Classification

- **Rule-Based Classification**
- **Naïve Bayes Classification**

Rule-Based Classification

Rule-Based Classifier
Using IF-THEN Rules for Classification

- Classify tuples by using a collection of "if…then…" rules.
  - Represent the knowledge in the form of IF-THEN rules

Rule: IF **(Condition)** THEN **Consequent**

- where
  - **Condition** is a conjunction of attributes (**rule antecedent** or **condition**)
  - **Consequent** is a class label (**rule consequent**)

- Examples of classification rules:

  IF **(*age*=youth AND *student*=yes)** THEN ***buys_computer = yes***

  IF **(*BloodType*=warm AND *LayEggs*=yes)** THEN ***class = bird***

Rule-Based Classifier: Rule Coverage and Accuracy

- A rule r **covers** an instance x if the attributes of the instance satisfy the condition of the rule.

Coverage of a rule:

- Fraction of tuples that satisfy the condition of a rule.

Accuracy of a rule:

- Fraction of tuples that satisfy the condition that also satisfy the consequent of a rule.

$\mathbf{n_{covers}}$: # of tuples covered by rule R

$\mathbf{n_{correct}}$: # of tuples correctly classified by rule R

$\mathbf{coverage(R) = n_{covers} / |D|}$ /* D: training data set */

$\mathbf{accuracy(R) = n_{correct} / n_{covers}}$

Rule-Based Classifier: Rule Coverage and Accuracy - Example

IF **(Status=Single)** THEN **Class=No**

Coverage = 4/10 = 40%

Accuracy = 2/4 = 50%

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	**Single**	125K	**No**
2	No	Married	100K	**No**
3	No	**Single**	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	**Single**	85K	**Yes**
9	No	Married	75K	**No**
10	No	**Single**	90K	**Yes**

Rule-Based Classifier - Example

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

R1: IF (Give Birth = no) AND (Can Fly = yes) THEN Class=Birds
R2: IF (Give Birth = no) AND (Live in Water = yes) THEN Class=Fishes
R3: IF (Give Birth = yes) AND (Blood Type = warm) THEN Class=Mammals
R4: IF (Give Birth = no) AND (Can Fly = no) THEN Class=Reptiles
R5: IF (Live in Water = sometimes) THEN Class=Amphibians

How a Rule-Based Classifier Works?

- **A rule-based classifier classifies a tuple based on the rule triggered by the tuple.**

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

R1: IF (Give Birth = no) AND (Can Fly = yes) THEN Class=Birds
R2: IF (Give Birth = no) AND (Live in Water = yes) THEN Class=Fishes
R3: IF (Give Birth = yes) AND (Blood Type = warm) THEN Class=Mammals
R4: IF (Give Birth = no) AND (Can Fly = no) THEN Class=Reptiles
R5: IF (Live in Water = sometimes) THEN Class=Amphibians

- A **lemur** triggers rule R3, so it is classified as a **mammal**
- A **turtle** triggers both R4 and R5
  - Since the classes predicted by the rules are contradictory (**reptiles** versus **amphibians**), their conflicting classes must be resolved.
- A **dogfish shark** triggers none of the rules
  - we need to ensure that the classifier can still make a reliable prediction even though a tuple is not covered by any rule.

Characteristics of Rule Sets

Mutually Exclusive Rules

- The rules in a rule set R are mutually exclusive if no two rules in R are triggered by the same tuple.
- Every tuple is covered by at most one rule in R.

Exhaustive Rules

- A rule set R has exhaustive coverage if there is a rule for each combination of attribute values.
- Every record is covered by at least one rule in R.

Characteristics of Rule Sets

Rules are not mutually exclusive:

- A tuple may trigger more than one rule
- ***Solution 1: Use ordered rule set***
  - The rules in a rule set are ordered in decreasing order of their priority (e.g., based on accuracy, coverage, or the order in which the rules are generated).
  - An ordered rule set is also known as a **decision list**.
  - When a tuple is presented, it is classified by the **highest-ranked rule** that covers the tuple.
- ***Solution 2: Use unordered rule set and a voting scheme***
  - A tuple triggers multiple rules and considers the consequent of each rule as a vote for a particular class.
  - The tuple is usually assigned to the class that receives the highest number of votes. In some cases, the vote may be weighted by the rule's accuracy

Characteristics of Rule Sets

Rules are not exhaustive

- A tuple may not trigger any rules
- ***Solution: Use a default class***
  - If the rule set is not exhaustive, then a default rule must be added to cover the remaining cases.

    DefaultRule: IF () THEN Class = ***defaultclass***
  - A default rule has an empty antecedent (TRUE) and is triggered when all other rules have failed.
  - ***defaultclass*** is known as the default class and is typically assigned to the majority class of training records not covered by the existing rules.

