Data Mining:

3. Klassifikation

D) Alternative Techniques II: Bayes Classifier etc.

- Naive Bayesian Classifier
  - Bayesian [Belief] Networks (BBN)
  - Artificial Neural Networks (ANN)
- Support Vector Machines (SVM)
- Ensemble Methods

Bayesian Classifiers

- A probabilistic framework for solving classification problems
- Let A, C be random variables.
- Conditional Probability:

$$P(C \mid A) = \frac{P(A, C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A, C)}{P(C)}$$

- Bayes' theorem says:

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

Bayesian Classifiers

- Consider each input attribute A_i, $i=1,\ldots,n$, and the class attribute C as random variables

- Given a record with attributes $(A_1=a_1, A_2=a_2, \ldots, A_n=a_n)$
  - Goal is to predict class C
  - Specifically, we want to find the value c of C that maximizes $P(C=c \mid A_1=a_1, A_2=a_2, \ldots, A_n=a_n)$

- Can we estimate $P(C \mid A_1, A_2, \ldots, A_n)$ directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Choose value c of C that maximizes

$$P(C=c \mid A_1=a_1, A_2=a_2, \ldots, A_n=a_n)$$

  - Equivalent to choosing value c of C that maximizes

$$P(A_1=a_1, A_2=a_2, \ldots, A_n=a_n \mid C=c)\, P(C=c)$$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1| C) P(A_2| C) \ldots P(A_n| C)$

- Estimate $P(A_i| C)$ for all values of C and of all A_i from the training data .

- Classify new record $(A_1=a_1, A_2=a_2, \ldots, A_n=a_n)$ as C=c iff $\Pi_i P(A_i=a_i| C=c) \cdot P(C=c)$ is maximal.

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Cheat
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

- Class: $P(C=c) = N_c/N$
  - e.g., $P(No) = 7/10$, $P(Yes) = 3/10$

    ["C=" omitted here and in the sequel]
- For discrete attributes:

  $$P(A_i=a|C=c) = N_{ac}/ N_c$$

  - where N_{ac} is the number of records that have attribute value $A_i=a$ and belong to class C=c
  - Examples: ["C=" omitted]

    P(Status=Married | No) = 4/7
    P(Refund=Yes | Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into intervals
    - introduce one ordinal attribute per interval
    - would violate independence assumption
  - Better: Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, use it to estimate the conditional probability $P(A_i|C)$ ……

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Cheat
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

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- For (Income, No):
  - Check Income values of records with Cheat=No
    - sample mean μ=110
    - sample variance σ^2=2975
    - sample std.deviation σ=54.54

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier Application

Given a test record:

$$X = (\text{Refund} = \text{No}, \text{Mar.St.} = \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:

If class=No:	sample mean=110	
	sample variance=2975	
If class=Yes:	sample mean=90	
	sample variance=25	

- P(X|No) = P(Refund=No|No) $\times$ P(Mar.St.=Married| No) $\times$ P(Income=120K| No)
  $= 4/7 \times 4/7 \times 0.0072 = 0.0024$
- P(X|Yes) = P(Refund=No| Yes) $\times$ P(Mar.St.=Married| Yes) $\times$ P(Income=120K| Yes)
  $= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Naïve Bayes Classifier

- If one of the conditional probabilities is zero, then the entire expression becomes zero
- Better use other probability estimations:

$$\text{Original}: P(A_i = a \mid c) = \frac{N_{ac}}{N_c}$$

$$\text{Laplace}: P(A_i = a \mid c) = \frac{N_{ac} + 1}{N_c + \#_i}$$

$$\text{m-Estimate}: P(A_i = a \mid c) = \frac{N_{ac} + mp}{N_c + m}$$

$\#_i$: number of possible values of the variable in question, here A_i

p: prior probability

m: parameter

- For the example above: same result, but more robust computation; for m=3, p=2/3: P(X|No)=0.0026; for m=3, p=1/3: P(X|Yes)=1.3×10^{-10}

Example of Naïve Bayes Classifier

Give Birth	Lay Eggs	Can Fly	Live in Water	Have Legs	Class
yes	no	no	yes	no	?

Name	Give Birth	Lay Eggs	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	no	yes	mammals
python	no	yes	no	no	no	non-mammals
salmon	no	yes	no	yes	no	non-mammals
whale	yes	no	no	yes	no	mammals
frog	no	yes	no	sometimes	yes	non-mammals
komodo	no	yes	no	no	yes	non-mammals
bat	yes	no	yes	no	yes	mammals
pigeon	no	yes	yes	no	yes	non-mammals
cat	yes	no	no	no	yes	mammals
leopard shark	yes	no	no	yes	no	non-mammals
turtle	no	yes	no	sometimes	yes	non-mammals
penguin	no	yes	no	sometimes	yes	non-mammals
porcupine	yes	no	no	no	yes	mammals
eel	no	yes	no	yes	no	non-mammals
salamander	no	yes	no	sometimes	yes	non-mammals
gila monster	no	yes	no	no	yes	non-mammals
platypus	no	yes	no	no	yes	mammals
owl	no	yes	yes	no	yes	non-mammals
dolphin	yes	no	no	yes	no	mammals
eagle	no	yes	yes	no	yes	non-mammals

Note: The attribute „Lay Eggs“ is redundant!

