

Probability and Learning

Review of Probability Theory

Conditional Probability

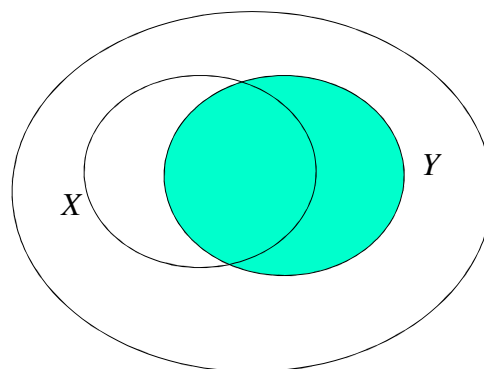
- Probability of an event given the occurrence of some other event.

E.g.,

What is the probability that a liberal Supreme Court Justice will be appointed in the next four years, given that Barack Obama was elected President?

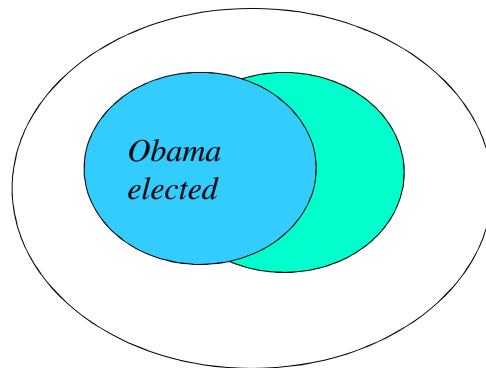
$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

event space



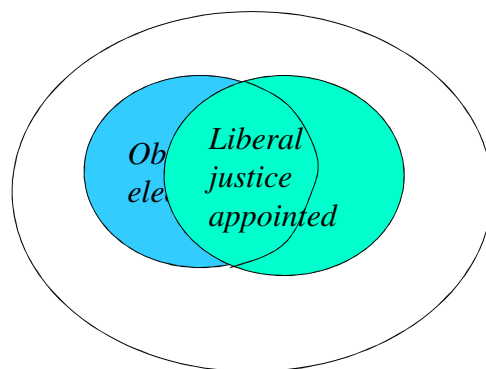
$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

event space



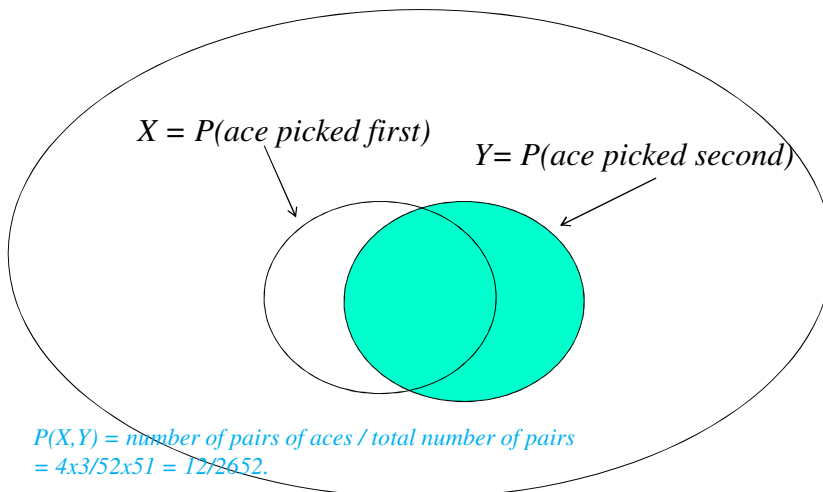
$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$

event space



- Consider choosing a card from a well-shuffled standard deck of 52 playing cards. Given that the first card chosen is an ace, what is the probability that the second card chosen will be an ace?

event space: all possible pairs of cards



$$P(X, Y) = \text{number of pairs of aces} / \text{total number of pairs} \\ = 4 \times 3 / 52 \times 51 = 12 / 2652.$$

$$P(Y) = 4 / 52$$

$$P(X | Y) = (12 / 2652) / (4 / 52) = 3 / 51.$$

**Relationships among joint, conditional,
posterior, and marginal probabilities**

$$P(X | Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$

$$P(X, Y) = P(X | Y)P(Y) = P(Y | X)P(X)$$

Bayes rule :

$$P(X | Y) = \frac{P(Y | X)P(X)}{P(Y)}$$

Multiplication (or “chain”) rule

$$P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2) \dots P(X_n | X_1, X_2, \dots, X_{n-1})$$

