?

Testing

$$P(\text{NOT} \mid 0,1,0,1,0,0) = \frac{P(0,1,0,1,0,0 \mid \text{NOT})P(\text{NOT})}{P(0,1,0,1,0,0)}$$

$$\underline{P(0,1,0,1,0,0) = P(0,1,0,1,0,0 \mid \text{SPAM}) + P(0,1,0,1,0,0 \mid \text{NOT})}$$

$$P(\text{SPAM}) = \frac{2}{3} \qquad P(\text{NOT}) = \frac{1}{3}$$

	SPAM	NOT
Mortgage	2/4	1/2
Bill	1/4	2/2
Payment	3/4	1/2
For	3/4	1/2
Your	3/4	1/2
Ticket	1/4	0/2

Testing

$$P(\text{NOT} \mid 0,1,0,1,0,0) = \frac{P(0,1,0,1,0,0 \mid \text{NOT})P(\text{NOT})}{P(0,1,0,1,0,0)}$$

$$P(0,1,0,1,0,0) = P(0,1,0,1,0,0 \mid \text{SPAM}) + P(0,1,0,1,0,0 \mid \text{NOT})$$

$$P(\text{SPAM}) = \frac{2}{3} \qquad P(\text{NOT}) = \frac{1}{3}$$

	SPAM	NOT
Mortgage	2/4	1/2
Bill	1/4	2/2
Payment	3/4	1/2
For	3/4	1/2
Your	3/4	1/2
Ticket	1/4	0/2

Algorithm

1. Calculate probabilities for each attribute conditioned on some class
2. Use the law of total probability to get the joint probability of the attributes
3. Use Bayes rule to calculate the desired probability for the class conditioned on the observed attributes

Testing

$$P(\mathsf{NOT}\,|\,0,1,0,1,0,0) = \frac{P(0,1,0,1,0,0\,|\,\mathsf{NOT})P(\mathsf{NOT})}{P(0,1,0,1,0,0)} = 0.87$$

The Flaw

$$P(\text{NOT} \,|\, \text{Bill for}) = 0.87$$

While

“Bill for” $\longrightarrow$ SPAM

Why?

