Data Mining
Unit 3

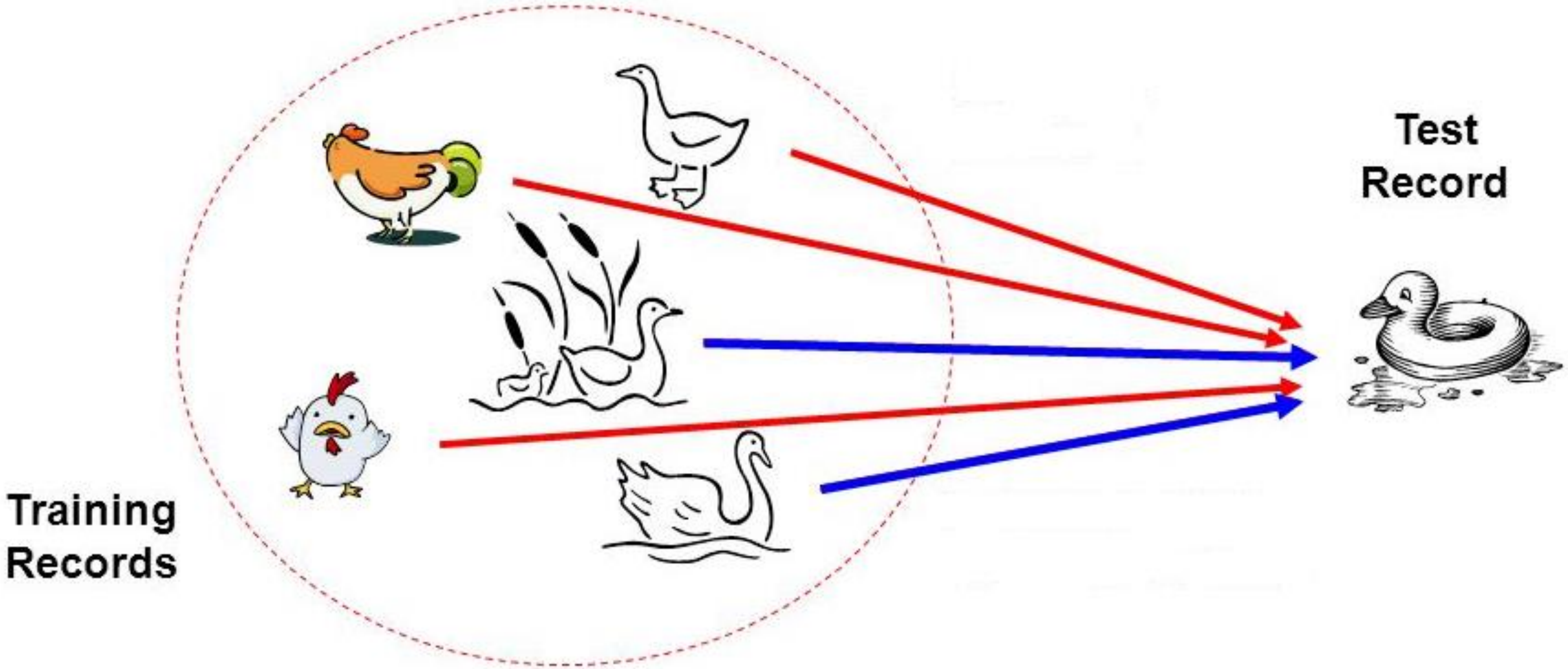
Statistical Classifier: Naïve Bayes’ Classifier

Slides credit: Dr. Debasis Samanta (IITkgp)

Bayesian Classifier

Bayesian Classifier

- Principle
  - If it walks like a duck, quacks like a duck, then it is probably a duck