Application to Machine Learning

- In machine learning we have a space H of hypotheses:
 h_1, h_2, \dots, h_n
- We also have a set D of data
- We want to calculate $P(h|D)$

Terminology

- **Prior probability of h :**
 - $P(h)$: Probability that hypothesis h is true given our prior knowledge
 - If no prior knowledge, all $h \in H$ are equally probable
- **Posterior probability of h :**
 - $P(h|D)$: Probability that hypothesis h is true, given the data D .
- **Likelihood of D :**
 - $P(D|h)$: Probability that we will see data D , given hypothesis h is true.

Bayes rule says:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

The Monty Hall Problem

You are a contestant on a game show.

There are 3 doors, A, B, and C. There is a new car behind one of them and goats behind the other two.

Monty Hall, the host, asks you to pick a door, any door. You pick door A.

Monty tells you he will open a door, different from A, that has a goat behind it. He opens door B: behind it there is a goat.

Monty now gives you a choice: Stick with your original choice A or switch to C.

Should you switch?

<http://math.ucsd.edu/~crypto/Monty/monty.html>

Bayesian probability formulation

Hypothesis space H :

h_1 = Car is behind door A

h_2 = Car is behind door B

h_3 = Car is behind door C

Data D = You chose A; Monty opened B, showed goat.

What is $P(h_1 | D)$?

What is $P(h_2 | D)$?

What is $P(h_3 | D)$?

Bayesian probability formulation

Hypothesis space H :

h_1 = Car is behind door A

h_2 = Car is behind door B

h_3 = Car is behind door C

Data D = You chose A; Monty opened B, showed goat.

What is $P(h_1 | D)$?

What is $P(h_2 | D)$?

What is $P(h_3 | D)$?

$$P(h_1 | D) = \frac{P(D | h_1)P(h_1)}{P(D)}$$

$$P(h_2 | D) = 0$$

$$P(h_3 | D) = \frac{P(D | h_3)P(h_3)}{P(D)}$$

$$P(h_1) = 1/3, P(h_3) = 1/3$$

$$P(D | h_1) = 1/2 \quad (\text{if car is behind door A, then Monty can open B or C})$$

$$P(D | h_3) = 1 \quad (\text{if car is behind door C, then Monty can open B})$$

$$\begin{aligned} P(D) &= P(h_1)P(D | h_1) + P(h_2)P(D | h_2) + P(h_3)P(D | h_3) \\ &= (1/3) \cdot (1/2) + (1/3) \cdot 0 + (1/3) \cdot 1 = 1/2 \end{aligned}$$

Thus:

$$P(h_1 | D) = \frac{P(D | h_1)P(h_1)}{P(D)} = \frac{(1/2) \cdot (1/3)}{(1/2)} = 1/3$$

$$P(h_3 | D) = \frac{P(D | h_3)P(h_3)}{P(D)} = \frac{(1) \cdot (1/3)}{(1/2)} = 2/3$$

Another example

It is known that about 0.0038 of the U.S. population is infected with HIV.

The current test for HIV has a 0.005 probability of returning a false positive, and a 0.002 probability of returning a false negative.

Joe takes the test; the result is positive.

Let h_1 be the hypothesis that Joe is actually HIV positive and h_2 be the hypothesis that Joe is actually HIV negative. Which of these is the maximum *a posteriori* hypothesis h_{MAP} , where

$$h_{MAP} = \arg \max_{h \in H} P(h | data) = \arg \max_{h \in H} \frac{P(data | h)P(h)}{P(data)} = \arg \max_{h \in H} P(data | h)P(h)$$

Given the information above, what is the probability that Joe is HIV positive?