Rule Ordering Schemes

- Rule ordering can be implemented on a rule-by-rule basis or on a class-by-class basis.

Rule-Based Ordering Scheme

- Individual rules are ranked based on their quality.
  - Every tuple is classified by the “best” rule covering it.
  - Lower-ranked rules are much harder to interpret because they assume the negation of the rules preceding them.

Class-Based Ordering Scheme

- Rules that belong to the same class appear together in the rule set.
- The rules are then collectively sorted on the basis of their class information.
  - The relative ordering among the rules from the same class is not important; as long as one of the rules fires, the class will be assigned to the test tuple.
  - A high-quality rule to be overlooked in favor of an inferior rule that happens to predict the higher-ranked class.

Rule Ordering Schemes

Rule-Based Ordering

(Skin Cover=feathers, Aerial Creature=yes) ⇒ Birds
(Body temperature=warm-blooded, Gives Birth=yes) ⇒ Mammals
(Body temperature=warm-blooded, Gives Birth=no) ⇒ Birds
(Aquatic Creature=semi) ⇒ Amphibians
(Skin Cover=scales, Aquatic Creature=no) ⇒ Reptiles
(Skin Cover=scales, Aquatic Creature=yes) ⇒ Fishes
(Skin Cover=none) ⇒ Amphibians

Class-Based Ordering

(Skin Cover=feathers, Aerial Creature=yes) ⇒ Birds
(Body temperature=warm-blooded, Gives Birth=no) ⇒ Birds
(Body temperature=warm-blooded, Gives Birth=yes) ⇒ Mammals
(Aquatic Creature=semi) ⇒ Amphibians
(Skin Cover=none) ⇒ Amphibians
(Skin Cover=scales, Aquatic Creature=no) ⇒ Reptiles
(Skin Cover=scales, Aquatic Creature=yes ⇒ Fishes

Building Classification Rules

- To build a rule-based classifier, we need to extract a set of rules that identifies key relationships between the attributes of a data set and the class label.

Direct Methods:

- Extract classification rules directly from data.
- Examples: RIPPER, CN2, …

Indirect Methods:

- Extract rules from other classification models (e.g. decision trees, etc).
- Examples: C4.5rules

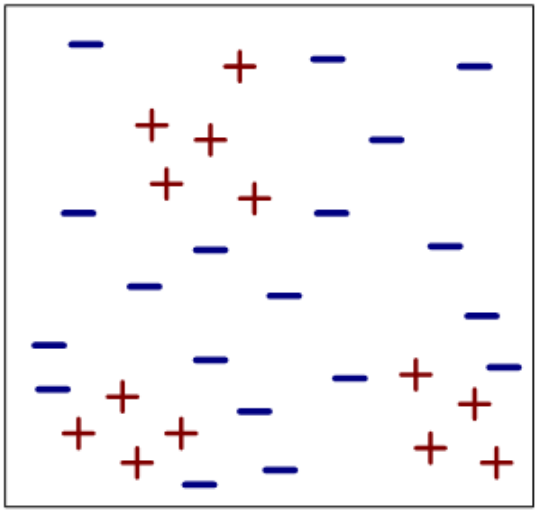
Direct Method: Sequential Covering

- **Sequential covering algorithm** extracts rules directly from training data
- Typical sequential covering algorithms: RIPPER, FOIL, CN2,
- Rules are learned sequentially, each for a given class C_i will cover many tuples of C_i but none (or few) of the tuples of other classes.

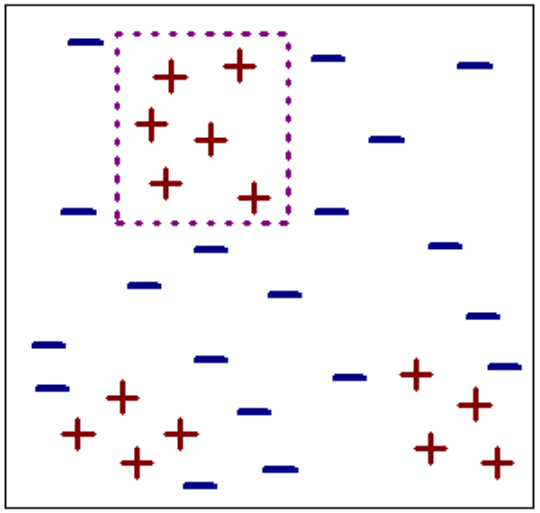
Sequential Covering Algorithm

1: Let E be the training records and A be the set of attribute-value pairs, $\{(A_j, v_j)\}$.
2: Let Y_o be an ordered set of classes $\{y_1, y_2, \ldots, y_k\}$.
3: Let $R = \{\ \}$ be the initial rule list.
4: **for** each class $y \in Y_o - \{y_k\}$ **do**
5: **while** stopping condition is not met **do**
6: $r \leftarrow$ Learn-One-Rule (E, A, y).
7: Remove training records from E that are covered by r.
8: Add r to the bottom of the rule list: $R \longrightarrow R \vee r$.
9: **end while**
10: **end for**
11: Insert the default rule, $\{\} \longrightarrow y_k$, to the bottom of the rule list R.