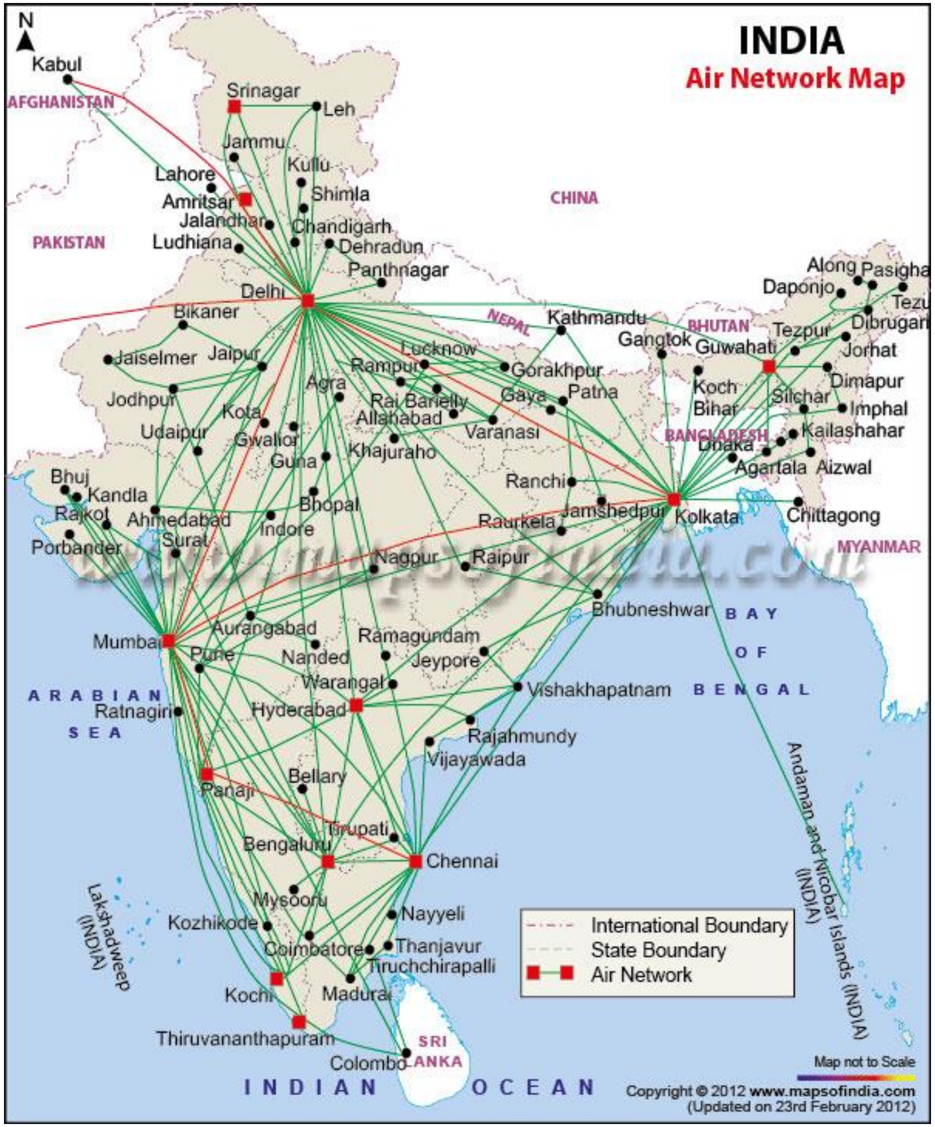


Bayesian Classifier

- A statistical classifier
  - Performs *probabilistic prediction*, *i.e.,* predicts class membership probabilities
- Foundation
  - Based on Bayes' Theorem.
- Assumptions
  1. The classes are mutually exclusive and exhaustive.
  2. The attributes are independent given the class.
- Called "Naïve" classifier because of these assumptions.
  - Empirically proven to be useful.
  - Scales very well.

Example: Bayesian Classification

- **Example 8.2:** Air Traffic Data
  - Let us consider a set observation recorded in a database
    - Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



Copyright © 2012 www.mapsofindia.com
(Updated on 23rd February 2012)

Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time

Cond. to next slide…

Air-Traffic Data

Cond. from previous slide…

Days	Season	Fog	Rain	Class
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Heavy	On Time

Air-Traffic Data

- In this database, there are four attributes

  A = [Day, Season, Fog, Rain]

  with 20 tuples.

- The categories of classes are:

  C= [On Time, Late, Very Late, Cancelled]

- Given this is the knowledge of data and classes, we are to find most likely classification for any other unseen instance, for example:

Week Day	**Winter**	**High**	**None**	**???**

- Classification technique eventually to map this tuple into an accurate class.

Bayesian Classifier

- In many applications, the relationship between the attributes set and the class variable is non-deterministic.
  - In other words, a test cannot be classified to a class label with certainty.
  - In such a situation, the classification can be achieved probabilistically.
- The Bayesian classifier is an approach for modelling probabilistic relationships between the attribute set and the class variable.
- More precisely, Bayesian classifier use Bayes' Theorem of Probability for classification.
- Before going to discuss the Bayesian classifier, we should have a quick look at the Theory of Probability and then Bayes' Theorem.

Bayes’ Theorem of Probability

Simple Probability

Definition 8.2: **Simple Probability**

If there are n elementary events associated with a random experiment and m of n of them are favorable to an event A, then the probability of happening or occurrence of A is

$$P(A) = \frac{m}{n}$$

Simple Probability

- Suppose, A and B are any two events and *P(A)*, *P(B)* denote the probabilities that the events *A* and *B* will occur, respectively.

- **Mutually Exclusive Events:**
  - Two events are mutually exclusive, if the occurrence of one precludes the occurrence of the other.

    Example: Tossing a coin (two events)

    Tossing a ludo cube (Six events)

💡 Can you give an example, so that two events are not mutually exclusive?

Hint: Tossing two identical coins, Weather (sunny, foggy, warm)

Simple Probability

- **Independent events:** Two events are independent if occurrences of one does not alter the occurrence of other.

Example: Tossing both coin and ludo cube together.

(How many events are here?)

💡 Can you give an example, where an event is dependent on one or more other events(s)?

Hint: Receiving a message (A) through a communication channel (B) over a computer (C), rain and dating.

Joint Probability

Definition 8.3: **Joint Probability**

If *P(A)* and *P(B)* are the probability of two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive, then $P(A \cap B) = 0$
If A and B are independent events, then $P(A \cap B) = P(A).P(B)$

Thus, for mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

Conditional Probability

Definition 8.2: **Conditional Probability**

If events are dependent, then their probability is expressed by conditional probability. The probability that A occurs given that B is denoted by $P(A|B)$.