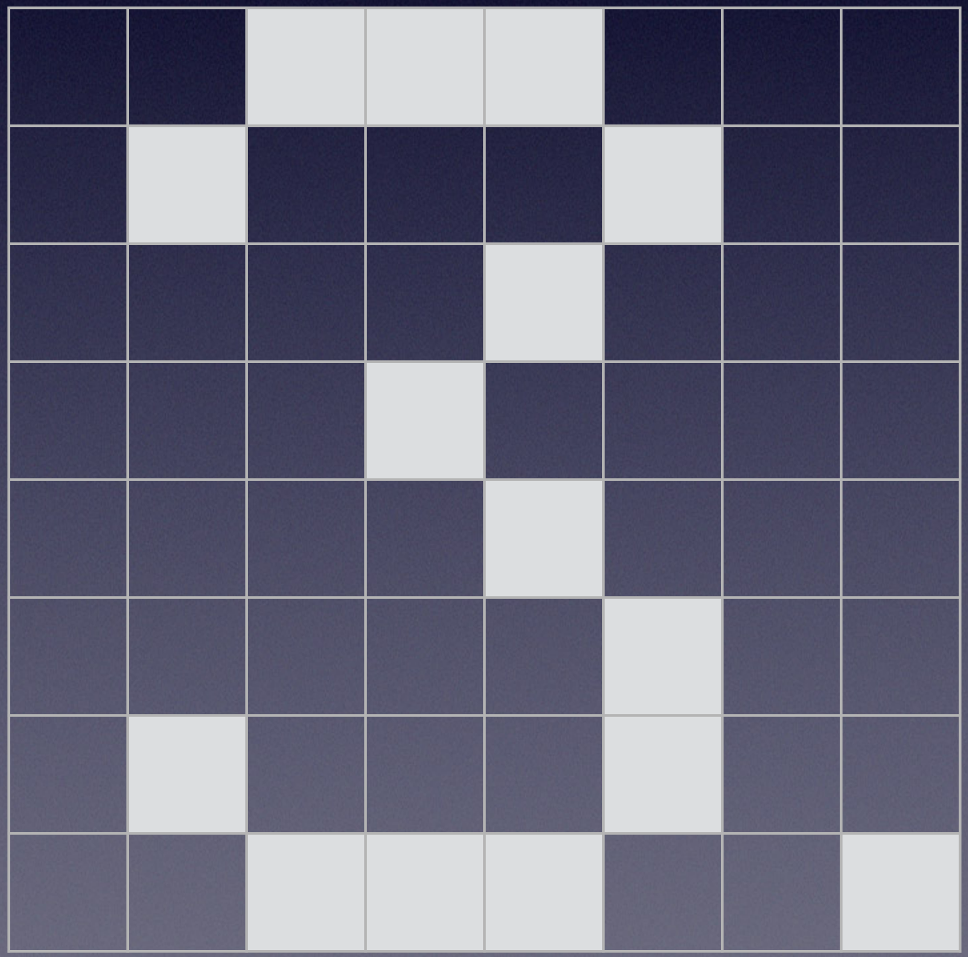
Zero Probability in Practice

P(features,C=2)

$P(C = 2) = 0.1$

P(features,C=3)

$P(C = 3) = 0.1$



Zero Probability in Practice

P(features,C=2)

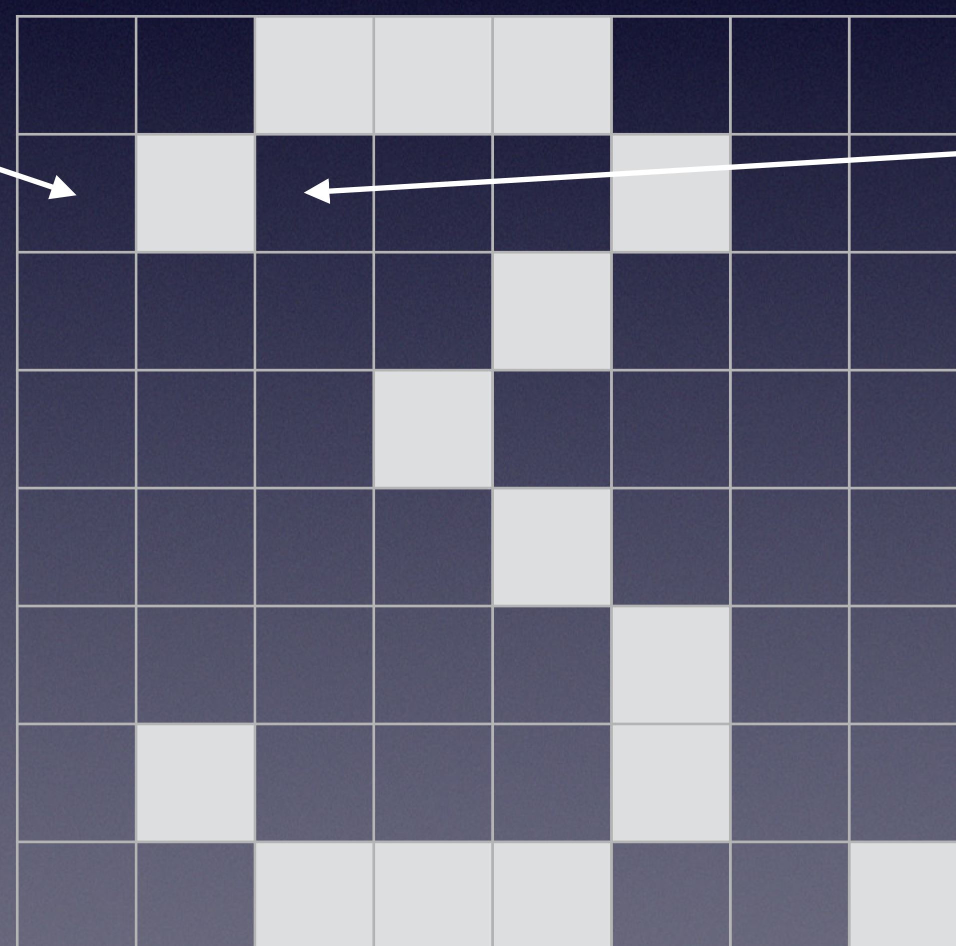
$P(C = 2) = 0.1$

$P(\text{on} \mid C = 2) = 0.8$

P(features,C=3)