A: given attribute values

M: mammals (7 of 20)

N: non-mammals (13 of 20)

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Naïve Bayes: Summary

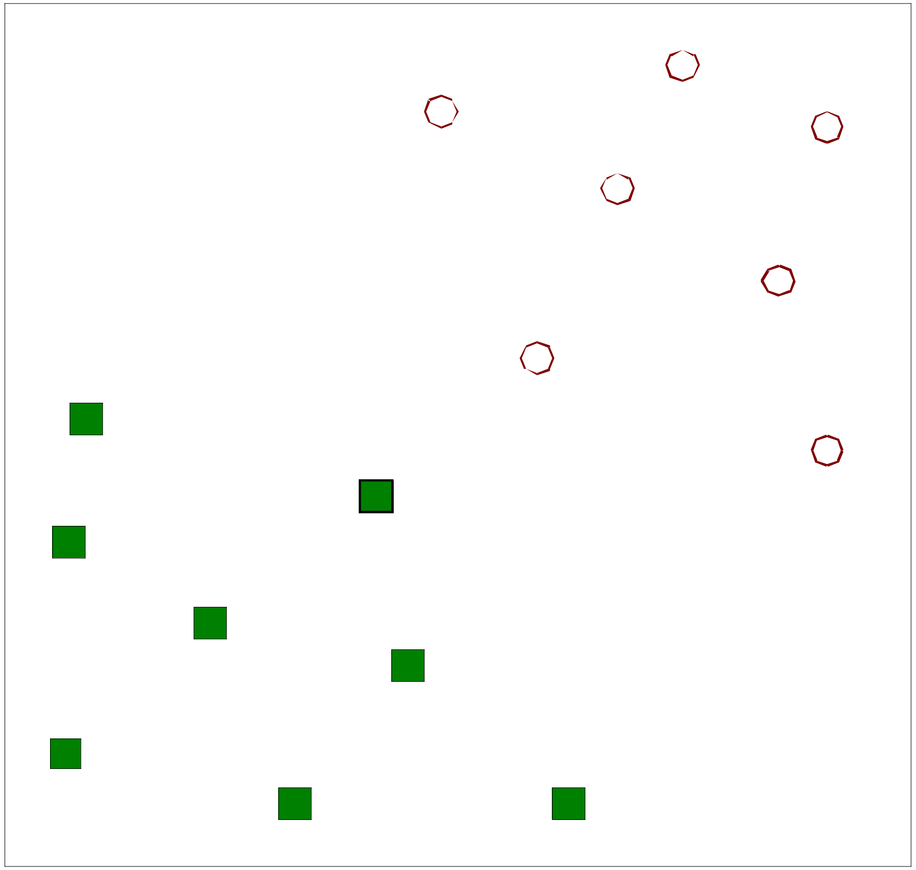
- Robust to isolated *noise points*

- Handle *missing values* by ignoring the instance during probability estimate calculations

- Robust to *irrelevant attributes,* since uniform distribution does not influence calculational decision.

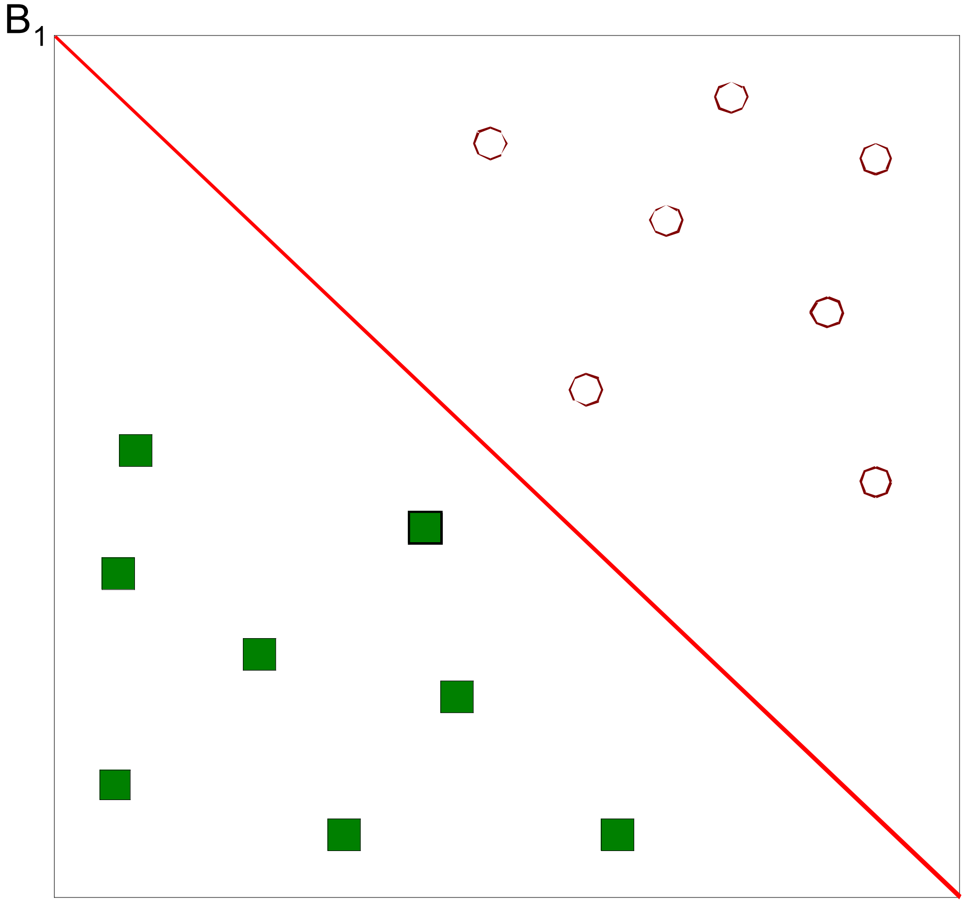
- Independence assumption may not hold for some attributes, i.e., *correlated attributes* can degrade classification success
  - Use other techniques such as Bayesian [Belief] Networks (BBN) encoding the dependencies among random variables