Suppose, A and B are two events associated with a random experiment. The probability of A under the condition that B has already occurred and $P(B) \neq 0$ is given by

$$P(A|B) = \frac{\text{Number of events in } B \text{ which are favourable to } A}{\text{Number of events in } B}$$

$$= \frac{\text{Number of events favourable to } A \cap B}{\text{Number of events favourable to } B}$$

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

Conditional Probability

Corollary 8.1: **Conditional Probability**

$$P(A \cap B) = P(A).P(B|A), \quad if\ P(A) \neq 0$$

or $$P(A \cap B) = P(B).P(A|B), \quad if\ P(B) \neq 0$$

For three events *A*, *B* and *C*

$$P(A \cap B \cap C) = P(A).P(B).P(C|A \cap B)$$

For *n* events A_1, A_2, …, A_n and if all events are mutually independent to each other

$$P(A_1 \cap A_2 \cap \ \dots\dots\dots\dots \cap A_n) = P(A_1).P(A_2)\dots\dots\dots\dots P(A_n)$$

Note:

$P(A|B) = 0$ if events are **mutually exclusive**

$P(A|B) = P(A)$ if *A* and *B* are **independent**

$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ otherwise**,**

$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{B} \cap \mathrm{A})$

Conditional Probability

- Generalization of Conditional Probability:

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

$$= \frac{\mathrm{P(B|A) \cdot P(A)}}{\mathrm{P(B)}} \qquad \because \mathrm{P(A \cap B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)}$$

By the law of total probability : P(B) = $\mathrm{P[(B \cap A) \cup (B \cap \bar{A})]}$, where $\bar{\mathrm{A}}$ denotes the compliment of event A. Thus,

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A) \cdot P(A)}}{\mathrm{P[(B \cap A) \cup (B \cap \bar{A})]}}$$

$$= \frac{\mathrm{P(B|A) \cdot P(A)}}{\mathrm{P(B|A) \cdot P(A) + P(B \mid \bar{A}) \cdot P(\bar{A})}}$$

Conditional Probability

In general,

$$\mathrm{P(A|D)} = \frac{\mathrm{P(A)} \ \cdot \mathrm{P(D|A)}}{\mathrm{P(A)} \cdot \mathrm{P(D|A)} + \mathrm{P(B)} \cdot \mathrm{P(D|B)} + \mathrm{P(C)} \cdot \mathrm{P}\left(\mathrm{D} \,\middle|\, \mathrm{C}\right)}$$

Total Probability

Definition 8.3: **Total Probability**

Let $E_1, E_2, \ldots \ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 or E_2 or $\ldots \ldots E_n$, then

$$P(A) = P(E_1).P(A|E_1) + P(E_2).P(A|E_2) + \cdots \ldots \ldots \ldots. + P(E_n).P(A|E_n)$$

Total Probability: An Example

Example 8.3

A bag contains 4 red and 3 black balls. A second bag contains 2 red and 4 black balls. One bag is selected at random. From the selected bag, one ball is drawn. What is the probability that the ball drawn is red?

This problem can be answered using the concept of Total Probability

E_1 =Selecting bag *I*

E_2 =Selecting bag *II*

A = Drawing the red ball

Thus, $P(A) = P(E_1).P(A|E_1) + P(E_2).P(A|E_2)$

where, $P(A|E_1)$ = Probability of drawing red ball when first bag has been chosen

and $P(A|E_2)$ = Probability of drawing red ball when second bag has been chosen

Reverse Probability

Example 8.3:

A bag (Bag I) contains 4 red and 3 black balls. A second bag (Bag II) contains 2 red and 4 black balls. You have chosen one ball at random. It is found as red ball. What is the probability that the ball is chosen from Bag I?

Here,

E_1 = Selecting bag I

E_2 = Selecting bag II

A = Drawing the red ball

We are to determine P(E_1|A). Such a problem can be solved using Bayes' theorem of probability.

Bayes' Theorem

Theorem 8.4: **Bayes' Theorem**

Let $E_1, E_2, \ldots \ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with $E_1\ or\ E_2\ or \ldots \ldots E_n$, then

$$P(E_i|A) = \frac{P(E_i).P(A|E_i)}{\sum_{i=1}^{n} P(E_i).P(A|E_i)}$$

Prior and Posterior Probabilities

- P(A) and P(B) are called prior probabilities
- P(A|B), P(B|A) are called posterior probabilities

Example 8.6: Prior versus Posterior Probabilities

- This table shows that the event *Y* has two outcomes namely *A* and *B*, which is dependent on another event *X* with various outcomes like x_1, x_2 and x_3.

- **Case1:** Suppose, we don't have any information of the event *A*. Then, from the given sample space, we can calculate $P(Y = A) = \frac{5}{10} = 0.5$
.

- **Case2:** Now, suppose, we want to calculate $P(X = x_2 | Y = A) = \frac{2}{5} = 0.4$.

The later is the conditional or posterior probability, where as the former is the prior probability.

X	*Y*
x_1	*A*
x_2	*A*
x_3	*B*
x_3	*A*
x_2	*B*
x_1	*A*
x_1	*B*
x_3	*B*
x_2	*B*
x_2	*A*

Naïve Bayesian Classifier

- Suppose, *Y* is a class variable and *X* = $\{X_1, X_2, \ldots\ldots, X_n\}$ is a set of attributes, with instance of *Y*.

INPUT (X)	CLASS(Y)
… … …	
… … …	…
$x_1, x_2, \ldots, x_n$	y_i
… … …	…

- The classification problem, then can be expressed as the class-conditional probability

$$P\big(Y = y_i | (X_1 = x_1) \text{ AND } (X_2 = x_2) \text{ AND } \ldots\ldots (X_n = x_n)\big)$$

Naïve Bayesian Classifier

- Naïve Bayesian classifier calculate this posterior probability using Bayes' theorem, which is as follows.

