Bayesian Learning

Reading:
C. Haruechaiyasak,
“A tutorial on naive Bayes classification”
(linked from class website)

Conditional Probability

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

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

Example

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?

Example

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?

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

$X =$ Second card is an ace

$Y =$ First card is an ace

$$P(Y) = \frac{4}{52}$$

$P(X,Y) =$ # possible pairs of aces / total # possible pairs

$$= \frac{4 \times 3}{52 \times 51} = \frac{12}{2652}$$

$$P(X \mid Y) = \frac{12}{2652} \div \frac{4}{42} = \frac{3}{51}$$

Deriving Bayes Rule

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

$$P(X, Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

Bayes rule :

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

Bayesian Learning

Application to Machine Learning

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

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:

Machine Learning Formulation

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

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, knows what is behind the doors. He 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

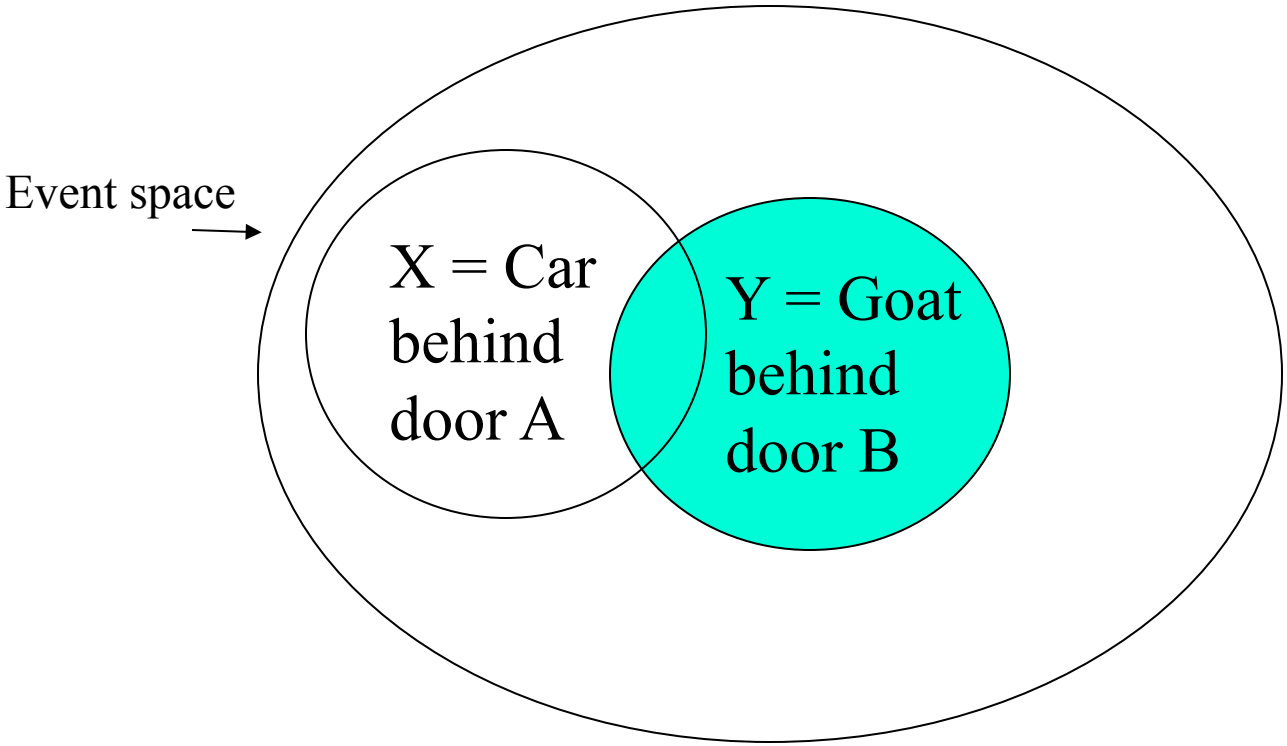
h_3 = Car is behind door C

Data D = Monty opened B to show a goat

What is $P(h_1 \mid D)$?