$$P(data | h^+)P(h^+) = (.998) \cdot (.0038) = .00379$$

$$P(data | h^-)P(h^-) = (.005)(.9962) = .004981$$

$$\begin{aligned} P(data) &= P(data | h^+)P(h^+) + P(data | h^-)P(h^-) \\ &= (.998)(.0038) + (.005)(.9962) = .00877 \end{aligned}$$

$$P(h^+ | D) = \frac{P(D | h^+)P(h^+)}{P(D)} = \frac{.00379}{.00877} = .43$$

$$P(h^- | D) = \frac{P(D | h^-)P(h^-)}{P(D)} = \frac{.004981}{.00877} = .57$$

Another Example

Suppose you receive an e-mail message with the subject "Hi". You have been keeping statistics on your e-mail, and have found that while only 10% of the total e-mail messages you receive are spam, 50% of the spam messages have the subject "Hi", and 2% of the non-spam messages have the subject "Hi". What is the probability this message is spam?

$$P(\text{spam} | \text{"Hi"}) = \frac{P(\text{"Hi"} | \text{spam})P(\text{spam})}{P(\text{"Hi"})} = \frac{(.5)(.1)}{(.5)(.1) + (.02)(.9)} = .735$$

Naive Bayes Classifier

Let $f(\mathbf{x})$ be a target function for classification: $f(\mathbf{x}) \in \{+, -\}$.

Let $\mathbf{x} = \langle a_1, a_2, \dots, a_n \rangle$

We want to find the most probable hypothesis
given the data \mathbf{x} :

$$\begin{aligned} h_{\text{MAP}} & \text{ (maximum } a \text{ posteriori hypothesis)} \\ & = \operatorname{argmax}_{class \in \{+, -\}} P(\text{class} \mid \mathbf{x}) \\ & = \operatorname{argmax}_{class \in \{+, -\}} P(\text{class} \mid a_1, a_2, \dots, a_n) \end{aligned}$$

By Bayes Theorem:

$$\begin{aligned} h_{\text{MAP}} & = \operatorname{argmax}_{class \in \{+, -\}} \frac{P(a_1, a_2, \dots, a_n \mid \text{class})P(\text{class})}{P(a_1, a_2, \dots, a_n)} \\ & = \operatorname{argmax}_{class \in \{+, -\}} P(a_1, a_2, \dots, a_n \mid \text{class})P(\text{class}) \end{aligned}$$

$P(\text{class})$ can be estimated from the training data. How?

How about estimating $P(a_1, a_2, \dots, a_n \mid \text{class})$ from training data? Would this work?

- Naive Bayes classifier: Assume each attribute is conditionally independent, given *class*.

$$P(a_1, a_2, \dots, a_n | class) = P(a_1 | class)P(a_2 | class) \cdots P(a_n | class)$$

Given this assumption, here's how to classify an instance

$$\mathbf{x} = \langle a_1, a_2, \dots, a_n \rangle:$$

For each a_i , estimate $P(a_i | class)$ from training data.

For new instance \mathbf{x} :

$$h_{NB} = \operatorname{argmax}_{class \in \{+, -\}} P(class) \prod_i P(a_i | class)$$

Example 1

Suppose you have a binary classification problem in which instances \mathbf{x} have three attributes: $\mathbf{x} = (a_1, a_2, a_3)$, where $a_i \in \{-, +\}$. Given the training set below, show how a naive Bayes classification algorithm would classify the new instance $\mathbf{x} = (0, 0, 0)$.

Training set:

$$\mathbf{x}_1 = (1, 0, 0), \text{ class} = +$$

$$\mathbf{x}_2 = (0, 1, 1), \text{ class} = +$$

$$\mathbf{x}_3 = (1, 1, 0), \text{ class} = -$$

$$\mathbf{x}_4 = (0, 0, 1), \text{ class} = -$$

$$\mathbf{x}_5 = (1, 1, 1), \text{ class} = +$$

$$h_{NB} = \operatorname{argmax}_{class \in \{+, -\}} P(class) \prod_i P(a_i | class)$$

$\mathbf{x}_1 = (1, 0, 0)$, class = +
 $\mathbf{x}_2 = (0, 1, 1)$, class = +
 $\mathbf{x}_3 = (1, 1, 0)$, class = -
 $\mathbf{x}_4 = (0, 0, 1)$, class = -
 $\mathbf{x}_5 = (1, 1, 1)$, class = +