$P(C = 3) = 0.1$

$P(\text{on} \mid C = 3) = 0.8$



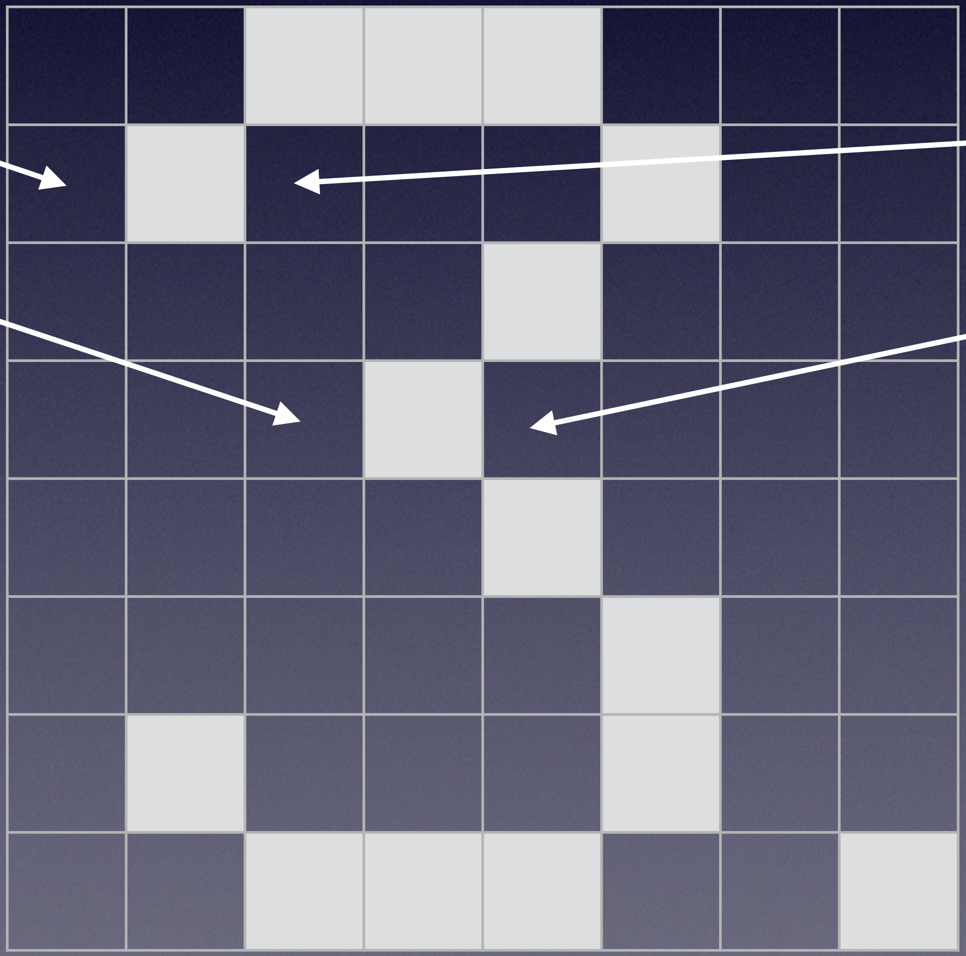
Zero Probability in Practice

P(features,C=2)

$P(C=2)=0.1$

P(features,C=3)

$P(C=3)=0.1$



$P(\text{on} \mid C=2)=0.8$

$P(\text{on} \mid C=2)=0.1$

$P(\text{on} \mid C=3)=0.8$

$P(\text{on} \mid C=3)=0.1$

Zero Probability in Practice

P(features,C=2)