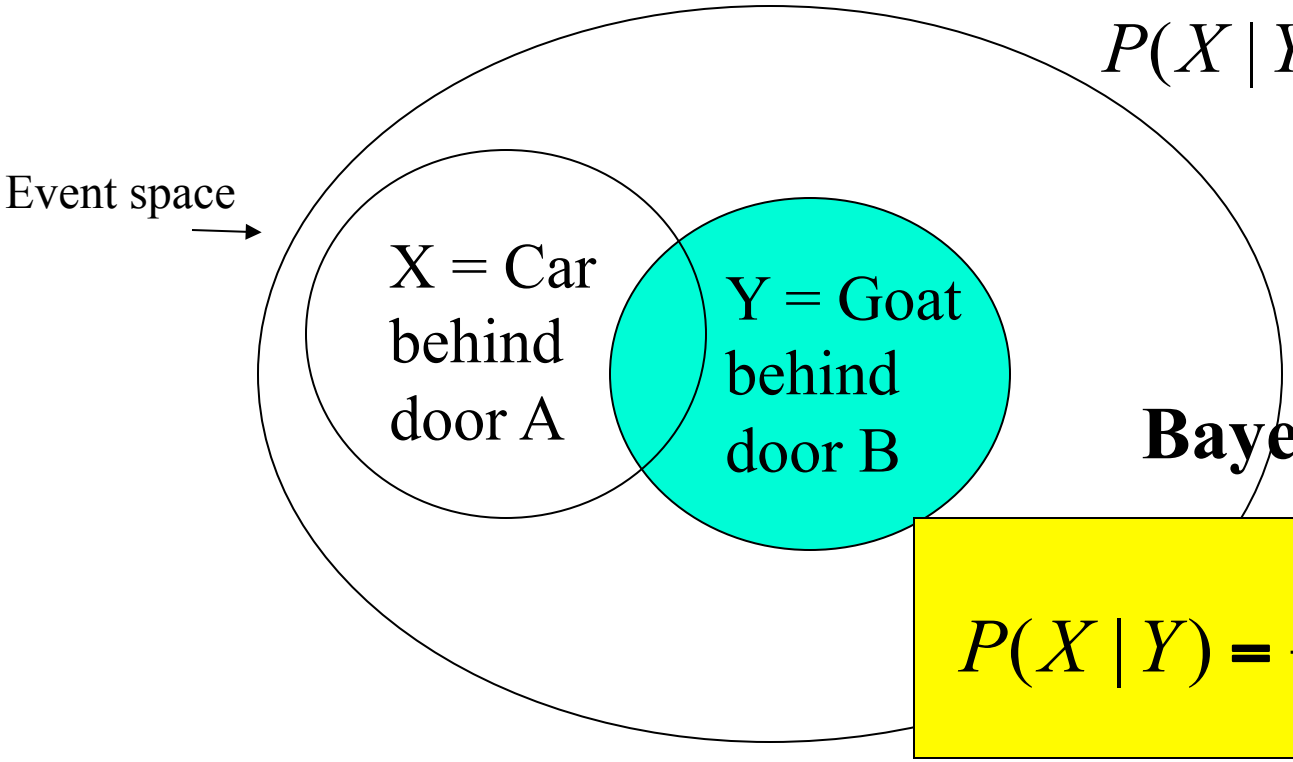
What is $P(h_2 \mid D)$?

What is $P(h_3 \mid D)$?



Event space = All possible configurations of cars and goats behind doors A, B, C

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



Bayes Rule:

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

Using Bayes' Rule to solve the Monty Hall problem

Using Bayes' Rule to solve the Monty Hall problem

You pick door A.
Data D = Monty opened door B to show a goat.

Hypothesis space H:

h_1 = Car is behind door A
h_2 = Car is behind door C
h_3 = Car is behind door B

What is P(h_1 | D)?
What is P(h_2 | D)?
What is P(h_3 | D)?

Prior probability:

P(h_1) = 1/3 P(h_2) =1/3 P(h_3) =1/3

Likelihood:

P(D | h_1) = 1/2

P(D | h_2) = 1

P(D | h_3) = 0

P(D) = p(D|h_1)p(h_1) + p(D|h_2)p(h_2) +
p(D|h_3)p(h_3) = 1/6 + 1/3 + 0 = 1/2

By Bayes rule:

P(h_1|D) = P(D|h_1)p(h_1) / P(D) = (1/2) (1/3) / (1/2) = 1/3

P(h_2|D) = P(D|h_2)p(h_2) / P(D) = (1)(1/3) / (1/2) = 2/3

So you should switch!