$$\begin{aligned}
 P(a_1 = 0 | +) &= \frac{1}{3} & P(a_1 = 1 | +) &= \frac{2}{3} \\
 P(a_1 = 0 | -) &= \frac{1}{2} & P(a_1 = 1 | -) &= \frac{1}{2} \\
 P(a_2 = 0 | +) &= \frac{1}{3} & P(a_2 = 1 | +) &= \frac{2}{3} \\
 P(a_2 = 0 | -) &= \frac{1}{2} & P(a_2 = 1 | -) &= \frac{1}{2} \\
 P(a_3 = 0 | +) &= \frac{1}{3} & P(a_3 = 1 | +) &= \frac{2}{3} \\
 P(a_3 = 0 | -) &= \frac{1}{2} & P(a_3 = 1 | -) &= \frac{1}{2}
 \end{aligned}$$

$$\begin{aligned}
 P(a_1 = 0, a_2 = 0, a_3 = 0 | +) &\propto P(+)^{\frac{3}{5}} P(a_1 = 0 | +)^{\frac{1}{3}} P(a_2 = 0 | +)^{\frac{1}{3}} P(a_3 = 0 | +)^{\frac{1}{3}} = \frac{1}{45} \\
 P(a_1 = 0, a_2 = 0, a_3 = 0 | -) &\propto P(-)^{\frac{2}{5}} P(a_1 = 0 | -)^{\frac{1}{2}} P(a_2 = 0 | -)^{\frac{1}{2}} P(a_3 = 0 | -)^{\frac{1}{2}} = \frac{1}{2}
 \end{aligned}$$

$h_{NB}(0,0,0) = -$

Student presentations

Example 2

- Use training data from decision tree example to classify the new instance:

<Outlook=sunny, Temperature=cool, Humidity=high,
Wind=strong>

| <u>Day</u> | <u>Outlook</u> | <u>Temp</u> | <u>Humidity</u> | <u>Wind</u> | <u>PlayTennis</u> |
|------------|----------------|-------------|-----------------|-------------|-------------------|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

$$h_{NB} = \operatorname{argmax}_{class \in \{+, -\}} P(class) \prod_i P(a_i | class)$$

$$\begin{array}{lll}
 P(\text{outlook} = \text{sunny} | +) = \frac{2}{9} & P(\text{outlook} = \text{overcast} | +) = \frac{4}{9} & P(\text{outlook} = \text{rain} | +) = \frac{3}{9} \\
 P(\text{outlook} = \text{sunny} | -) = \frac{3}{5} & P(\text{outlook} = \text{overcast} | -) = \frac{0}{5} & P(\text{outlook} = \text{rain} | -) = \frac{2}{5} \\
 P(\text{temperature} = \text{hot} | +) = \frac{2}{9} & P(\text{temperature} = \text{mild} | +) = \frac{4}{9} & P(\text{temperature} = \text{cool} | +) = \frac{3}{9} \\
 P(\text{temperature} = \text{hot} | -) = \frac{2}{5} & P(\text{temperature} = \text{mild} | -) = \frac{2}{5} & P(\text{temperature} = \text{cool} | -) = \frac{1}{5} \\
 P(\text{humidity} = \text{high} | +) = \frac{3}{9} & P(\text{humidity} = \text{normal} | +) = \frac{6}{9} & \\
 P(\text{humidity} = \text{high} | -) = \frac{4}{5} & P(\text{humidity} = \text{normal} | -) = \frac{1}{5} & \\
 P(\text{wind} = \text{strong} | +) = \frac{3}{9} & P(\text{wind} = \text{weak} | +) = \frac{6}{9} & \\
 P(\text{wind} = \text{strong} | -) = \frac{3}{5} & P(\text{wind} = \text{weak} | -) = \frac{2}{5} &
 \end{array}$$

$$\begin{aligned}
 &P(+ | \text{outlook} = \text{sunny}, \text{temperature} = \text{cool}, \text{humidity} = \text{high}, \text{wind} = \text{strong}) \\
 &\propto P(+)P(\text{outlook} = \text{sunny} | +)P(\text{temperature} = \text{cool} | +)P(\text{humidity} = \text{high} | +)P(\text{wind} = \text{strong} | +) \\
 &= \left(\frac{9}{14}\right)\left(\frac{2}{9}\right)\left(\frac{3}{9}\right)\left(\frac{3}{9}\right)\left(\frac{3}{9}\right) = .00529
 \end{aligned}$$

$$\begin{aligned}
 &P(- | \text{outlook} = \text{sunny}, \text{temperature} = \text{cool}, \text{humidity} = \text{high}, \text{wind} = \text{strong}) \\
 &\propto P(-)P(\text{outlook} = \text{sunny} | -)P(\text{temperature} = \text{cool} | -)P(\text{humidity} = \text{high} | -)P(\text{wind} = \text{strong} | -) \\
 &= \left(\frac{5}{14}\right)\left(\frac{3}{5}\right)\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)\left(\frac{3}{5}\right) = .02057
 \end{aligned}$$

$$h_{NB}(\text{outlook} = \text{sunny}, \text{temperature} = \text{cool}, \text{humidity} = \text{high}, \text{wind} = \text{strong}) = \text{no}$$