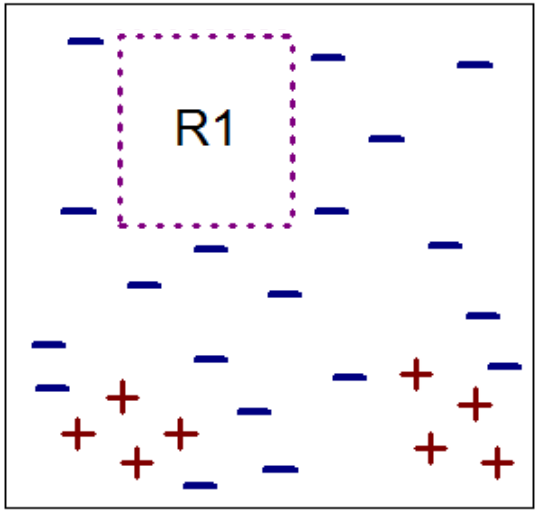
Sequential Covering - Example



original data set



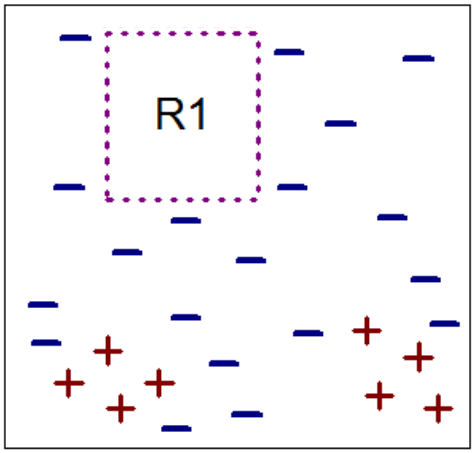
tuples covered by R1

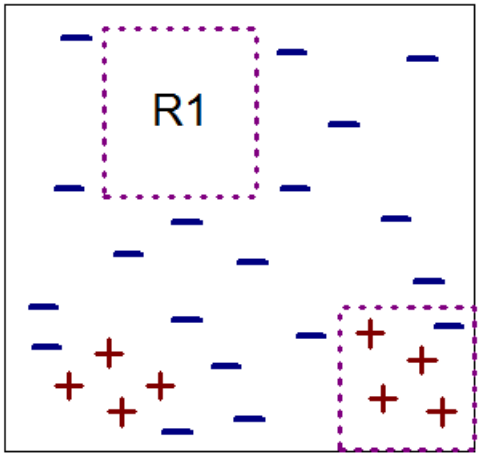


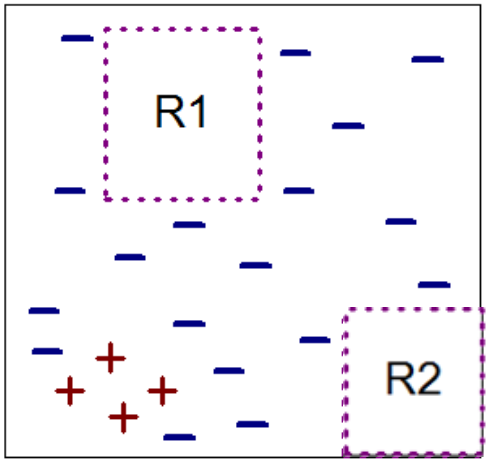
remove tuples covered by R1

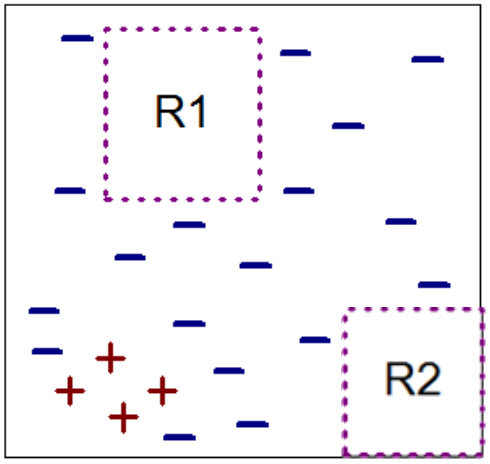
- Find the first rule R1 and remove the tuples covered by R1 from the training data set.
- Add the rule to the rule list.