Support Vector Machines



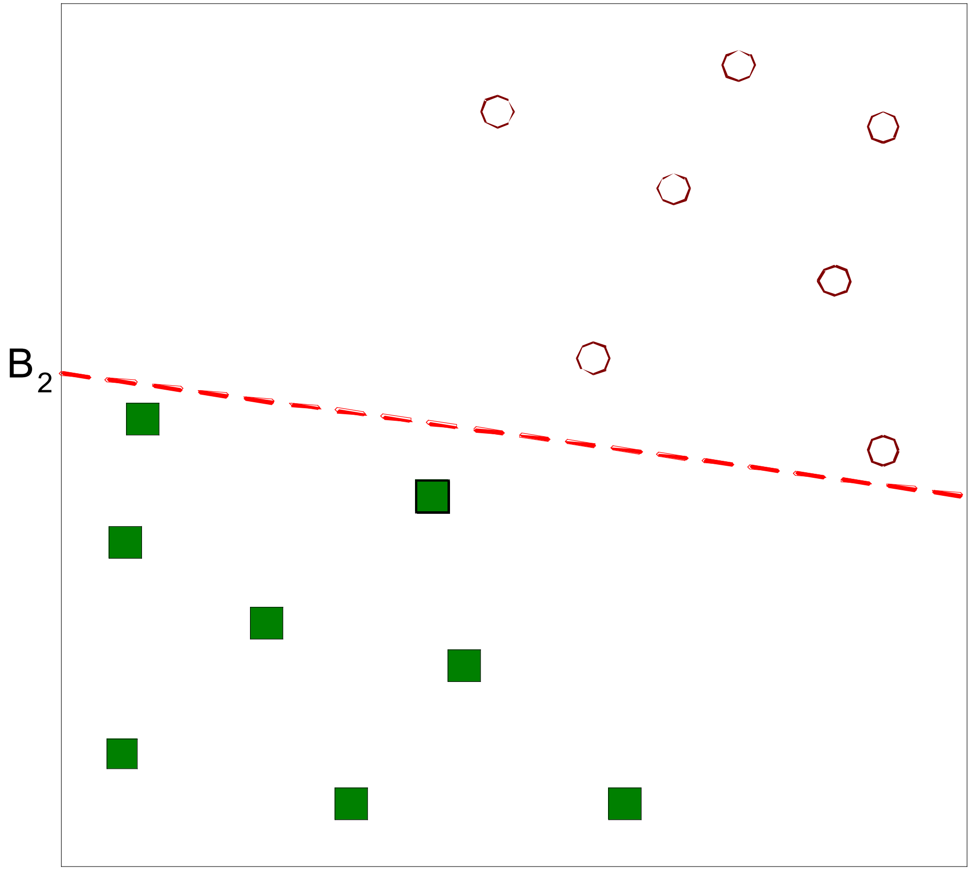
- Find a linear hyperplane (decision boundary) that will separate the data