- From Bayes' theorem on conditional probability, we have

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

$$= \frac{P(X|Y)\cdot P(Y)}{P(X|Y=y_1)\cdot P(Y=y_1)+\cdots+P(X|Y=y_k)\cdot P(Y=y_k)}$$

where,

$$P(X) = \sum_{i=1}^{k} P(X|Y=y_i)\cdot P(Y=y_i)$$

Note:

- $P(X)$ is called the evidence (also the total probability) and it is a constant.

- The probability *P(Y|X)* (also called class conditional probability) is therefore proportional to *P(X|Y)*· $P(Y)$.

- Thus, *P(Y|X)* can be taken as a measure of *Y* given that *X*.

$$P(Y/X) \approx P(X|Y)\cdot P(Y)$$

Naïve Bayesian Classifier

- Suppose, for a given instance of *X* (say *x* = (X_1 $= x_1$) and ….. ($X_n$$= x_n$)).

- There are any two class conditional probabilities namely *P(Y=* y_i */X=x)* and *P(Y=* y_j | *X=x)*.

- If *P(Y=* y_i | *X=x)* > *P(Y=* y_j | *X=x)*, then we say that y_i is more stronger than y_j for the instance *X* = *x*.

- The strongest y_i is the classification for the instance *X* = *x*.

Naïve Bayesian Classifier

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	½ = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	½ = 0.5	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	Holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
Season	Spring	4/14 = 0.29	0/2 = 0	0/3 = 0	0/1 = 0
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3= 0.33	0/1 = 0
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

Naïve Bayesian Classifier

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	None	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	Heavy	1/14 = 0.07	1/2 = 0.5	2/3 = 0.67	1/1 = 1
Prior Probability		14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

Naïve Bayesian Classifier

Instance:

Week Day	**Winter**	**High**	**Heavy**	**???**

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.50 = 0.0125

Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

Case3 is the strongest; Hence correct classification is **Very Late**

Naïve Bayesian Classifier

Algorithm: Naïve Bayesian Classification

Input: Given a set of *k* mutually exclusive and exhaustive classes *C* = $\{c_1, c_2, \ldots\ldots, c_k\}$, which have prior probabilities *P*(C_1), *P*(C_2),….. *P*(C_k).

There are *n*-attribute set *A* = $\{A_1, A_2, \ldots\ldots, A_n\}$, which for a given instance have values A_1 = a_1, A_2 = a_2,….., A_n = a_n

Step: For each c_i ∈ C, calculate the class condition probabilities, *i* = 1,2,…..,*k*

$$p_i = P(C_i) \times \prod_{j=1}^{n} P(A_j = a_j | C_i)$$

$$p_x = \max\{p_1, p_2, \ldots\ldots, p_k\}$$

Output: C_x is the classification

Note: $\boldsymbol{\sum p_i \neq 1}$, because they are not probabilities rather proportion values (to posterior probabilities)

Naïve Bayesian Classifier

Pros and Cons

- The Naïve Bayes’ approach is a very popular one, which often works well.
- However, it has a number of potential problems
    - It relies on all attributes being categorical.
    - If the data is less, then it estimates poorly.

Naïve Bayesian Classifier

Approach to overcome the limitations in Naïve Bayesian Classification

- Estimating the posterior probabilities for continuous attributes
  - In real life situation, all attributes are not necessarily be categorical, In fact, there is a mix of both categorical and continuous attributes.
  - In the following, we discuss the schemes to deal with continuous attributes in Bayesian classifier.

1. We can discretize each continuous attributes and then replace the continuous values with its corresponding discrete intervals.

2. We can assume a certain form of probability distribution for the continuous variable and estimate the parameters of the distribution using the training data. A Gaussian distribution is usually chosen to represent the posterior probabilities for continuous attributes. A general form of Gaussian distribution will look like

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

where, μ and σ^2 denote mean and variance, respectively.

Naïve Bayesian Classifier

For each class C_i, the posterior probabilities for attribute A_j (it is the numeric attribute) can be calculated following Gaussian normal distribution as follows.

$$P(A_j = a_j | C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(a_j - \mu ij)^2}{2\sigma_{ij}^2}}$$

Here, the parameter μ_{ij} can be calculated based on the sample mean of attribute value of A_j for the training records that belong to the class C_i.

Similarly, σ_{ij}^2 can be estimated from the calculation of variance of such training records.

Naïve Bayesian Classifier

M-estimate of Conditional Probability

- The M-estimation is to deal with the potential problem of Naïve Bayesian Classifier when training data size is too poor.
  - If the posterior probability for one of the attribute is zero, then the overall class-conditional probability for the class vanishes.
  - In other words, if training data do not cover many of the attribute values, then we may not be able to classify some of the test records.
- This problem can be addressed by using the M-estimate approach.

M-estimate Approach

- M-estimate approach can be stated as follows

$$P(A_{j=}a_j|C_i) = \frac{n_{c_i}+mp}{n+m}$$

where, n = total number of instances from class C_i

n_{c_i} = number of training examples from class C_i that take the value $A_{j=}a_j$

m = it is a parameter known as the equivalent sample size, and

p = is a user specified parameter.

Note:

If n = 0, that is, if there is no training set available, then $P(a_i/C_i)$ = p,

so, this is a different value, in absence of sample value.