$P(C = 2) = 0.1$

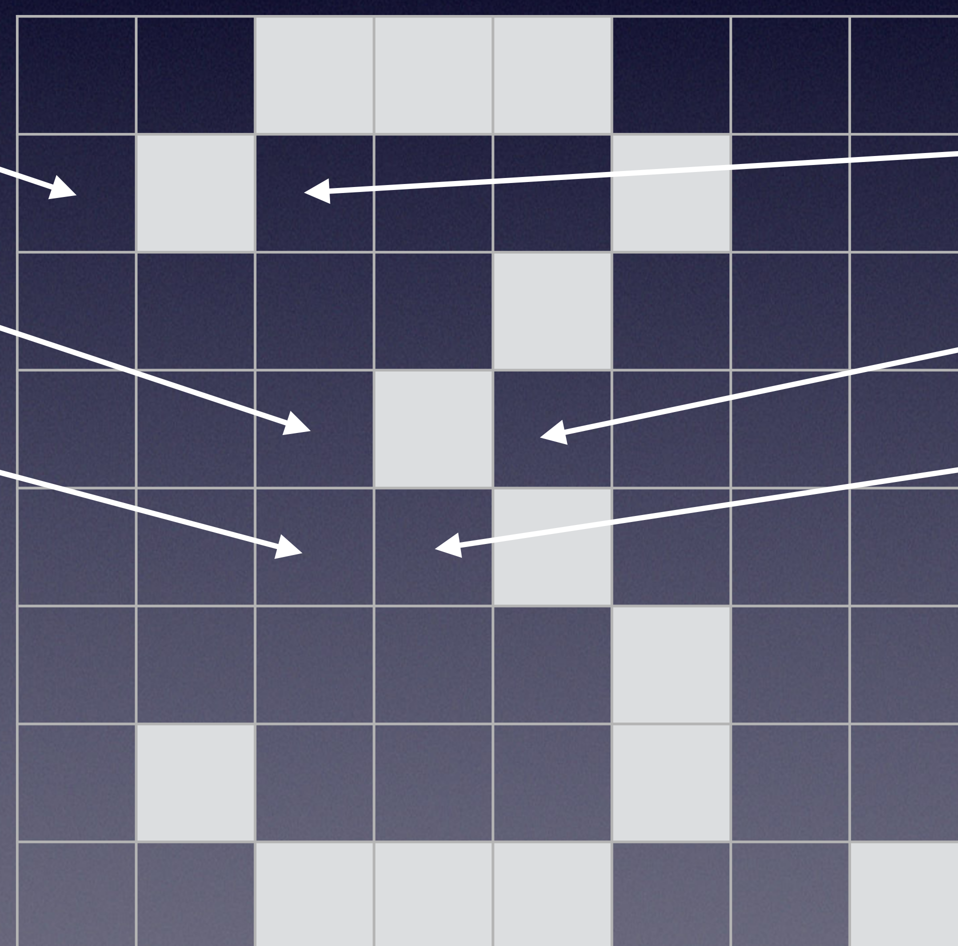
P(features,C=3)