Support Vector Machines



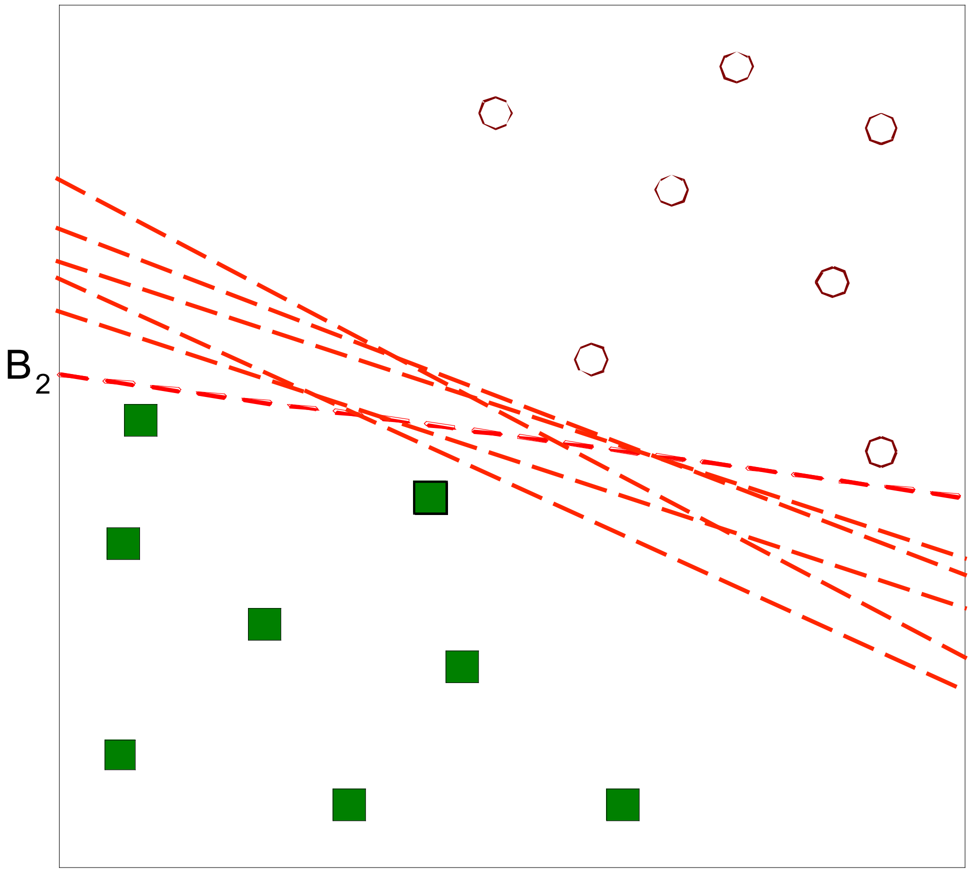
- One Possible Solution

Support Vector Machines



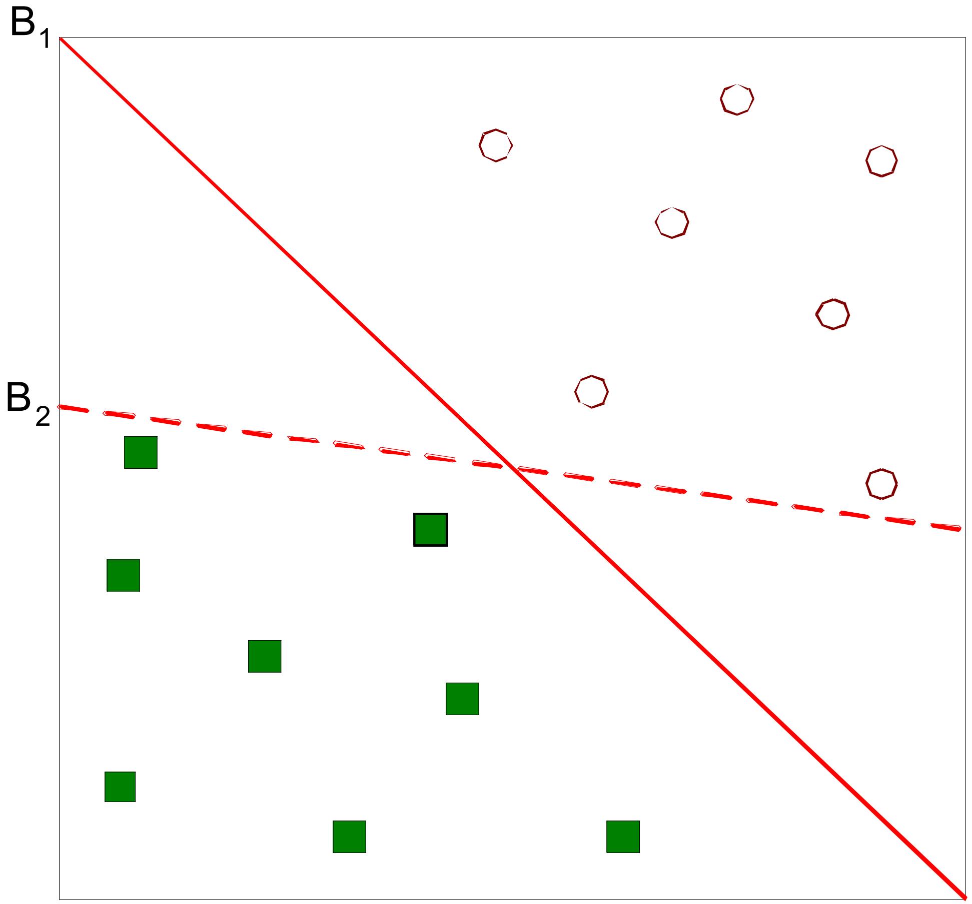
- Another possible solution

Support Vector Machines



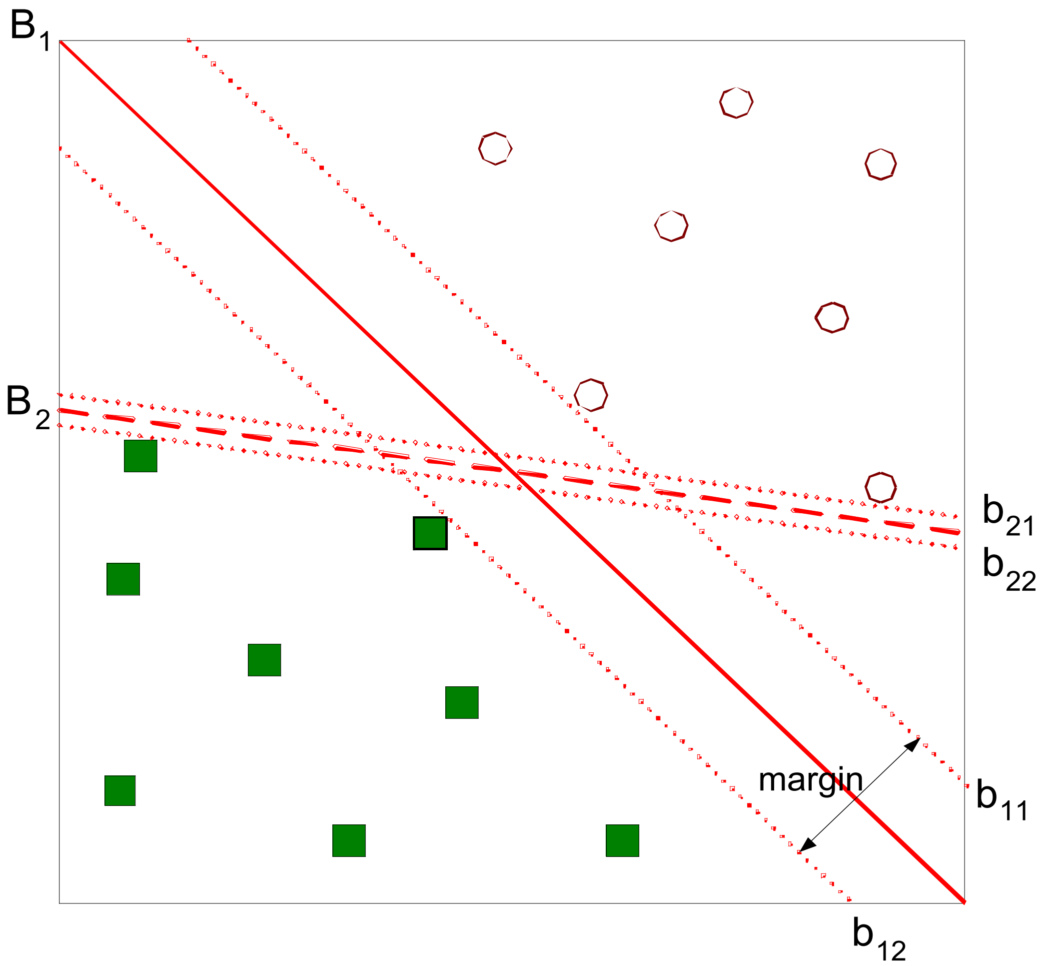
- Other possible solutions

Support Vector Machines



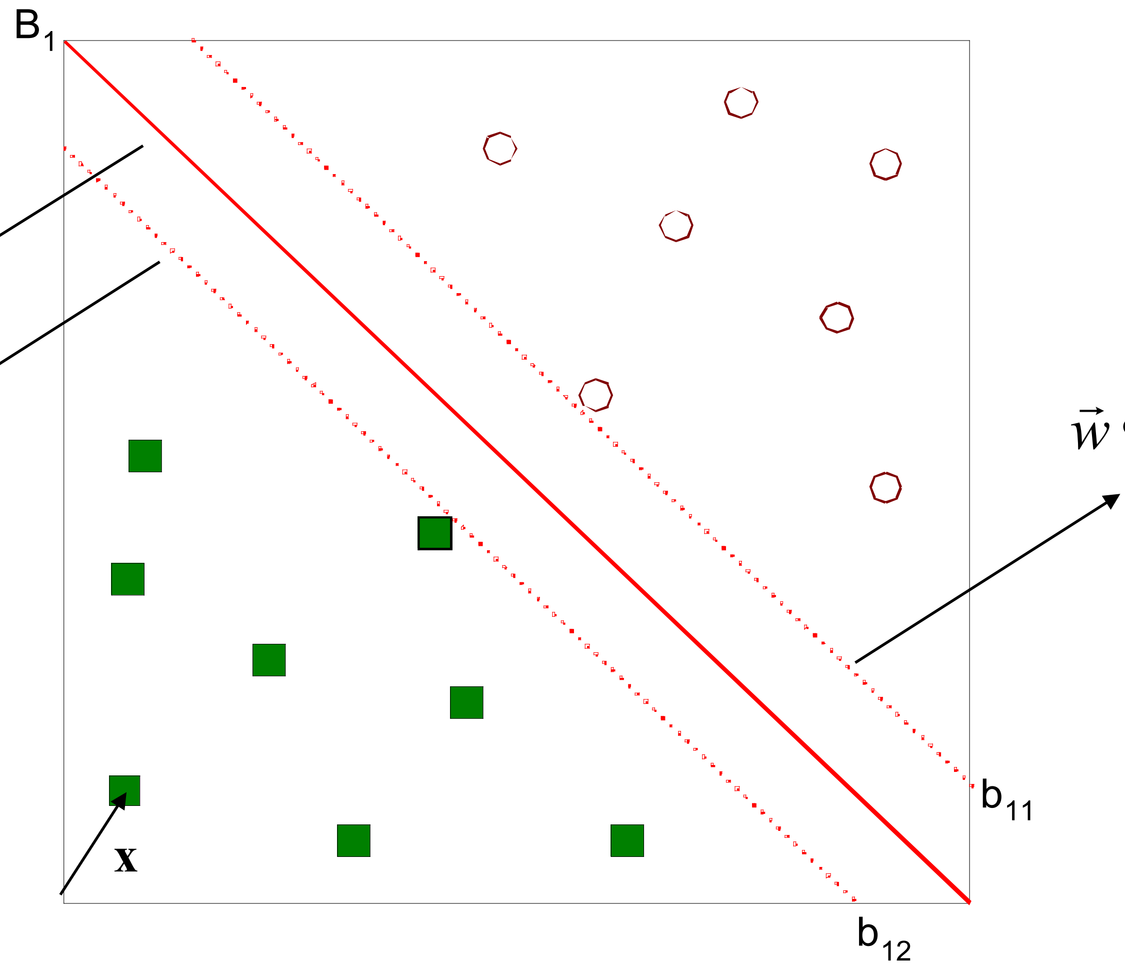
- Which one is better? B1 or B2?
- How do you define better?

Support Vector Machines



- Find hyperplane that maximizes the margin => B1 is better than B2

Support Vector Machines