MAP ("maximum a posteriori") Learning

Bayes rule: $P(h \mid D) = \dfrac{P(D \mid h)P(h)}{P(D)}$

Goal of learning: Find maximum a posteriori hypothesis h_{MAP}:

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\, P(h \mid D)$$

$$= \underset{h \in H}{\text{argmax}}\, \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\text{argmax}}\, P(D \mid h)P(h)$$

because $P(D)$ is a constant independent of h.

Note: If every $h \in H$ is equally probable, then

$$h_{\text{MAP}} = \underset{h \in H}{\text{argmax}}\ P(D \mid h)$$

This is called the "maximum likelihood hypothesis".

A Medical Example

Toby takes a test for leukemia. The test has two outcomes: positive and negative. It is known that if the patient has leukemia, the test is positive 98% of the time. If the patient does not have leukemia, the test is positive 3% of the time. It is also known that 0.008 of the population has leukemia.

Toby's test is positive.

Which is more likely: Toby has leukemia or Toby does not have leukemia?

- **Hypothesis space:**

  h_1 = T. has leukemia

  h_2 = T. does not have leukemia

- **Prior:** 0.008 of the population has leukemia. Thus

  $P(h_1) = 0.008$

  $P(h_2) = 0.992$

- **Likelihood:**

  $P(+ \mid h_1) = 0.98, P(- \mid h_1) = 0.02$

  $P(+ \mid h_2) = 0.03, P(- \mid h_2) = 0.97$

- **Posterior knowledge:**

  Blood test is + for this patient.

- In summary

$$P(h_1) = 0.008,\ P(h_2) = 0.992$$

$$P(+ \mid h_1) = 0.98,\ P(- \mid h_1) = 0.02$$

$$P(+ \mid h_2) = 0.03,\ P(- \mid h_2) = 0.97$$

- Thus:

$$h_{MAP} = \underset{h \in H}{argmax}\ P(D \mid h)P(h)$$

$$P(+ \mid leukemia)P(leukemia) = (0.98)(0.008) = 0.0078$$

$$P(+ \mid \neg leukemia)P(\neg leukemia) = (0.03)(0.992) = 0.0298$$

$$h_{MAP} = \neg leukemia$$

- What is P(leukemia|+)?

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

So,

$$P(leukemia \mid +) = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$

$$P(\neg leukemia \mid +) = \frac{0.0298}{0.0078 + 0.0298} = 0.79$$

These are called the “posterior” probabilities.

In-Class Exercises

1-2

Quiz 5 (this Thursday)

What you need to know:

- Definition of Information Gain and how to calculate it
- Bayes rule and how to derive it from definition of conditional probability
- How to solve problems similar to in-class exercises using Bayes rule
- **Note:** Naïve Bayes algorithm won't be on this quiz.
- No calculator needed

In-class exercises 2b-2c

Bayesianism vs. Frequentism

- Classical probability: **Frequentists**
  - Probability of a particular event is defined relative to its *frequency* in a sample space of events.
  - E.g., probability of "the coin will come up heads on the next trial" is defined relative to the *frequency* of heads in a sample space of coin tosses.
- **Bayesian** probability:
  - Combine measure of "prior" belief you have in a proposition with your subsequent observations of events.
- **Example:** Bayesian can assign probability to statement "There was life on Mars a billion years ago" but frequentist cannot.

Independence and Conditional Independence

- Two random variables, X and Y, are independent if

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

- Two random variables, X and Y, are independent *given* Z if

$$P(X,Y \mid C) = P(X \mid C)P(Y \mid C)$$

- Examples?

Naive Bayes Classifier

Let $f(\mathbf{x})$ be a target function for classification: $f(\mathbf{x}) \in \{+1, -1\}$.

Let $\boldsymbol{x} = \langle x_1, x_2, ..., x_n \rangle$

We want to find the most probable class value, h_{MAP}, given the data $\mathbf{x}$:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class \mid D)$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class \mid x_1, x_2, ..., x_n)$$

By Bayes Theorem:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\text{argmax}} \frac{P(x_1, x_2, ..., x_n \mid class) P(class)}{P(x_1, x_2, ..., x_n)}$$

$$= \underset{class \in \{+1,-1\}}{\text{argmax}} P(x_1, x_2, ..., x_n \mid class) P(class)$$

$P(class)$ can be estimated from the training data. How?