A Practice Example

Example 8.4

Class:
C1:buys_computer = ‘yes’
C2:buys_computer = ‘no’

Data instance
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = fair)

age	income	student	credit_rating	comp
<=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

A Practice Example

- $P(C_i)$: P(buys_computer = “yes”) = 9/14 = 0.643
  P(buys_computer = “no”) = 5/14= 0.357

- Compute $P(X|C_i)$ for each class
  P(age = “<=30” | buys_computer = “yes”) = 2/9 = 0.222
  P(age = “<= 30” | buys_computer = “no”) = 3/5 = 0.6
  P(income = “medium” | buys_computer = “yes”) = 4/9 = 0.444
  P(income = “medium” | buys_computer = “no”) = 2/5 = 0.4
  P(student = “yes” | buys_computer = “yes) = 6/9 = 0.667
  P(student = “yes” | buys_computer = “no”) = 1/5 = 0.2
  P(credit_rating = “fair” | buys_computer = “yes”) = 6/9 = 0.667
  P(credit_rating = “fair” | buys_computer = “no”) = 2/5 = 0.4

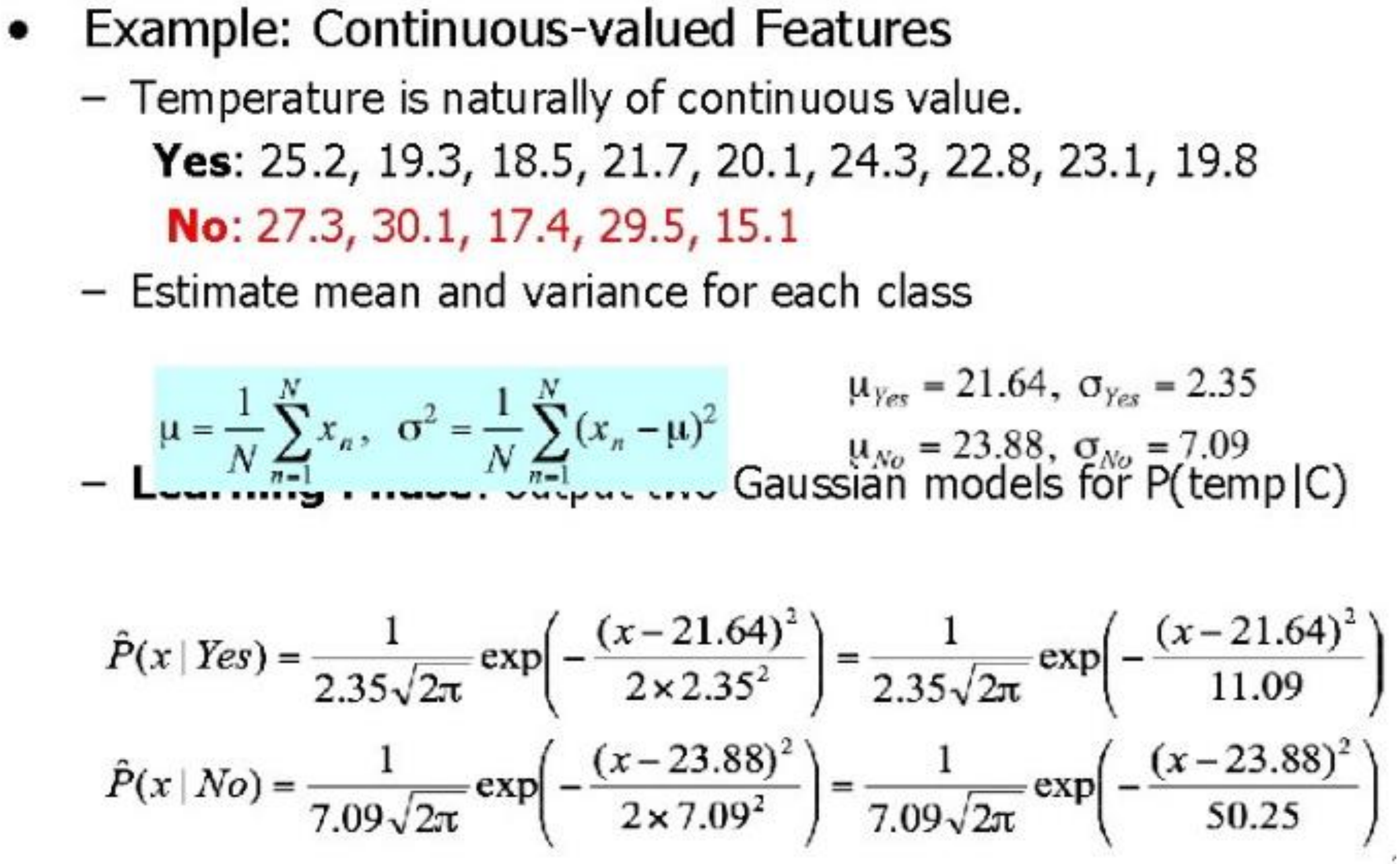
- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = “yes”) = 0.222 × 0.444 × 0.667 × 0.667 = 0.044
P(X|buys_computer = “no”) = 0.6 × 0.4 × 0.2 × 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028
P(X|buys_computer = “no”) * P(buys_computer = “no”) = 0.007

Therefore, X belongs to class (“buys_computer = yes”)

Estimating conditional probability for continuous values: example



- Example: Continuous-valued Features
  - Temperature is naturally of continuous value.

    Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

    No: 27.3, 30.1, 17.4, 29.5, 15.1
  - Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$

$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

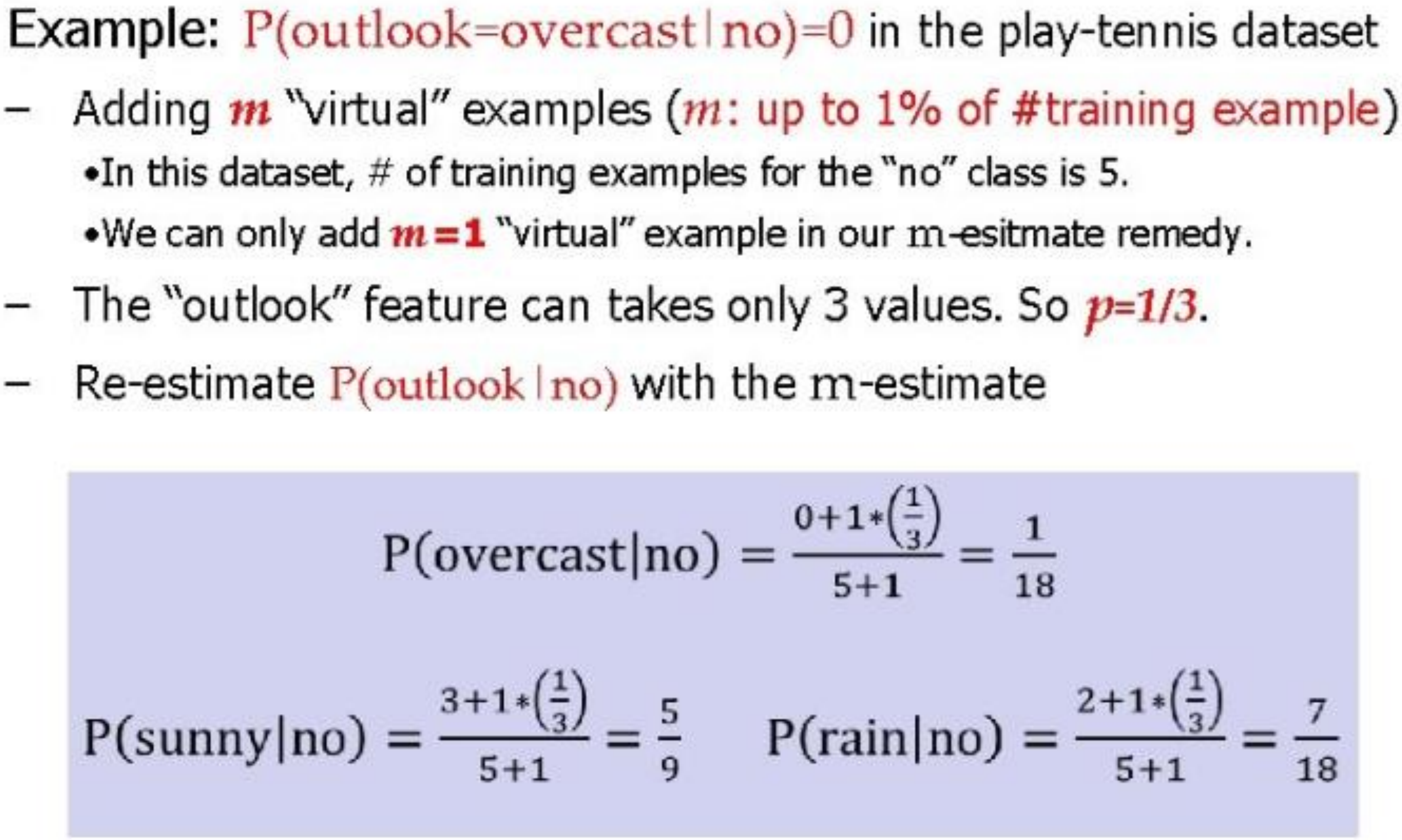
  - **Learning Phase:** output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}}\exp\left(-\frac{(x-21.64)^2}{2\times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}}\exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}}\exp\left(-\frac{(x-23.88)^2}{2\times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}}\exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

M-estimate: example

p: prior estimate of the probability
m: equivalent sample size (constant) In the absence of other information, assume a uniform prior: p =1/k where k is the number of values that the attribute a_i can take.