Linear classifier $f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \leq -1 \end{cases}$

$$\text{Margin} = \frac{2}{\|\vec{w}\|^2}$$

Support Vector Machines

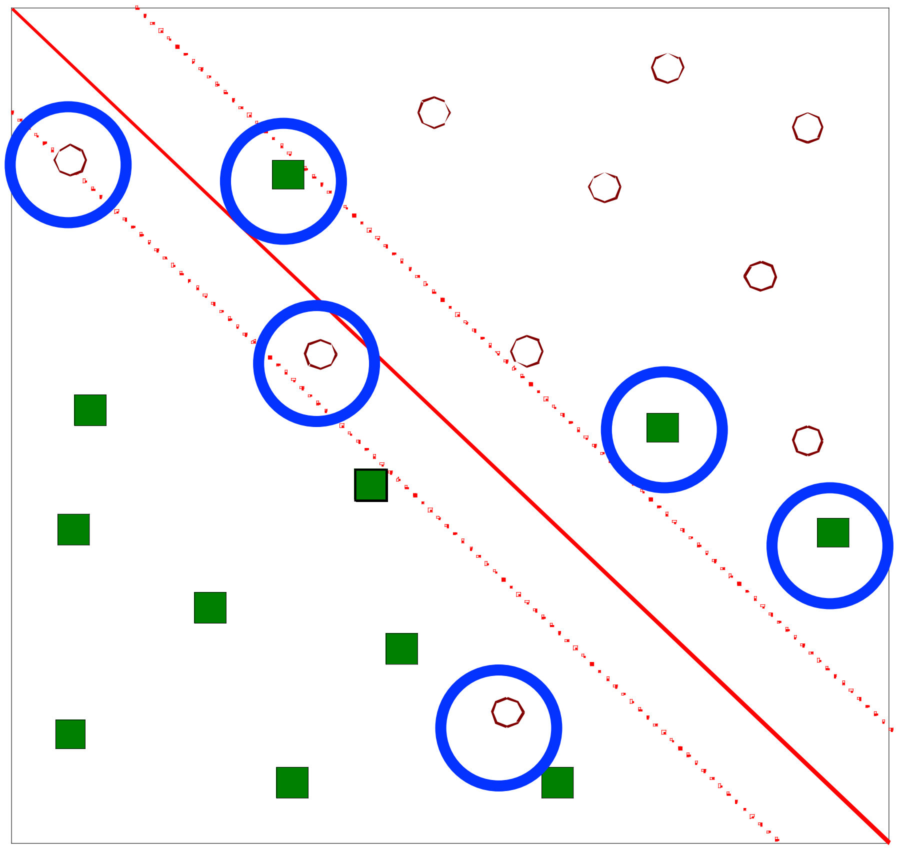
- We have to estimate the parameters w and b for the linear classifier f from the training data.
- We want to find a w that maximizes: $\text{Margin} = \dfrac{2}{\|\vec{w}\|^2}$
  - which is equivalent to minimizing: $L(w) = \dfrac{\|\vec{w}\|^2}{2}$
  - but subjected to the following constraint for the training data x_i:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \leq -1 \end{cases}$$