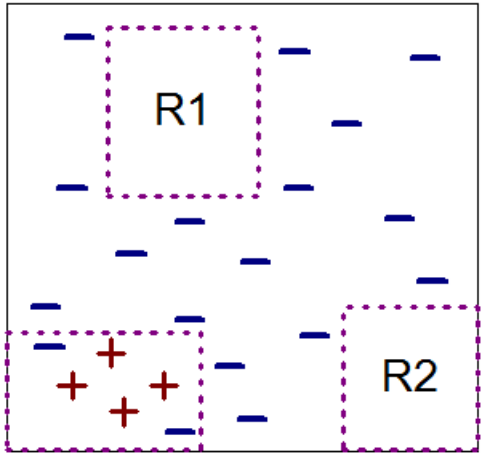
Sequential Covering - Example

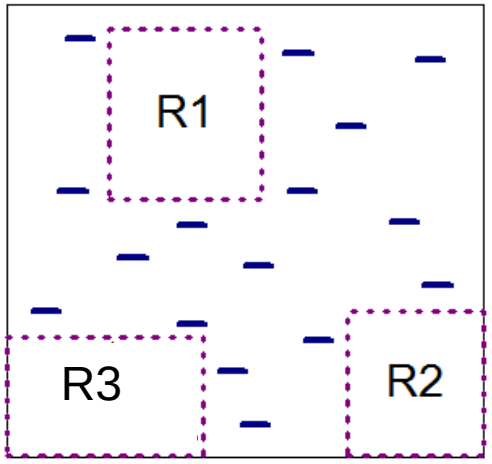












Sequential Covering
How to Learn-One-Rule?

- The objective of Learn-One-Rule function is to extract a classification rule that covers many of the positive examples and none (or very few) of the negative examples in the training set.
- Finding an optimal rule is computationally expensive given the exponential size of the search space.
- Learn-One-Rule function addresses the exponential search problem by growing the rules in a greedy fashion.
  - It generates an initial rule r and keeps refining the rule until a certain stopping criterion is met.
  - The rule is then pruned to improve its generalization error.

Learn-One-Rule: Rule-Growing Strategy

- Two common ***rule-growing strategies***: **general-to-specific** or **specific-to-general**.

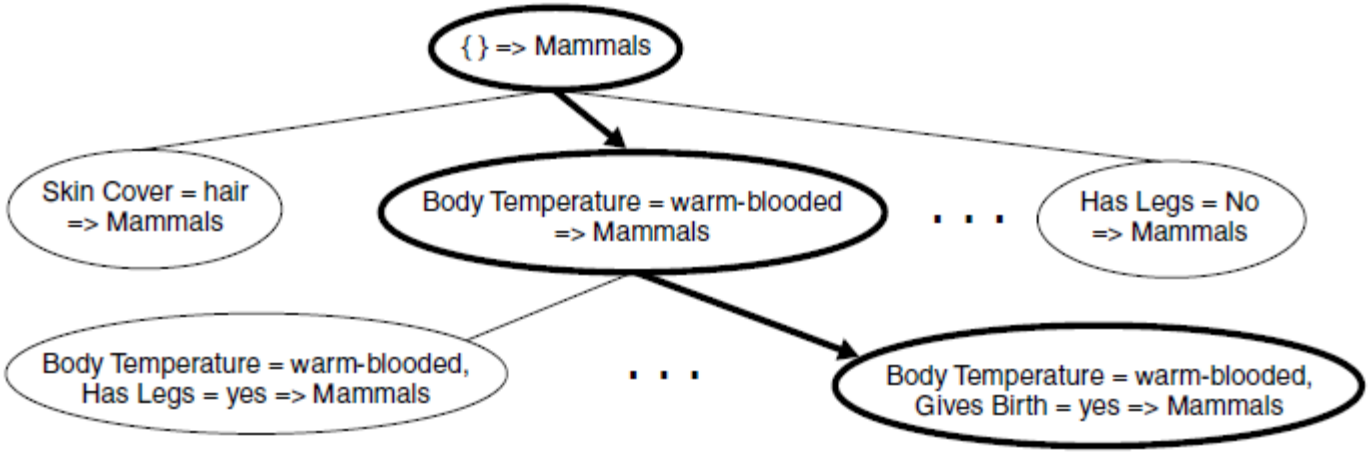
General-to-specific:

- An initial rule r : {} → y is created, where the left-hand side is an empty set and the right-hand side contains the target class.
  - The rule has poor quality because it covers all the examples in the training set.
- New conjuncts are subsequently added to improve the rule's quality.
  - Algorithm then explores all possible candidates and greedily chooses the next conjunct.
  - Process continues until the stopping criterion is met (e.g., when the added conjunct does not improve the quality of the rule).