$P(C = 3) = 0.1$

$P(\text{on} \mid C = 2) = 0.8$

$P(\text{on} \mid C = 2) = 0.1$

$P(\text{off} \mid C = 2) = 0.1$



$P(\text{on} \mid C = 3) = 0.8$

$P(\text{on} \mid C = 3) = 0.1$

$P(\text{off} \mid C = 3) = 0.6$

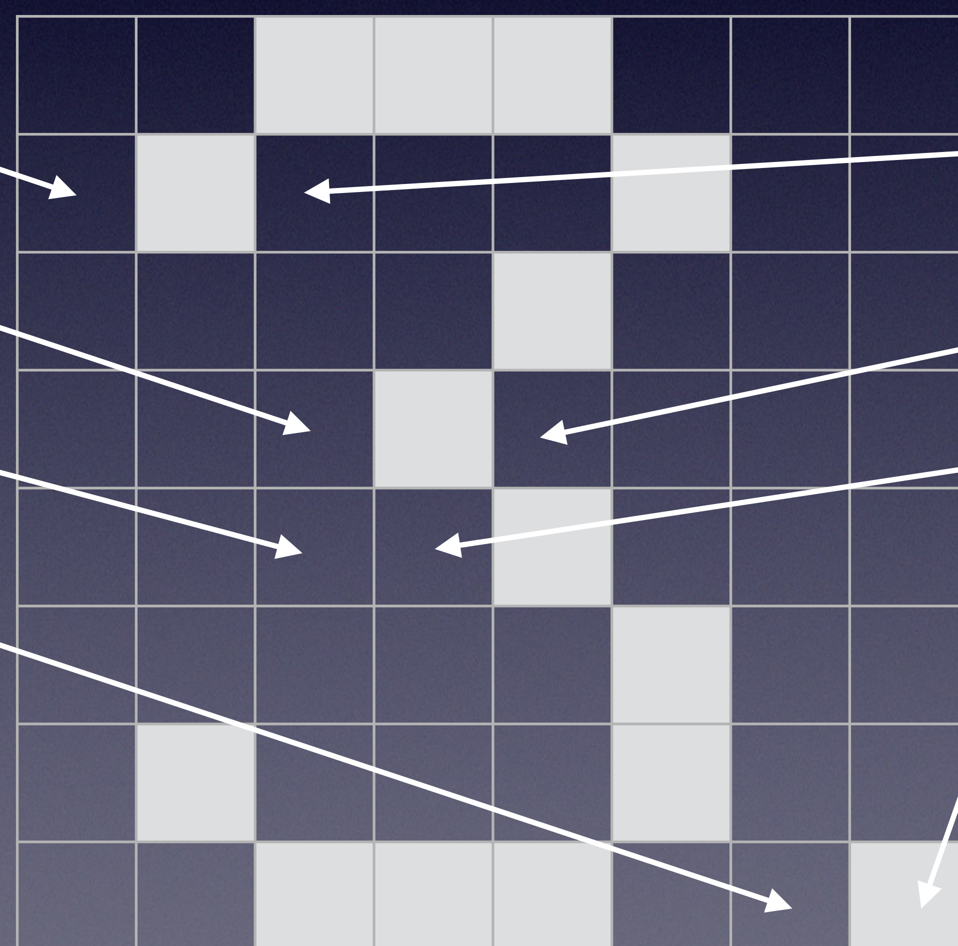
Zero Probability in Practice

P(features,C=2)

$P(C = 2) = 0.1$

P(features,C=3)