However, in general, not practical to use training data to estimate $P(x_1, x_2, ..., x_n \mid class)$. Why not?

- Naive Bayes classifier: Assume

$$P(x_1, x_2, ..., x_n \mid class) = P(x_1 \mid class)P(x_2 \mid class)\cdots P(x_n \mid class)$$

Is this a good assumption?

Given this assumption, here's how to classify an instance $\boldsymbol{x} = \langle x_1, x_2, ..., x_n \rangle$:

Naive Bayes classifier:

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class)\prod_i P(x_i \mid class)$$

Estimate the values of these various probabilities over the training set.

Training data:

Day	Outlook	Temp	Humidity	Wind	PlayTennis
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

Test data:

D15	Sunny	Cool	High	Strong	?

In practice, use training data to compute a probablistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

In practice, use training data to compute a probablistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

In practice, use training data to compute a probablistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)$$

In-class exercise 3a

Estimating probabilities / Smoothing

- **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?

- Let $n(c)$ be the number of training instances with class c, and $n(x_i = a_i, c)$ be the number of training instances with attribute value $x_i = a_i$ and class c. Then

$$P(x_i = a_i \mid c) = \frac{n(x_i = a_i, c)}{n(c)}$$

- **Problem with this method:** If $n(c)$ is very small, gives a poor estimate.

- E.g., $P(Outlook = Overcast \mid no) = 0$.

- Now suppose we want to classify a new instance:

  <*Outlook=overcast*, *Temperature=cool*, *Humidity=high*, *Wind=strong*>. Then:

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

This incorrectly gives us zero probability due to small sample.

One solution: Laplace smoothing (also called "add-one" smoothing)

For each class c_j and attribute x_i with value a_i, add one "virtual" instance.

That is, recalculate:

$$P(x_i = a_i \mid c_j) \approx \frac{n(x_i = a_i\ ,\ c_j) + 1}{n(c_j) + k}$$

where k is the number of possible values of attribute a.

Training data:

Day	Outlook	Temp	Humidity	Wind	PlayTennis
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

Add virtual instances for *Outlook*:

Outlook=Sunny: **Yes** *Outlook=Overcast*: **Yes** *Outlook=Rain*: **Yes**

Outlook=Sunny: **No** *Outlook=Overcast*: **No** *Outlook=Rain*: **No**

P(*Outlook=Overcast*| **No**) = 0 / 5 → 0 + 1 / 5 + 3 = 1/8

$$P(Outlook = Sunny \mid Yes) = 2/9 \rightarrow 3/12 \quad P(Outlook = Sunny \mid No) = 3/5 \rightarrow 4/8$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \rightarrow 5/12 \quad P(Outlook = Overcast \mid No) = 0/5 \rightarrow 1/8$$
$$P(Outlook = Rain \mid Yes) = 3/9 \rightarrow 4/12 \quad P(Outlook = Rain \mid No) = 2/5 \rightarrow 3/8$$

$$P(Humidity = High \mid Yes) = 3/9 \rightarrow 4/11 \quad P(Humidity = High \mid No) = 4/5 \rightarrow 5/7$$
$$P(Humidity = Normal \mid Yes) = 6/9 \rightarrow 7/11 \quad P(Humidity = Normal \mid No) = 1/5 \rightarrow 2/7$$

Etc.

In-class exercise 3b

Naive Bayes on continuous-valued attributes

- How to deal with continuous-valued attributes?