Estimating probabilities

- **Recap:** In previous example, we had a training set and a new example,
<Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong>

We asked: What classification is given by a naive Bayes classifier?

- We needed to estimate

$$P(\text{Temperature} = \text{high} \mid \text{PlayTennis} = \text{yes}) \quad P(\text{Temperature} = \text{high} \mid \text{PlayTennis} = \text{no})$$

$$P(\text{Temperature} = \text{mild} \mid \text{PlayTennis} = \text{yes}) \quad P(\text{Temperature} = \text{mild} \mid \text{PlayTennis} = \text{no})$$

$$P(\text{Temperature} = \text{cool} \mid \text{PlayTennis} = \text{yes}) \quad P(\text{Temperature} = \text{cool} \mid \text{PlayTennis} = \text{no})$$

- Using frequencies of occurrence for different events, as in the previous example, can be inaccurate if there are a small number of examples in the event space.
- E.g., suppose we didn't have training example D6. Then
 $P(\text{temperature}=\text{cool} \mid \text{PlayTennis} = \text{no}) = 0$.

- Now suppose we want to classify our new instance:
<Outlook=sunny, Temperature=cool, Humidity=high,
Wind=strong>. Then:

$$P(\text{no}) \prod_i P(a_i \mid \text{no}) = 0$$

This incorrectly gives us zero probability due to sampling error on one attribute.

Here is one solution: m-estimate of probabilities

- Let n be the number of training examples with class c . For attribute a , let n_v be the number of training examples with class c and $a = v$.

- We defined
$$P(a = v \mid \text{class} = c) \approx \frac{n_v}{n_c}$$

- Define **m-estimate of probability** as:

$$P(a = v \mid \text{class} = c) \approx \frac{n_v + mp}{n_c + m}$$

where p = prior estimate of $P(a = v \mid \text{class} = c)$ and m is the weight given to p .

- The idea here is to pretend that we have an extra m training examples with class = c , with $p \times m$ of them having $a = v$.
- Usually, unless you have good prior information, set $p = 1/(\text{number of values of attribute } a)$. m is usually set to 2.

- Example:

$$P(\text{humidity} = \text{normal} \mid \text{no}) = 1/5 = .2$$

- Let $p = 1/(\text{number of values of attribute humidity}) = 1/2$
- Now, pretend we had two additional training instances with class no , one half of them having humidity=normal.

- **m-estimate with $m = 2$:**

$$P(\text{humidity} = \text{normal} \mid \text{no}) = \frac{1 + 2\left(\frac{1}{2}\right)}{5 + 2} = 2/7 = .28$$

Naive Bayes with continuous data

- How to deal with continuous-valued attributes?
- Example: Suppose your training examples had the following values for humidity:

Humidity: 5 23 55 60 72 85

– Need to discretize!

Simplest discretization method

For each attribute A_i , create k equal-length bins in interval from A_i^{min} to A_i^{max} .

E.g., from previous example, $\text{Humidity}^{min} = 5$, $\text{Humidity}^{max} = 85$,

Suppose $k = 4$ bins: Then bin size is 20



Questions: What should k be? What if some bins have very few instances?

Problem with balance between *discretization bias* and *variance*.

The more bins, the lower the bias, but the higher the variance, due to small sample size.

Alternative simple (but effective)
discretization method
(Yang & Webb, 2001)

Let n = number of training examples. For each attribute A_i , create $\approx \sqrt{n}$ bins. Sort values of A_i in ascending order, and put $\approx \sqrt{n}$ of them in each bin.

Don't need m-estimate of probability

E.g., suppose you have 1000 training examples. You would use $\sqrt{1000} \approx 32$ bins, and put ≈ 32 examples in each bin.