    - This is a constrained optimization problem
      - There are numerical approaches to solve it (e.g., quadratic programming).

Support Vector Machines

- What if the problem is linear, but nonseparable?



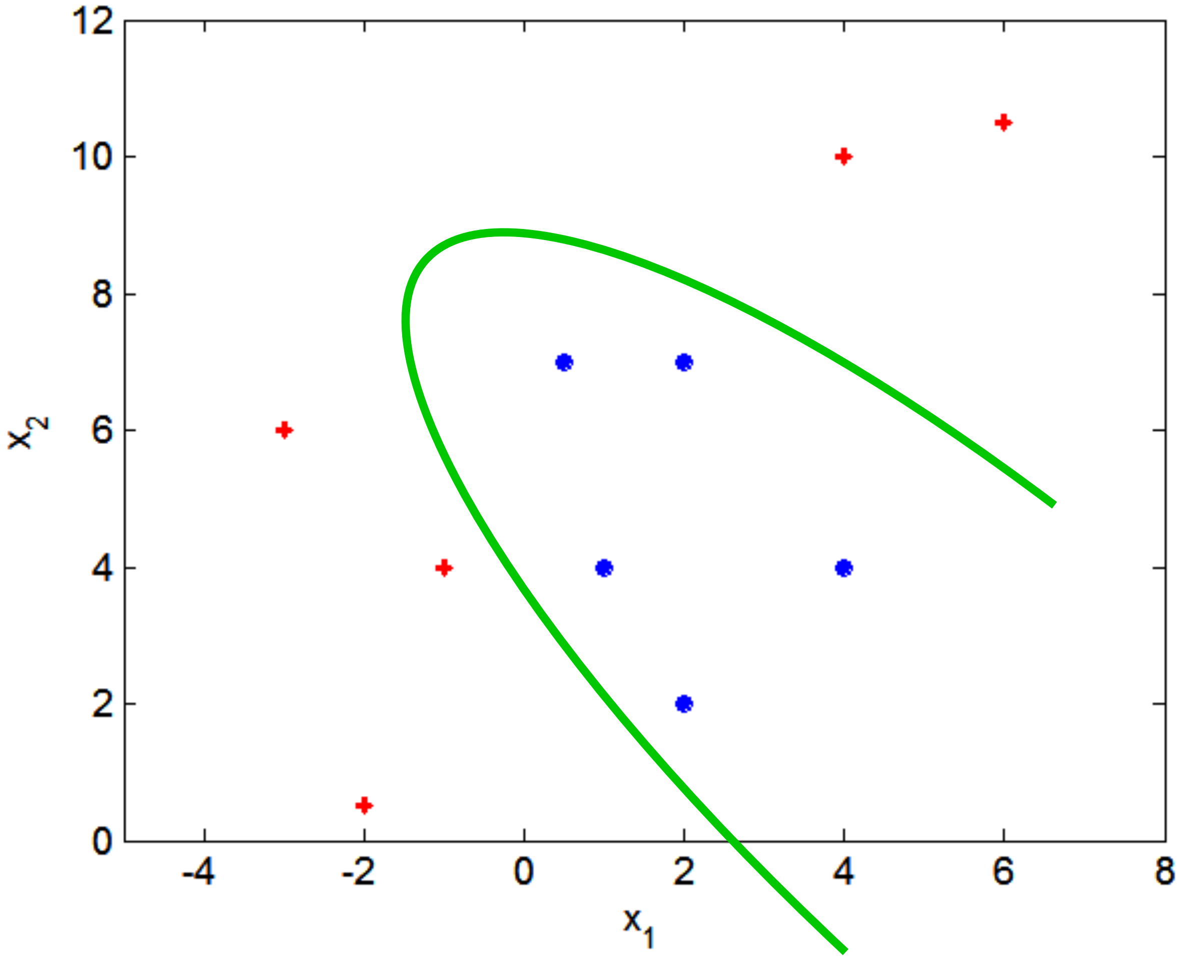
Support Vector Machines

- What if the problem is linear, but nonseparable?
  - Introduce slack variables (for deviations)
    - Need to minimize: $$L(w) = \frac{\| \vec{w} \|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i^k\right)$$
    - Subject to:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \geq 1 - \xi_\mathrm{i} \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \leq -1 + \xi_\mathrm{i} \end{cases}$$