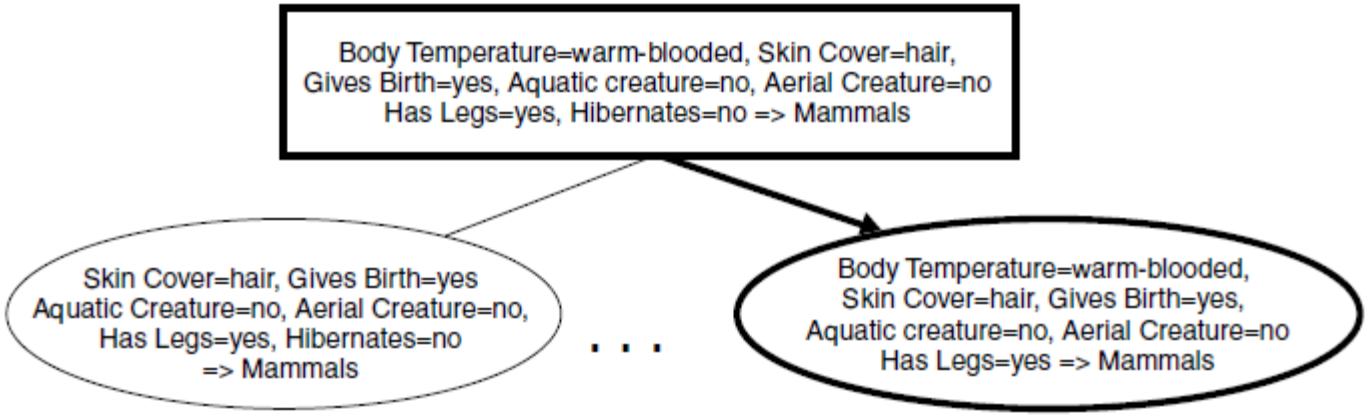
Specific-to-general

- One of the positive examples is randomly chosen as the initial seed for the rule-growing process.
- During the refinement step, the rule is generalized by removing one of its conjuncts so that it can cover more positive examples.

Learn-One-Rule: Rule-Growing Strategy - Example



General-to-specific



Specific-to-general

Learn-One-Rule: Rule Evaluation

- An **evaluation metric** is needed to determine which conjunct should be added (or removed) during the rule-growing process.
  - **Accuracy** is an obvious choice because it explicitly measures the fraction of training examples classified correctly by the rule.
  - A potential limitation of accuracy is that it does not take into account the ***rule's coverage***.
    - Training data set has: 60 positive examples, 100 negative examples
    - R1: covers 50 positive examples and 5 negative examples
    - R2: covers 2 positive examples and no negative examples.
    - Although the accuracy of R2 (100%) is higher than the accuracy of R1 (90.9%), R1 is a better rule because of the coverage of the rule.
- We need evaluation metrics take into account the ***rule's coverage:***
  - **Foil's Information Gain,** and other metrics.

Learn-One-Rule: Rule Evaluation

- **Foil's Information Gain** is a ***rule-quality measure*** which considers both coverage and accuracy.

R0: {A} ⇒ class (initial rule)

R1: {A and B} ⇒ class (rule after adding conjunct B)

$$\textbf{Foil's Information Gain(R0, R1)} = \mathbf{p1} \times \left(\mathbf{log_2} \frac{\mathbf{p1}}{\mathbf{p1+n1}} - \mathbf{log_2} \frac{\mathbf{p0}}{\mathbf{p0+n0}} \right)$$

where

p0: number of positive instances covered by R0

n0: number of negative instances covered by R0

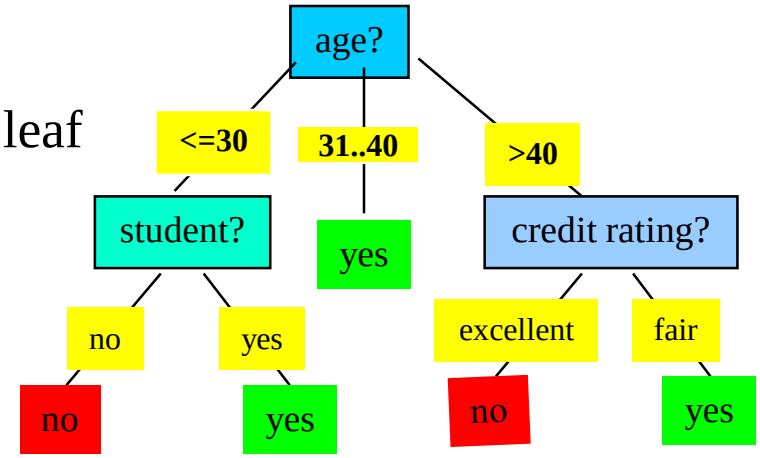
p1: number of positive instances covered by R1

n1: number of negative instances covered by R1

Indirect Method for Rule Extraction

Rule Extraction from a Decision Tree

- Rules are easier to understand than large trees
- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction: the leaf holds the class prediction
- Rules are mutually exclusive and exhaustive