This gives good balance between discretization bias and variance.

Beyond Independence: Conditions for the
Optimality of the Simple Bayesian Classifier
(P. Domingos and M. Pazzani)

- Recap of naive Bayes classifier:

Let instance $\mathbf{x} = \langle a_1, a_2, \dots, a_n \rangle$. Let the possible classes for x be $V = \{v_1, v_2, \dots, v_n\}$. Then:

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

Called "naive" because it assumes attributes are independent of one another.

- This paper asks: why does the naive (“simple”) Bayes classifier, SBC, do so well in domains with clearly dependent attributes?

Experiments

- Compare five classification methods on 30 data sets from the UCI ML database.

SBC = Simple Bayesian Classifier

Default = “Choose class with most representatives in data”

C4.5 = Quinlan’s decision tree induction system

PEBLS = An instance-based learning system

CN2 = A rule-induction system

- For SBC, numeric values were discretized into ten equal-length intervals.

| Domain | SBC | Default | C4.5 | PEBLs | CN2 |
|------------------|-----------|------------------------|------------------------|------------------------|------------------------|
| Audiology | 73.9±5.3 | 21.3±2.6 ¹ | 72.5±5.8 ⁶ | 75.8±5.4 ⁴ | 71.0±5.1 ² |
| Annealing | 93.5±2.7 | 76.4±1.8 ¹ | 91.3±2.3 ³ | 98.7±0.9 ¹ | 81.2±5.4 ¹ |
| Breast cancer | 68.7±5.4 | 67.6±7.6 ⁶ | 70.1±5.6 ⁴ | 65.8±4.7 ³ | 67.9±7.1 ⁶ |
| Credit screening | 85.2±1.7 | 57.4±3.8 ¹ | 85.0±2.0 ⁶ | 81.3±2.0 ¹ | 82.0±2.2 ¹ |
| Chess endgames | 88.0±1.4 | 52.0±1.9 ¹ | 99.2±0.1 ¹ | 96.9±0.7 ¹ | 98.1±1.0 ¹ |
| Pima diabetes | 74.4±3.0 | 66.0±2.3 ¹ | 72.4±2.8 ⁴ | 71.4±2.4 ¹ | 73.8±2.7 ⁶ |
| Echocardiogram | 66.7±7.4 | 67.8±6.6 ⁶ | 65.8±6.2 ⁶ | 64.1±6.1 ⁵ | 68.2±7.2 ⁶ |
| Glass | 50.4±15.9 | 31.7±5.5 ¹ | 66.1±8.4 ¹ | 65.8±7.3 ¹ | 63.8±5.5 ¹ |
| Heart disease | 83.1±3.2 | 55.0±3.4 ¹ | 74.2±4.2 ¹ | 79.2±3.8 ¹ | 79.7±2.9 ¹ |
| Hepatitis | 81.2±3.7 | 78.1±3.1 ² | 78.7±4.7 ⁴ | 79.9±6.6 ⁶ | 80.3±4.2 ⁶ |
| Horse colic | 77.8±4.2 | 63.6±3.0 ¹ | 83.6±4.1 ¹ | 76.3±4.4 ⁵ | 82.5±4.2 ¹ |
| Thyroid disease | 97.3±0.7 | 95.3±0.6 ¹ | 99.1±0.2 ¹ | 97.3±0.4 ⁶ | 98.8±0.4 ¹ |
| Iris | 89.0±12.8 | 26.5±5.2 ¹ | 93.4±2.4 ⁵ | 91.7±3.7 ⁶ | 93.3±3.6 ⁵ |
| Labor neg. | 92.6±7.9 | 65.0±9.5 ¹ | 79.7±7.1 ¹ | 91.6±4.3 ⁶ | 82.1±6.9 ¹ |
| Lung cancer | 46.4±14.7 | 26.8±12.3 ¹ | 40.9±16.3 ⁶ | 42.3±17.3 ⁶ | 38.6±13.5 ⁴ |
| Liver disease | 61.8±6.9 | 58.1±3.4 ³ | 63.7±4.3 ⁶ | 60.1±3.6 ⁶ | 65.0±3.8 ⁴ |
| LED | 66.8±5.9 | 8.0±2.7 ¹ | 61.2±8.4 ² | 55.3±6.1 ¹ | 58.6±8.1 ¹ |
| Lymphography | 81.5±5.6 | 57.3±5.4 ¹ | 75.3±4.8 ¹ | 82.9±5.6 ⁶ | 78.8±4.9 ³ |
| Post-operative | 61.8±9.8 | 71.2±5.2 ¹ | 70.2±4.9 ¹ | 58.8±8.1 ⁶ | 60.8±8.2 ⁶ |
| Promoters | 87.6±6.0 | 43.1±4.2 ¹ | 74.3±7.8 ¹ | 91.7±5.9 ¹ | 75.9±8.8 ¹ |
| Primary tumor | 44.9±5.4 | 24.6±3.2 ¹ | 35.9±5.8 ¹ | 30.9±4.7 ¹ | 39.8±5.2 ¹ |
| Solar flare | 68.0±3.1 | 25.2±4.4 ¹ | 70.6±2.9 ¹ | 67.6±3.5 ⁶ | 70.4±3.0 ¹ |
| Sonar | 24.1±8.7 | 50.8±7.6 ¹ | 64.7±7.2 ¹ | 73.3±7.5 ¹ | 66.2±7.5 ¹ |
| Soybean | 100.0±0.0 | 30.0±14.3 ¹ | 95.0±9.0 ³ | 100.0±0.0 ⁶ | 96.9±5.9 ³ |
| Splice junctions | 95.4±0.6 | 52.4±1.6 ¹ | 93.4±0.8 ¹ | 94.3±0.5 ¹ | 81.5±5.5 ¹ |
| Voting records | 91.2±1.6 | 60.5±3.1 ¹ | 96.3±1.3 ¹ | 94.9±1.2 ¹ | 95.8±1.6 ¹ |
| Wine | 90.9±13.3 | 36.4±9.9 ¹ | 91.7±5.6 ⁶ | 96.9±2.2 ⁴ | 90.8±4.7 ⁶ |
| Zoology | 91.9±3.6 | 39.4±6.4 ¹ | 89.6±4.7 ¹ | 94.6±4.3 ¹ | 90.6±5.0 ⁵ |