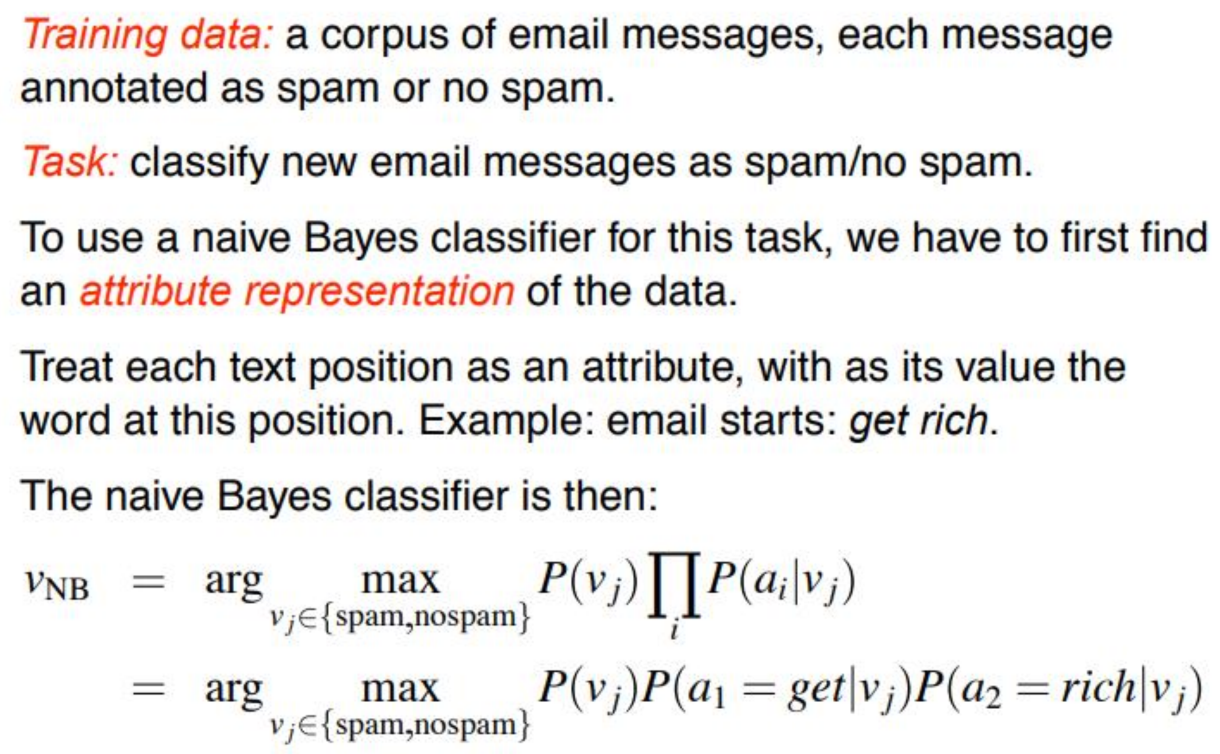
Example: P(outlook=overcast|no)=0 in the play-tennis dataset

- Adding *m* "virtual" examples (*m*: up to 1% of #training example)
  - In this dataset, # of training examples for the "no" class is 5.
  - We can only add **m=1** "virtual" example in our m-esitmate remedy.
- The "outlook" feature can takes only 3 values. So *p=1/3*.
- Re-estimate P(outlook|no) with the m-estimate

$$P(\text{overcast}|\text{no}) = \frac{0+1*\left(\frac{1}{3}\right)}{5+1} = \frac{1}{18}$$

$$P(\text{sunny}|\text{no}) = \frac{3+1*\left(\frac{1}{3}\right)}{5+1} = \frac{5}{9} \qquad P(\text{rain}|\text{no}) = \frac{2+1*\left(\frac{1}{3}\right)}{5+1} = \frac{7}{18}$$

E-mail classification



Training data: a corpus of email messages, each message annotated as spam or no spam.

Task: classify new email messages as spam/no spam.

To use a naive Bayes classifier for this task, we have to first find an *attribute representation* of the data.

Treat each text position as an attribute, with as its value the word at this position. Example: email starts: *get rich*.

The naive Bayes classifier is then:

$$\begin{aligned} v_{\mathrm{NB}} &= \arg \max_{v_j \in \{\mathrm{spam},\mathrm{nospam}\}} P(v_j) \prod_i P(a_i | v_j) \\ &= \arg \max_{v_j \in \{\mathrm{spam},\mathrm{nospam}\}} P(v_j) P(a_1 = get | v_j) P(a_2 = rich | v_j) \end{aligned}$$

Using naive Bayes means we assume that **words are independent of each** other. Clearly incorrect, but doesn’t hurt a lot for our task.

The classifier uses $P(a_i = w_k|v_j)$, i.e., the probability that the i-th word in the email is the k-word in our vocabulary, given the email has been classified as v_j.

Simplify by assuming that **position is irrelevant**: estimate $P(w_k|v_j)$, i.e., the probability that word w_k occurs in the email, given class v_j.

Create a **vocabulary:** make a list of all words in the training corpus, discard words with very high or very low frequency.

Training: estimate priors:

$$P(v_j) = \frac{n}{N}$$

Estimate likelihoods using the ***m*-estimate:**

$$P(w_k|v_j) = \frac{n_k+1}{n+|Vocabulary|}$$

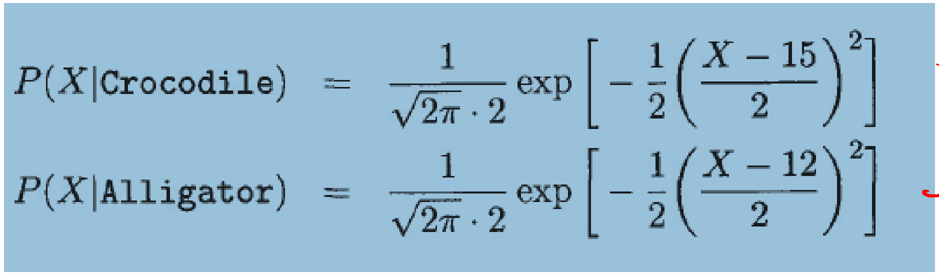
N: total number of words in all emails
n: number of words in emails with class v_j
n_k: number of times word w_k occurs in emails with class v_j
$|Vocabulary|$: size of the vocabulary

Testing: to classify a new email, assign it the class with the highest posterior probability. Ignore unknown words.