- Example: Rule extraction from our *buys_computer* decision-tree

IF *age* = young AND *student* = *no* THEN *buys_computer* = *no*

IF *age* = young AND *student* = *yes* THEN *buys_computer* = *yes*

IF *age* = mid-age THEN *buys_computer* = *yes*

IF *age* = old AND *credit_rating* = *excellent* THEN *buys_computer* = *no*

IF *age* = old AND *credit_rating* = *fair* THEN *buys_computer* = *yes*

Indirect Method for Rule Extraction
C4.5 Rules

- Extract rules from an unpruned decision tree
- For each rule, r: A $\rightarrow$ y,
  - Consider an alternative rule r′: A′ $\rightarrow$ y where A′ is obtained by removing one of the conjuncts in A
  - Compare the pessimistic error rate for r against all r's
  - Prune if one of the alternative rules has lower pessimistic error rate
  - Repeat until we can no longer improve generalization error
- Instead of ordering the rules, order subsets of rules (class ordering)
  - Each subset is a collection of rules with the same rule consequent (class)

Naïve Bayes Classification

Bayesian Classification

- **Bayesian classifiers** are ***statistical classifiers***.
  - They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
- Bayesian classification is based on **Bayes Theorem.**
- A simple Bayesian classifier known as the **Naïve Bayesian Classifier** to be *comparable in performance* with decision tree and selected neural network classifiers.
  - Bayesian classifiers exhibits high accuracy and speed when applied to large databases.
- Even when **Bayesian methods** are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

Bayes Theorem

- P(A) is **prior probability** (**unconditional probability**) of event A.

- P(A|B) is **posterior probability** (**conditional probability**) of event A given that event B holds.

- P(A,B) is the **joint probability** of two events A and B.
  - The (unconditional) probability of the events A and B occurring together.
  - P(A,B) = P(B,A)

Bayes Theorem

$P(A|B) = P(A,B) / P(B)$ ➔ $P(A,B) = P(A|B)*P(B)$

$P(B|A) = P(B,A) / P(A)$ ➔ $P(B,A) = P(B|A)*P(A)$

Since $P(A,B) = P(B,A)$, we have $P(A|B)*P(B) = P(B|A)*P(A)$

Thus, we have **Bayes Theorem**

$P(A|B) = P(B|A)*P(A) / P(B)$

$P(B|A) = P(A|B)*P(B) / P(A)$

Bayes Theorem - Example

Bayes Theorem

P(A|B) = P(B|A)*P(A) / P(B)

P(B|A) = P(A|B)*P(B) / P(A

Sample Space for events A and B

A holds	T	T	F	F	T	F	T
B holds	T	F	T	F	T	F	F

P(A) = 4/7 P(B) = 3/7 P(A,B) = P(B,A) = 2/7

P(B|A) = 2/4 P(A|B) = 2/3

Is Bayes Theorem correct?

P(B|A) = P(A|B)*P(B) / P(A) = (2/3 * 3/7) / 4/7 = 2/4 ➔ CORRECT

P(A|B) = P(B|A)*P(A) / P(B) = (2/4 * 4/7) / 3/7 = 2/3 ➔ CORRECT

Bayes Theorem - Example

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time

    P(S|M) = 0.5
  - Prior probability of any patient having meningitis is 1/50,000

    P(M) = 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

    P(S) = 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis? **P(M|S) ?**

$$\mathbf{P(M|S) = \frac{P(S|M)\,P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002}$$

Independence of Events

- The events A and B are **INDEPENDENT** if and only if **P(A,B) = P(A)*P(B)**

Example: Bit strings of length 3 is {000,001,010,011,100,101,110,111}

Event A: A randomly generated bit string of length three begins with a 1.

Event B: A randomly generated bit string of length three ends with a 1.

P(A) = 4/8 100,101,110,111 P(B) = 4/8 001,011,101,111

P(A,B) = 2/8 101,111 Are A and B independent?

P(A)*P(B) = (4/8) * (4/8) = 16/64 = 2/8 = P(A,B)

➔ A and B are independent.

Event C: A randomly generated bit string of length three contains with two 1s.

P(C) = 3/8 011,101,110

P(A,C) = 2/8 101,110 Are A and C independent?