Table 1: Empirical results: average accuracies and standard deviations. Superscripts denote significance levels for the difference in accuracy between the SBC and the corresponding algorithm, using a one-tailed paired *t* test: 1 is 0.005, 2 is 0.01, 3 is 0.025, 4 is 0.05, 5 is 0.1, and 6 is above 0.1.

Number of domains in which SBC was more accurate versus less accurate than corresponding classifier

Same as line 1, but significant at 95% confidence

Table 2. Summary of accuracy results.

| Measure | SBC | C4.5 | PEBLS | CN2 |
|---------------|------|-------|-------|-------|
| No. wins | - | 16-12 | 15-11 | 18-10 |
| No. sig. wins | - | 12-9 | 7-9 | 12-8 |
| Rank | 2.32 | 2.54 | 2.79 | 2.68 |

Average rank over all domains (1 is best in each domain)

Measuring Attribute Dependence

They used a simple, pairwise mutual information measure:

For attributes A_m and A_n dependence is defined as

$$D(A_m, A_n | C) = H(A_m | C) + H(A_n | C) - H(A_m A_n | C)$$

where

$A_m A_n$ is a “derived attribute”, whose values consist of the possible combinations of values of A_m and A_n

$H(A_j | C)$ = “conditional entropy”

$$= \sum_i P(C_i) \sum_k -P(C_i \wedge A_j = v_k) \log_2 P(C_i \wedge A_j = v_k)$$

Note: If A_m and A_n are independent, then $D(A_m, A_n | C) = 0$.

Table 3: Empirical measures of attribute dependence.