$P(C = 3) = 0.1$



$P(\text{on} \mid C = 2) = 0.8$

$P(\text{on} \mid C = 2) = 0.1$

$P(\text{off} \mid C = 2) = 0.1$

$P(\text{on} \mid C = 2) = 0.01$

$P(\text{on} \mid C = 3) = 0.8$

$P(\text{on} \mid C = 3) = 0.1$

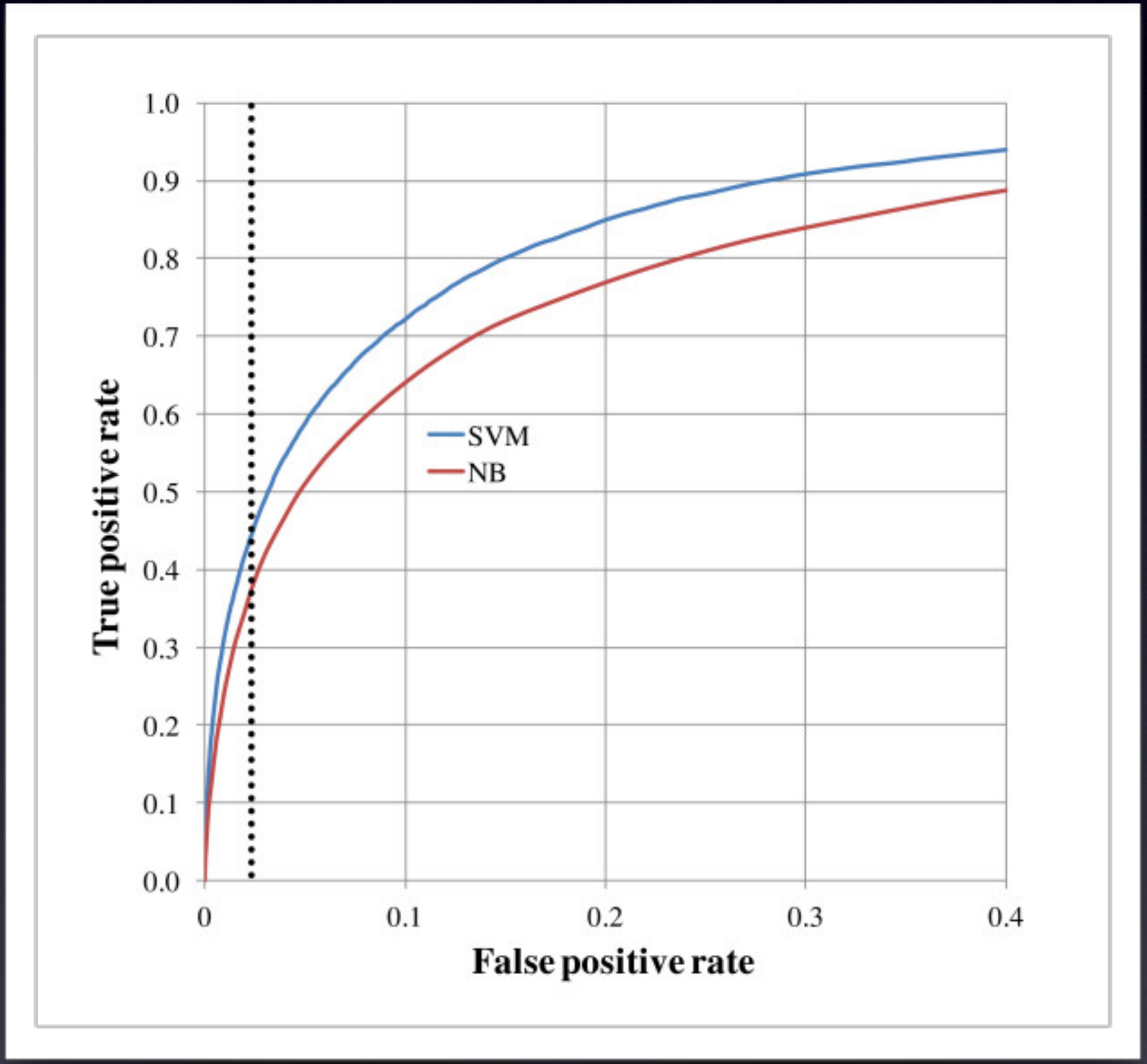
$P(\text{off} \mid C = 3) = 0.6$