Bayes Error rate

- Suppose we know the true probability distribution that governs P(XlY).
- The Bayesian classification method allows us to determine the ideal decision boundary for the classification task
- Example:
  - Consider the task of identifying alligators and crocodiles based on their respective lengths. The average length of an adult crocodile is about 15 feet, while the average length of an adult alligator is about 12 feet

- Assuming that their length x follows a Gaussian distribution with a standard deviation equal to 2 feet, we can express their class-conditional probabilities as follows:



$$P(X|\texttt{Crocodile}) = \frac{1}{\sqrt{2\pi}\cdot 2}\exp\left[-\frac{1}{2}\left(\frac{X-15}{2}\right)^2\right]$$

$$P(X|\texttt{Alligator}) = \frac{1}{\sqrt{2\pi}\cdot 2}\exp\left[-\frac{1}{2}\left(\frac{X-12}{2}\right)^2\right]$$

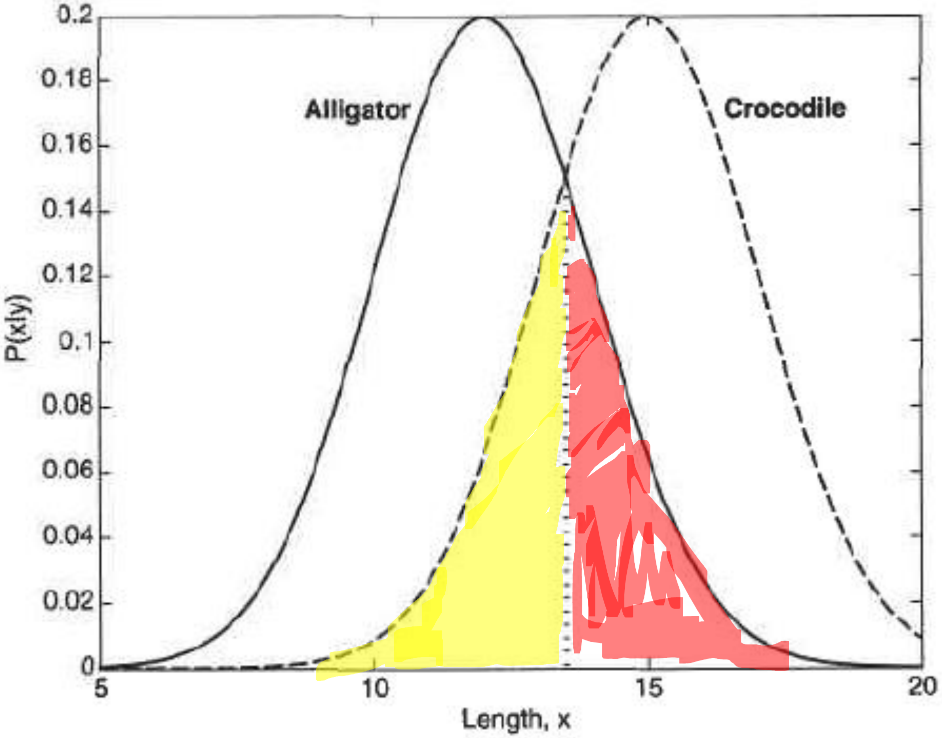
- Assuming that their prior probabilities are the same, the ideal decision boundary is located at some length $\hat{x}$ such that

$$P(X=\hat{x}|\texttt{Crocodile}) = P(X=\hat{x}|\texttt{Alligator})$$

Using above equations:

$$\left(\frac{\hat{x}-15}{2}\right)^2 = \left(\frac{\hat{x}-12}{2}\right)^2,$$

which can be solved to yield $\hat{x} = 13.5$.
The decision boundary for this example
is located halfway between the two means.



What happens if the prior probabilities are different?

- The ideal decision boundary in the preceding example classifies all creatures whose lengths are less than $\hat{x}$ as alligators and those whose lengths are greater than $\hat{x}$ as crocodiles.
- The error rate of the classifier is given by the sum of the area under the posterior probability curve for crocodiles (from length 0 to $\hat{x}$) and the area under the posterior probability curve for alligators (from $\hat{x}$ to ∞):

$$\text{Error} = \int_0^{\hat{x}} P(\texttt{Crocodile}|X)dX + \int_{\hat{x}}^{\infty} P(\texttt{Alligator}|X)dX.$$

- The total error rate is known as the **Bayes error rate**.

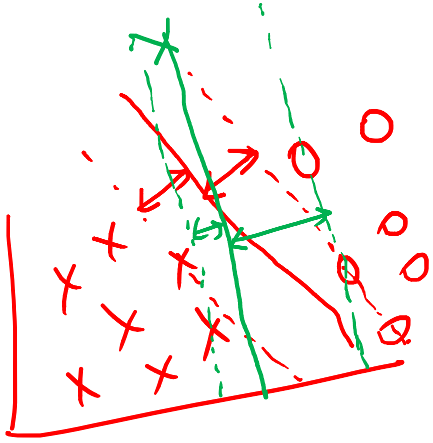
Multiclass Classification

- One vs. rest approach
- One against one approach

K classes

For each class $y_i \in Y$ a binary classifier. B_i

$(1 - r)$ approach $\equiv$



SVM