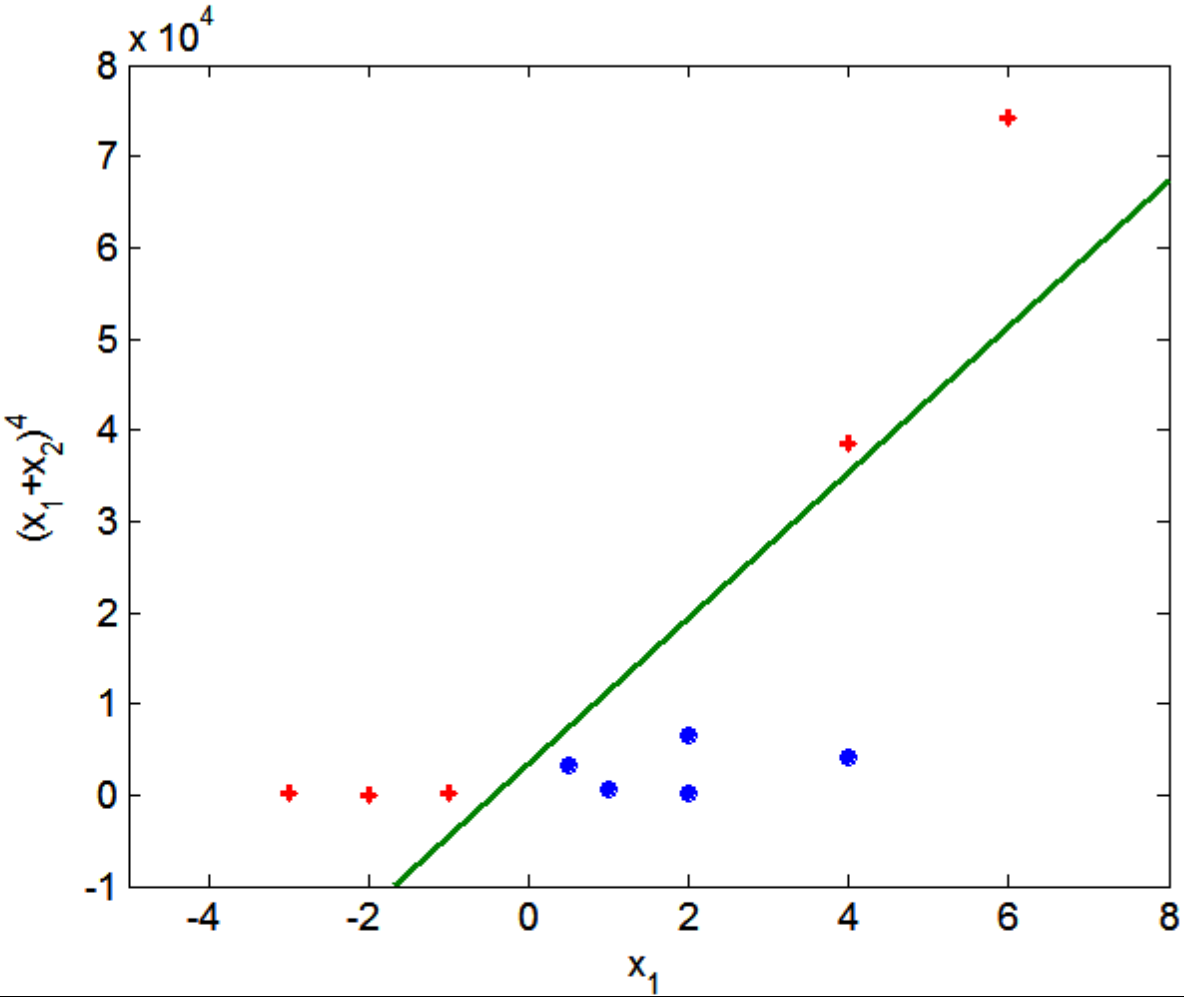
Nonlinear Support Vector Machines

- What if decision boundary is not linear?