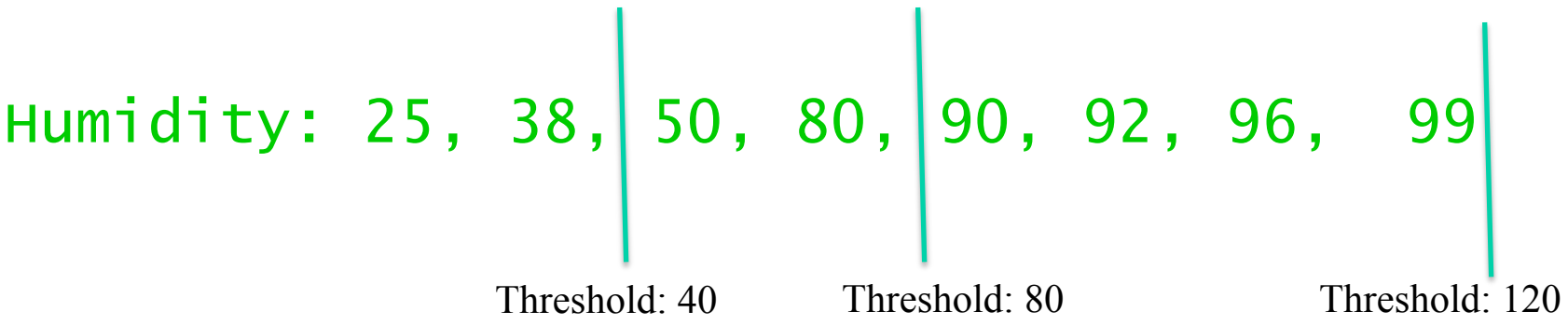
Two possible solutions:

- Discretize
- Assume particular probability distribution of classes over values (estimate parameters from training data)

Simplest discretization method

For each attribute x_i, create k equal-size bins in interval from $min(x_i)$ to $max(x_i)$.

Choose thresholds in between bins.



***P*(*Humidity* < 40 | *yes*)** **$P(40 <= Humidity < 80 \mid yes)$** **$P(80 <= Humidity < 120 \mid yes)$**

***P*(*Humidity* < 40 | *no*)** **$P(40 <= Humidity < 80 \mid no)$** **$P(80 <= Humidity < 120 \mid no)$**

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 add-one smoothing of probabilities

This gives good balance between discretization bias and variance.

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 add-one smoothing of probabilities

This gives good balance between discretization bias and variance.

Humidity: 25, 38, 50, | 80, 90, 92, | 96, 99

$\sqrt{8} \approx 3$ bins, 3 items per bin

In-class exercise 4

Gaussian Naïve Bayes

Assume that within each class, values of each numeric feature are normally distributed:

$$p(x_i = a_i \mid c_j) = N(x_i; \mu_{i,c_j}, \sigma_{i,c_j})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\mu_{i,c}$ is the mean of feature i given the class c_j, and $\sigma_{i,c}$ is the standard deviation of feature i given the class c_j

We estimate $\mu_{i,c}$ and $\sigma_{i,c}$ from training data.

Example

f_1	f_2	**Class**
3.0	5.1	+
4.1	6.3	+
7.2	9.8	+
2.0	1.1	−
4.1	2.0	−
8.1	9.4	−

Example

f_1	f_2	Class
3.0	5.1	+
4.1	6.3	+
7.2	9.8	+
2.0	1.1	−
4.1	2.0	−
8.1	9.4	−

$$\mu_{1,+} = \frac{(3.0 + 4.1 + 7.2)}{3} = 4.8$$

$$\sigma_{1,+} = \sqrt{\frac{(3.0 - 4.8)^2 + (4.1 - 4.8)^2 + (7.2 - 4.8)^2}{3}} = 1.8$$

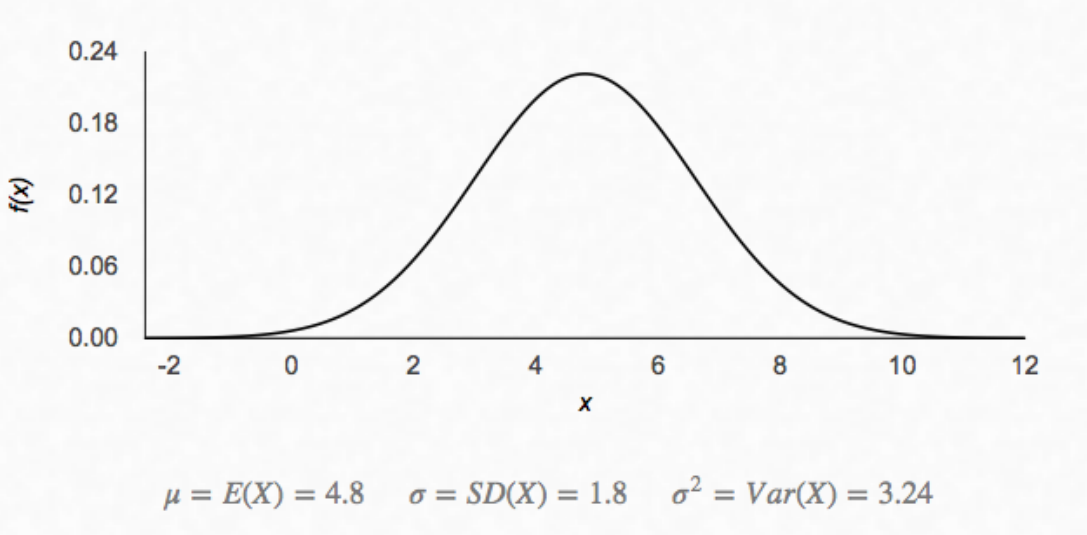
$$\mu_{1,-} = \frac{(2.0 + 4.1 + 8.1)}{3} = 4.7$$