$P(\text{on} \mid C = 3) = 0$

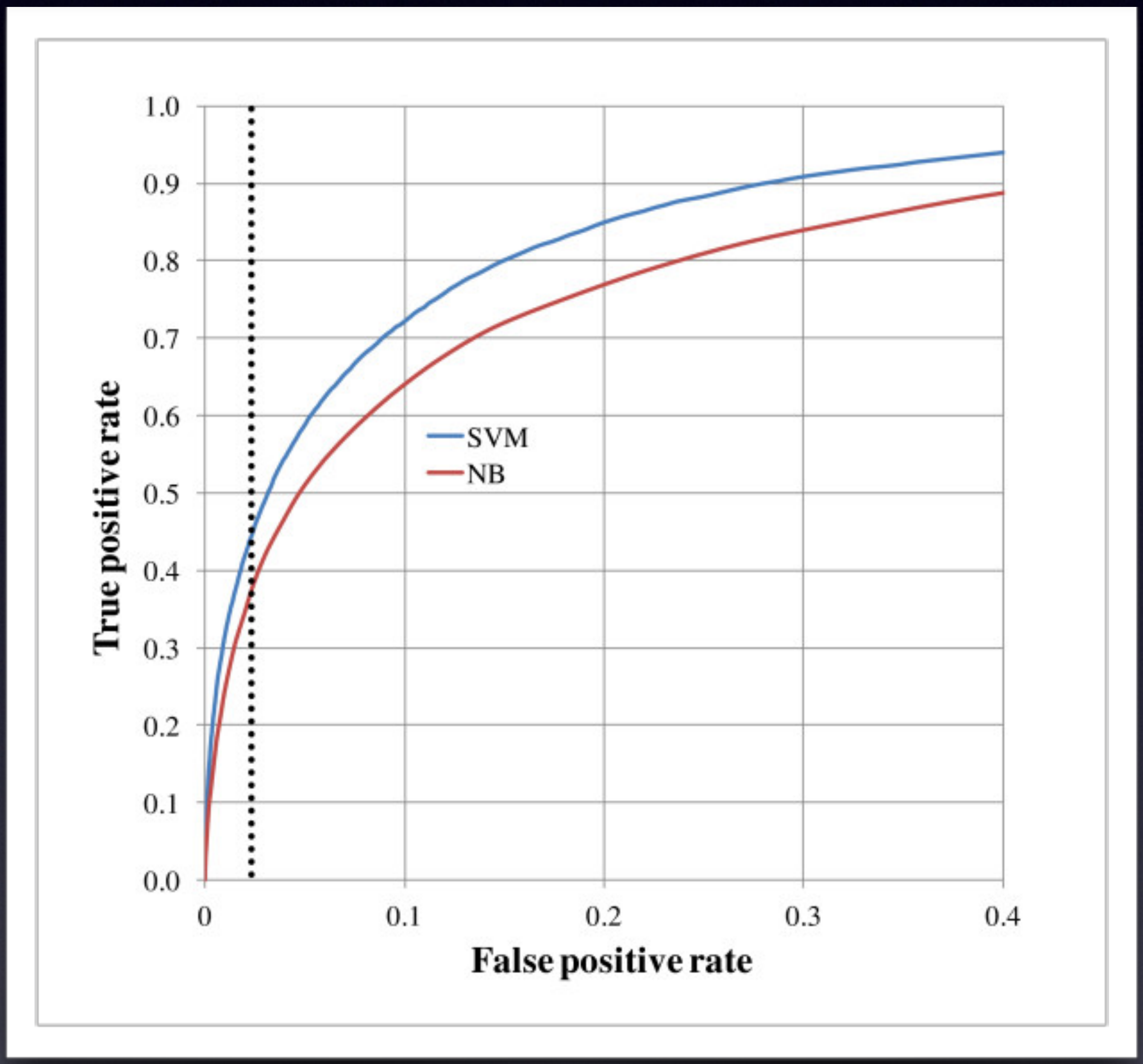
Major Area of Use

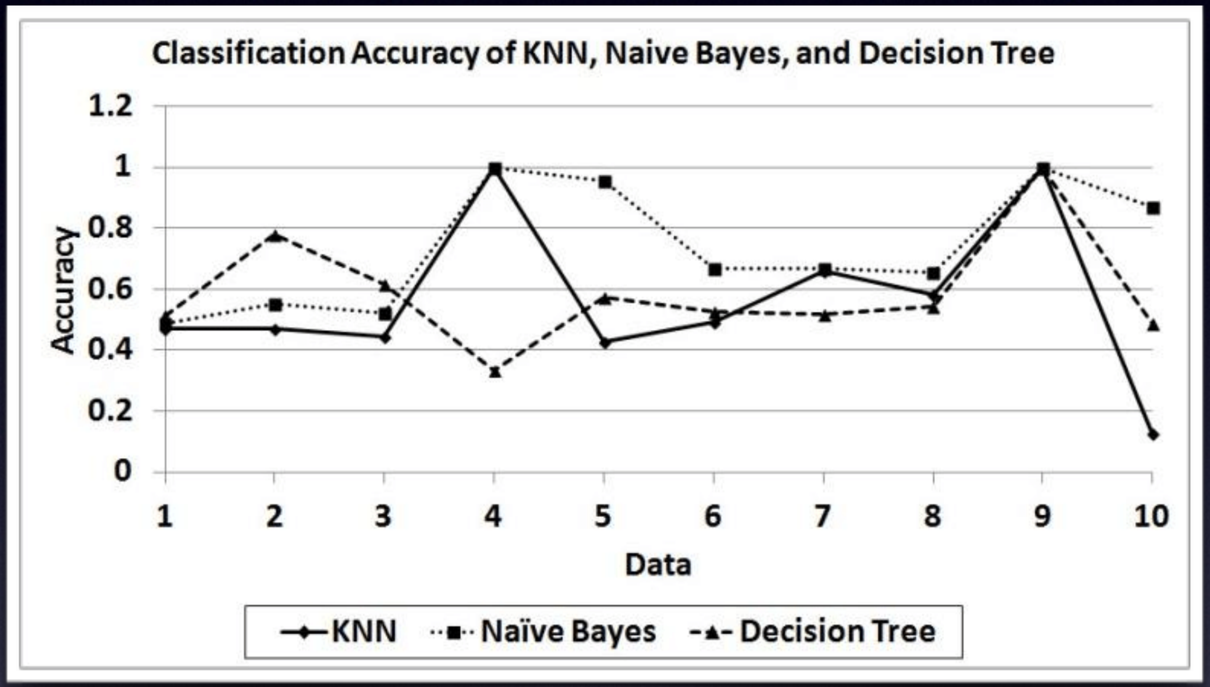
- Text Categorization: Spam filtering, Sentiment Analysis, etc.
- Automatic Medical Diagnosis and other Recommendation Systems.
- Real Time Prediction.