Nonlinear Support Vector Machines

- Transform data into higher dimensional space

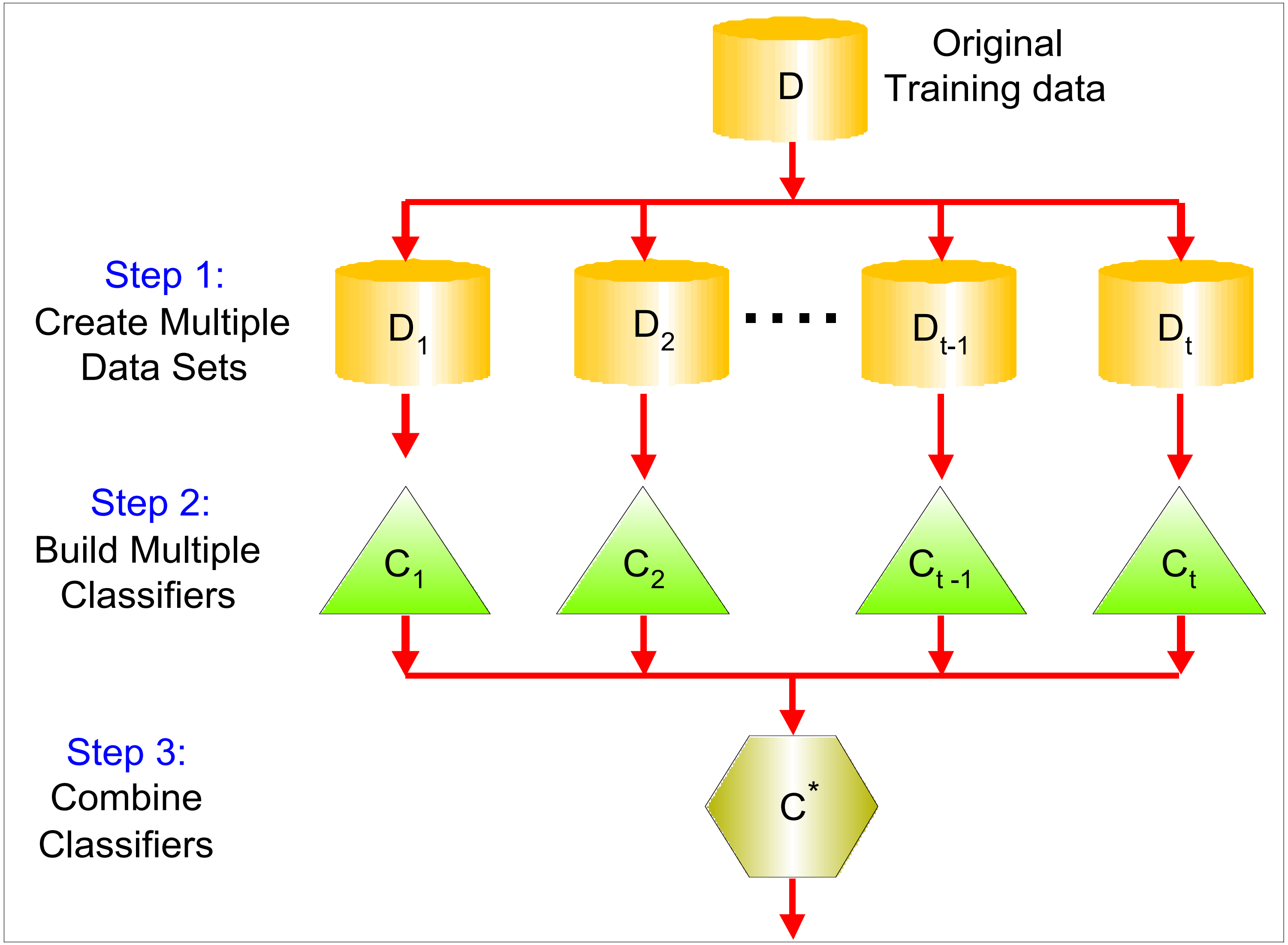


Ensemble Methods

- Construct a set of classifiers from the training data
  - by varying the training set
  - by varying the input attributes
  - by varying the class labels
    (binary labels built on top of multi-class labels)
  - by varying the classifier algorithm

- Predict class label of previously unseen records by aggregating predictions made by those multiple classifiers, e.g. using majority voting

General Idea



Why does it work?

- Suppose there are 25 base classifiers
  - Each classifier has error rate, $\varepsilon = 0.35$
  - Assume classifiers are independent
  - Probability that the ensemble classifier makes a wrong prediction:
    - i.e. the majority = 13 or more classifiers are wrong

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^{i} (1-\varepsilon)^{25-i} = 0.06 \quad < 0.35$$

How to generate an ensemble of classifiers? Cross Validation

Since classification labels are needed for training and validation, the training set is often multi-used:

- Holdout
  - → Reserve 2/3 for training and 1/3 for testing
  - – But: reduced training set
- Better: Random subsampling
  - → Repeated random holdout
  - – But: no control how often a record is used for training
- k-Fold Cross validation:
  - → Partition data into k disjoint subsets;
    train on k-1 partitions, test on the remaining one,
    do this for all partitions and combine the classifiers !
  - → Special case “leave-one-out”: k=number of training records

How to generate an ensemble of classifiers? Bagging

- Bagging (bootstrap aggregating) repeatedly samples (with replacement) from a training data set according to a uniform distribution:

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Build classifier on each bootstrap sample

- Each sample has probability $(1 – 1/n)^n$ of being selected; converges for large n to 63%.

Bagging (Example)

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 2:

x	0.1	0.2	0.3	0.4	0.5	0.8	0.9	1	1	1
y	1	1	1	-1	-1	1	1	1	1	1

x <= 0.65 ==> y = 1
x > 0.65 ==> y = 1

Bagging Round 3:

x	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 4:

x	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.3 ==> y = 1
x > 0.3 ==> y = -1

Bagging Round 5:

x	0.1	0.1	0.2	0.5	0.6	0.6	0.6	1	1	1
y	1	1	1	-1	-1	-1	-1	1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 6:

x	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 7:

x	0.1	0.4	0.4	0.6	0.7	0.8	0.9	0.9	0.9	1
y	1	-1	-1	-1	-1	1	1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 8:

x	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.8	0.9	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 9:

x	0.1	0.3	0.4	0.4	0.6	0.7	0.7	0.8	1	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 10:

x	0.1	0.1	0.1	0.1	0.3	0.3	0.8	0.8	0.9	0.9
y	1	1	1	1	1	1	1	1	1	1

x <= 0.05 ==> y = -1
x > 0.05 ==> y = 1

Figure 5.35. Example of bagging.

In this example, one-level decision trees are generated as classifiers.

Bagging (Example)

Round	x=0.1	x=0.2	x=0.3	x=0.4	x=0.5	x=0.6	x=0.7	x=0.8	x=0.9	x=1.0
1	1	1	1	-1	-1	-1	-1	-1	-1	-1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	-1	-1	-1	-1	-1	-1	-1
4	1	1	1	-1	-1	-1	-1	-1	-1	-1
5	1	1	1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	1	1	1
8	-1	-1	-1	-1	-1	-1	-1	1	1	1
9	-1	-1	-1	-1	-1	-1	-1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
Sum	2	2	2	-6	-6	-6	-6	2	2	2
Sign	1	1	1	-1	-1	-1	-1	1	1	1
True Class	1	1	1	-1	-1	-1	-1	1	1	1

Figure 5.36. Example of combining classifiers constructed using the bagging approach.

How to generate an ensemble of classifiers? Boosting

- An iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records
  - Initially, all N records are assigned equal weights
  - Unlike bagging, weights may change at the end of boosting round
- Experiments on famous data sets have shown that ensemble classifiers (generated by bagging or boosting) generally outperform a single deci-sion tree wrt accuracy.

Boosting

- Records that are wrongly classified will have their weights increased
- Records that are classified correctly will have their weights decreased

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

- Example 4 is hard to classify
- Its weight is increased, therefore it is more likely to be chosen again in subsequent rounds