| Domain | Rank | D_{Max} | % Hi. | D_{Avg} |
|------------------|------|-----------|-------|-----------|
| Breast cancer | 2 | 0.548 | 66.7 | 0.093 |
| Credit screening | 1 | 0.790 | 46.7 | 0.060 |
| Chess endgames | 4 | 0.383 | 25.0 | 0.015 |
| Pima diabetes | 1 | 0.483 | 62.5 | 0.146 |
| Echocardiogram | 3 | 0.769 | 85.7 | 0.360 |
| Glass | 4 | 0.836 | 100.0 | 0.363 |
| Heart disease | 1 | 0.388 | 53.8 | 0.085 |
| Hepatitis | 1 | 0.589 | 52.6 | 0.089 |
| Horse colic | 3 | 0.510 | 95.5 | 0.157 |
| Thyroid disease | 3 | 0.516 | 44.0 | 0.054 |
| Iris | 4 | 0.731 | 100.0 | 0.469 |
| Labor neg. | 1 | 1.189 | 100.0 | 0.449 |
| Lung cancer | 1 | 1.226 | 98.2 | 0.165 |
| Liver disease | 3 | 0.513 | 100.0 | 0.243 |
| LED | 1 | 0.060 | 0.0 | 0.025 |
| Lymphography | 2 | 0.410 | 55.6 | 0.076 |
| Post-operative | 3 | 0.181 | 0.0 | 0.065 |
| Promoters | 2 | 0.394 | 98.2 | 0.149 |
| Primary tumor | 1 | 0.098 | 0.0 | 0.023 |
| Solar flare | 3 | 0.216 | 16.7 | 0.041 |
| Sonar | 5 | 1.471 | 100.0 | 0.491 |
| Soybean | 1 | 0.726 | 31.4 | 0.016 |
| Splice junctions | 1 | 0.084 | 0.0 | 0.017 |
| Voting records | 4 | 0.316 | 25.0 | 0.052 |
| Wine | 3 | 0.733 | 100.0 | 0.459 |
| Zoology | 2 | 0.150 | 0.0 | 0.021 |

Results:

(1) SBC is more successful than more complex methods, even when there is substantial dependence among attributes.

(2) No correlation between degree of attribute dependence and SBC's rank.

But why????

An Example

- Let $C = \{+, -\}$, and attributes = $\{A, B, C\}$.
- Let $P(+)=P(-)=1/2$.
- Suppose A and C are completely independent, and A and B are completely dependent (i.e., $A=B$).
- Optimal classification procedure:

$$\begin{aligned}
 v_{MAP} &= \operatorname{argmax}_{v_j \in \{+, -\}} P(A, B, C | v_j) P(v_j) \\
 &= \operatorname{argmax}_{v_j \in \{+, -\}} P(A | v_j) P(C | v_j)
 \end{aligned}$$

- This leads to the following conditions:
 If $P(A|+) P(C|+) > P(A|-) P(C|-)$
 then class = +
 else class = -
- In the paper, the authors use Bayes Theorem to rewrite these conditions, and plot the “decision boundaries” for the optimal classifier and for the SBC.

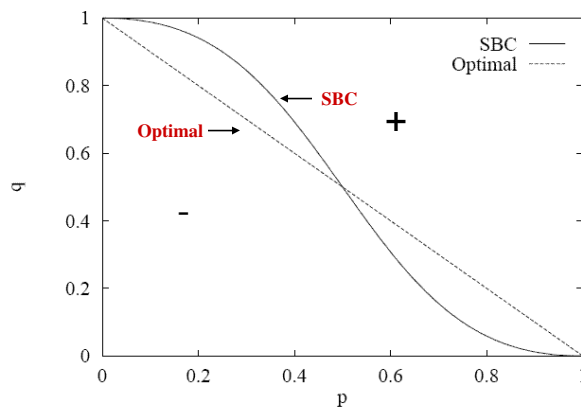


Figure 1: Decision boundaries for the SBC and the optimal classifier.

$$p = P(+ | A)$$

$$q = P(+ | C)$$

Even though A and B are *completely* dependent, and the SBC assumes they are completely independent, the SBC gives the optimal classification in a very large part of the problem space! *But why?*

- *Explanation:*

Suppose $V = \{+, -\}$ are the possible classes. Let E be a new example with attributes $\langle a_1, a_2, \dots, a_n \rangle$.

What the naive Bayes classifier does is calculates two probabilities,

$$P(+ | E) \sim P(+)\prod_i P(a_i | +)$$

$$P(- | E) \sim P(-)\prod_i P(a_i | -)$$

and returns the class that has the maximum probability given E .

- The probability calculations are correct only if the independence assumption is correct.
- However, the classification is correct in all cases in which the relative ranking of the two probabilities, as calculated by the SBC, is correct!
- The latter covers a lot more cases than the former.
- Thus, the SBC is effective in many cases in which the independence assumption does not hold.