Comparison with Other Methods

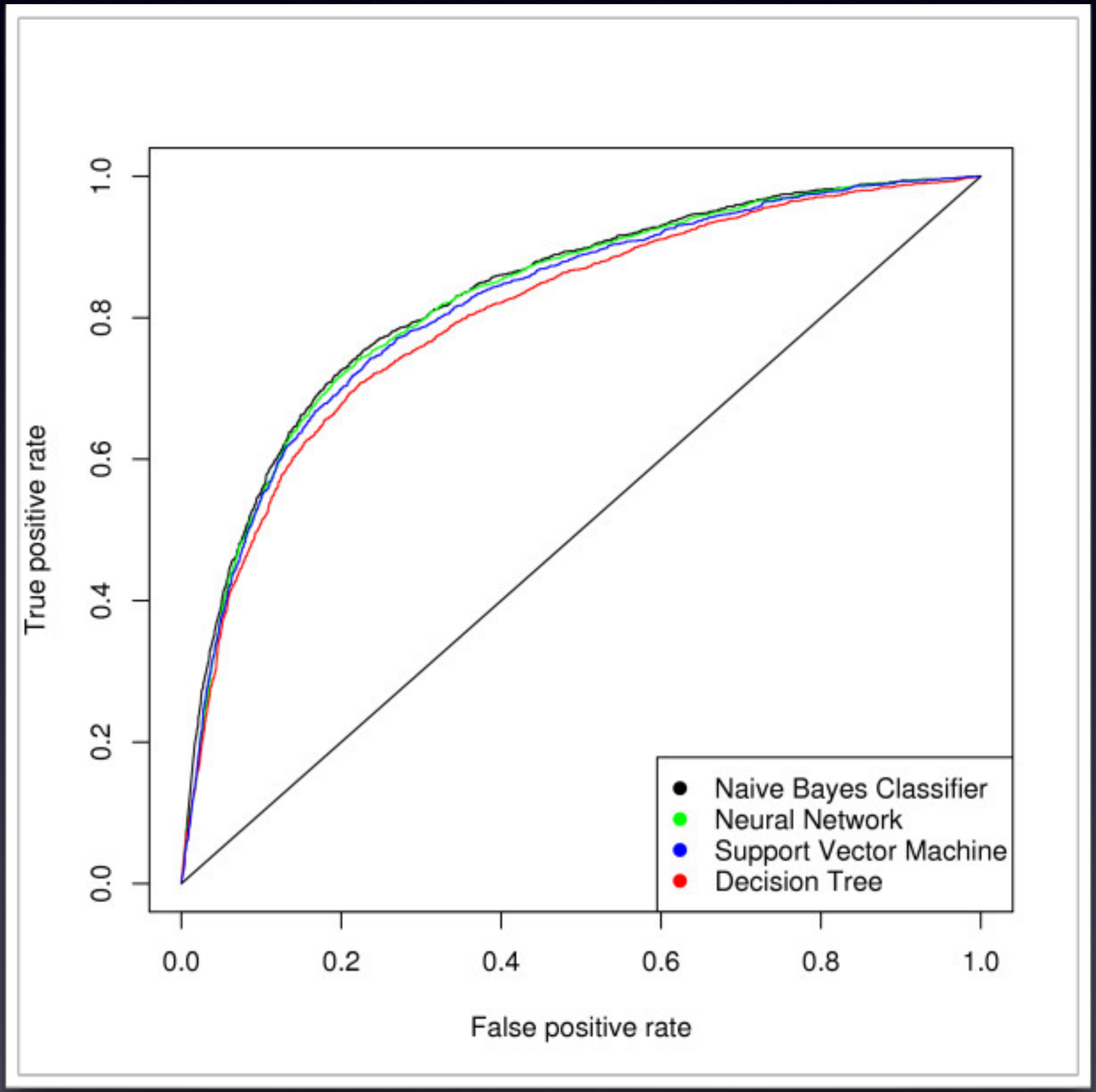


Comparison with Other Methods





Comparison with Other Methods



Comparison with Other Methods

Back at the Titanic

	Model
Score	
92.82	Random Forest
92.82	Decision Tree
87.32	KNN
81.14	Logistic Regression
80.81	Support Vector Machines
80.70	Perceptron
77.10	Naïve Bayes
76.99	Stochastic Gradient Decent

Other Types

- Gaussian Naïve Bayes - continuous variables
- Multinomial Naïve Bayes - vectors represent the frequency of different events
- Bernoulli Naïve Bayes - Boolean data

Pros and Cons

- Algorithm is resource efficient - fast and scales well, good for large datasets.
- When independence assumption holds, it preforms better than logistic regression and less training data is required.
- Zero frequency: Data not observed in training will be assigned a 0.
- Not sensitive to irrelevant attributes.
- Easy to build as it is a simple algorithm.