P(A)*P(C) = (4/8)*(3/8) = 12/64 = 3/16 ≠ 2/8

➔ A and C are NOT independent.

Bayes Theorem for Prediction

- Let **X** be a **data sample**: its class label is unknown.
- Let H be a **hypothesis** that X belongs to class C.
- **Classification** is to determine P(H|**X**), (i.e., **posteriori probability**): the probability that the hypothesis holds given the observed data sample **X.**
- P(H) (**prior probability**): the initial probability of H
  - E.g., **X** will buy computer, regardless of age, income, …
- P(**X**): probability that sample data is observed
- P(**X**|H) (**likelihood**): the probability of observing the sample **X**, given that the hypothesis H holds.

Bayes Theorem: $$\mathbf{P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}}$$

- Predicts **X** belongs to C_i iff the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|X)$ for all the *k* classes.

Naïve Bayes Classifier

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an attribute vector $(x_1, x_2, \ldots, x_n)$
  - Attributes $A_1, A_2, \ldots, A_n$ have values $A_1=x_1, A_2=x_2, \ldots, A_n=x_n$
- Suppose there are *m* classes $C_1, C_2, \ldots, C_m$.
- We are looking the classification of the tuple $(x_1, x_2, \ldots, x_n)$.
- **The classification of this tuple will be the class C_i that maximizes the following conditional probability**.

$$\mathbf{P(C_i \mid x_1, x_2, \ldots, x_n)}$$

Naïve Bayes Classifier

- To compute $P(C_i \mid x_1, x_2, \ldots, x_n)$ is almost impossible for a real data set.
- We use Bayes Theorem to find this conditional probability.

$$P(C_i \mid x_1, x_2, \ldots, x_n) = P(x_1, x_2, \ldots, x_n \mid C_i) * P(C_i) \; / \; P(x_1, x_2, \ldots, x_n)$$

- **Since $P(x_1, x_2, \ldots, x_n)$ is constant for all classes, we only look at the class C_i that maximizes the following formula.**

$$\mathbf{P(x_1, x_2, \ldots, x_n \mid C_i) * P(C_i)}$$

- We should compute $P(x_1, x_2, \ldots, x_n \mid C_i)$ and $P(C_i)$ from the training dataset.

Naïve Bayes Classifier
Computing Probabilities

- To compute $P(C_i)$ from the dataset is easy.

  $P(C_i) = N_{Ci} / N$ where N_{Ci} is the number of tuples belong to class C_i and N is the number of the total tuples in the dataset.

- But, to compute $P(x_1, x_2, \ldots, x_n | C_i)$ from the dataset is NOT easy.
  - In fact, it is almost impossible for a dataset with many attributes.
  - If we have n binary attributes, the number of possible tuples is 2^n.

Naïve Bayes Classifier
Computing Probabilities – Independence Assumption

- In order to compute $P(x_1, x_2, \ldots, x_n \mid C_i)$, we make independence assumption for attributes although this assumption may not be true.

  Independence Assumption: Attributes are conditionally independent (i.e., no dependence relation between attributes)

  $$\mathbf{P(x_1, x_2, \ldots, x_n \mid C_i) = P(x_1 \mid C_i) * P(x_2 \mid C_i) * \ldots * P(x_n \mid C_i)}$$

- If $\mathbf{A_k}$ **is categorical**,

  $P(x_k|C_i)$ is the # of tuples in C_i having value x_k for A_k divided by

  $|C_i|$ (# of tuples of C_i in the dataset)

Naïve Bayes Classifier

Computing Probabilities – continuous-valued attribute

- If **A_k is a continuous-valued attribute**,

  $P(x_k|C_i)$ is usually computed based on Gaussian distribution with a mean μ and standard deviation σ

$$\mathbf{g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

and $P(x_k|C_i)$ is $\mathbf{g(x_k, \mu_{C_i}, \sigma_{C_i})}$

- Or we can discretize the continuous-valued attribute first.

Naïve Bayes Classifier

Computing Probabilities from Training Dataset

Dataset has 14 tuples.

Two classes:

buyscomputer=yes

buyscomputer=no

P(bc=yes) = 9/14

P(bc=no) = 5/14

age	income	student	creditrating	buyscomputer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayes Classifier