$$\sigma_{1,-} = \sqrt{\frac{(2.0 - 4.7)^2 + (4.1 - 4.7)^2 + (8.1 - 4.7)^2}{3}} = 2.5$$

Fill in the rest…

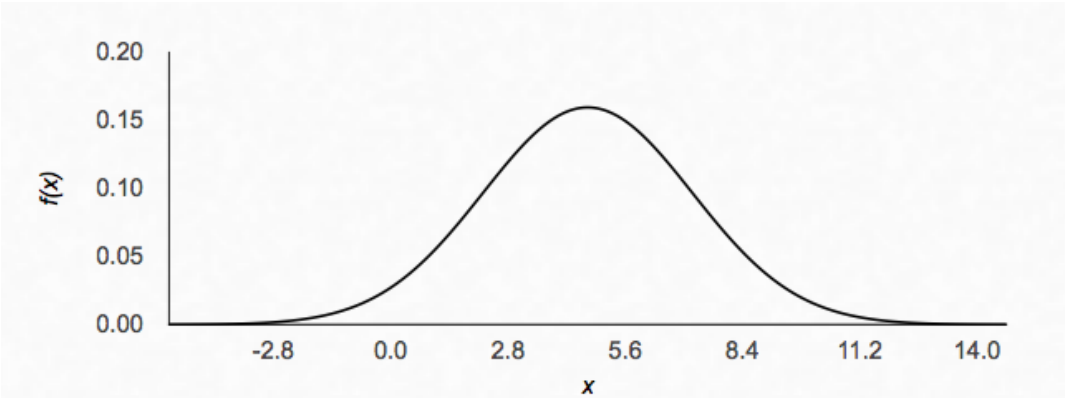
http://homepage.stat.uiowa.edu/~mbognar/applets/normal.html

$N_{1,+} = N(x;\ 4.8,\ 1.8)$



$\mu = E(X) = 4.8 \quad \sigma = SD(X) = 1.8 \quad \sigma^2 = Var(X) = 3.24$

$N_{1,-} = N(x;\ 4.7,\ 2.5)$



$\mu = E(X) = 4.7 \quad \sigma = SD(X) = 2.5 \quad \sigma^2 = Var(X) = 6.25$

Now, suppose you have a new example **x**, with $f_1 = 5.2, f_2 = 6.3$.

What is $class_{NB}$ (**x**) ?

Now, suppose you have a new example **x**, with $f_1 = 5.2, f_2 = 6.3$.

What is $class_{NB}$ (**x**) ?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)$$

Now, suppose you have a new example **x**, with $f_1 = 5.2, f_2 = 6.3$.

What is $class_{NB}$ (**x**) ?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)$$

$$p(x_i = a_i \mid c_j) = N(x_i; \mu_{i,c_j}, \sigma_{i,c_j})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifer

(P. Domingos and M. Pazzani)

Naive Bayes classifier is 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±3.9 [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

$$D(A_m, A_n \mid C)$$
$$= Entropy(A_m \mid C) + Entropy(A_n \mid C) - Entropy(A_m A_n \mid C)$$

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

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

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????

- Explanation:

Suppose $C = \{+1,-1\}$ are the possible classes. Let **x** be a new example with attributes $<a_1, a_2, ..., a_n>$..

What the naive Bayes classifier does is calculates two probabilities,

$$P(+ \mid \mathbf{x}) \sim P(+) \prod_i P(x_i = a_i \mid +)$$

$$P(- \mid \mathbf{x}) \sim P(-) \prod_i P(a_i \mid -)$$

and returns the class that has the maximum probability given **x**.

- 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.