Computing Probabilities from Training Dataset

P(age=b31|bc=yes)=2/9
P(age=i31|bc=yes)=4/9
P(age=g40|bc=yes)=3/9

P(age=b31|bc=no)=3/5
P(age=i31|bc=no)=0
P(age=g40|bc=no)=2/5

P(inc=high|bc=yes)=2/9
P(inc=med|bc=yes)=4/9
P(inc=low|bc=yes)=3/9

P(inc=high|bc=no)=2/5
P(inc=med|bc=no)=2/5
P(inc=low|bc=no)=1/5

P(std=yes|bc=yes)=6/9
P(std=no|bc=yes)=3/9

P(std=yes|bc=no)=1/5
P(std=no|bc=no)=4/5

P(cr=exc|bc=yes)=3/9
P(cr=fair|bc=yes)=6/9

P(cr=exc|bc=no)=3/5
P(cr=fair|bc=no)=2/5

P(bc=yes) = 9/14
P(bc=no) = 5/14

age	income	student	creditrating	buyscomputer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayes Classifier
Finding Classification

X: (age <= 30 , income = medium, student = yes, creditrating = fair)

P(X|bc=yes)=P(age=b30|bc=yes)*P(inc=med|bc=yes)*P(std=yes|bc=yes)*P(cr=fair|bc=yes)

= 2/9 * 4/9 * 6/9 * 6/9 = 0.044

P(X|bc=no)=P(age=b30|bc=no)*P(inc=med|bc=no)*P(std=yes|bc=no)*P(cr=fair|bc=no)

= 3/5 * 2/5 * 1/5 * 2/5 = 0.019

P(X|bc=yes)*P(bc=yes) = 0.044 * 9/14 = 0.028

P(X|bc=no)*P(bc=no) = 0.019 * 5/14 = 0.007

➔ Therefore, X belongs to class "buyscomputer = yes"

Confidence of the classification: 0.028/(0.028+0.007) = 0.80 80%

Naïve Bayes Classifier - Example

- Compute all probabilities for Naïve Bayes Classifier.
- Find the classification of the tuple (A=T,B=F)
- What is the confidence of that classification?

A	B	Class
T	F	Yes
T	F	No
T	T	Yes
T	F	Yes
F	T	Yes
F	T	No
F	F	No

Naïve Bayes Classifier - Example

P(yes) = 4/7 P(no) = 3/7

P(A=T|yes) = 3/4 P(A=T|no) = 1/3
P(A=F|yes) = 1/4 P(A=F|no) = 2/3

P(B=T|yes) = 2/4 P(B=T|no) = 1/3
P(B=F|yes) = 2/4 P(B=F|no) = 2/3

P(A=T,B=F|yes) = P(A=T|yes)*P(B=F|yes) = 3/4 * 2/4 = 6/16
P(A=T,B=F|no) = P(A=T|no)*P(B=F|no) = 1/3 * 2/3 = 2/9

P(A=T,B=F|yes)*P(yes) = 6/16 * 4/7 = 24/112 = 0.214
P(A=T,B=F|no)*P(no) = 2/9 * 3/7 = 6/63 = 0.095

Classification is YES

Confidence: 0.214 / (0.214+0.095) = 0.69 69%

A	B	Class
T	F	Yes
T	F	No
T	T	Yes
T	F	Yes
F	T	Yes
F	T	No
F	F	No

Avoiding the Zero-Probability Problem

- Naïve Bayesian prediction requires each conditional probability to be a **non-zero value**. Otherwise, the predicted probability will be zero

$$P(x_1, x_2, \ldots, x_n \mid C_i) = P(x_1 \mid C_i) * P(x_2 \mid C_i) * \ldots * P(x_n \mid C_i)$$

- In order to avoid zero probability values, we apply **smoothing techniques**.
- One of these smoothing techniques is **add-one smoothing (Laplacian correction)**.

	Smoothed Values
$P(A=v1 \mid C_i) = N_{v1Ci} / N_{ci}$	$\mathbf{P(A=v1 \mid C_i) = (N_{v1Ci} + 1) / (N_{ci}+3)}$
$P(A=v2 \mid C_i) = N_{v2Ci} / N_{ci}$	$\mathbf{P(A=v2 \mid C_i) = (N_{v2Ci} + 1) / (N_{ci}+3)}$
$P(A=v3 \mid C_i) = N_{v3Ci} / N_{ci}$	$\mathbf{P(A=v1 \mid C_i) = (N_{v3Ci} + 1) / (N_{ci}+3)}$

Avoiding the Zero-Probability Problem

P(age=b31|bc=no)=3/5

P(age=i31|bc=no)=0

P(age=g40|bc=no)=2/5

After add-one smoothing:

P(age=b31|bc=no)=(3+1)/(5+3) = 4/8

P(age=i31|bc=no)=(0+1)/(5+3) = 1/8

P(age=g40|bc=no)=(2+1)/(5+3) = 3/8

Naïve Bayes Classifier: Comments

- Advantages
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables
    - E.g., hospitals: patients: Profile: age, family history, etc.
      Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
    - Dependencies among these cannot be modeled by Naïve Bayes Classifier
- How to deal with these dependencies? Bayesian Belief Networks