DATA MINING
SUPERVISED LEARNING

Classification

- Nearest Neighbor Classifier
- Support Vector Machines (SVM)
- Naïve Bayes

NEAREST NEIGHBOR CLASSIFICATION

Instance-Based Classifiers

Set of Stored Cases

Atr1	………	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

- Store the training records
- Use training records to predict the class label of unseen cases

Unseen Case

Atr1	………	AtrN

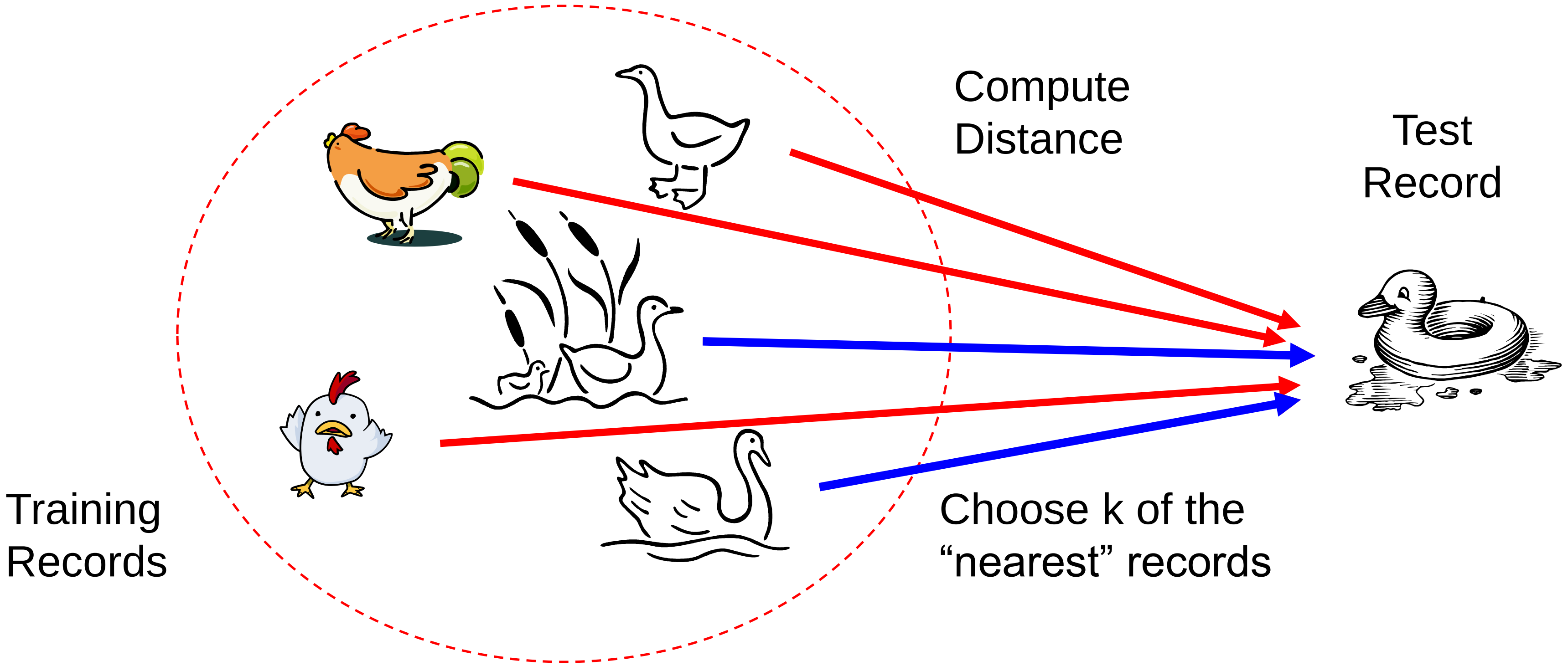
Instance Based Classifiers

- Examples:
  - Rote-learner（机械学习）
    - Memorizes entire training data and performs classification only if attributes of record match one of the training examples exactly
  - Nearest neighbor classifier
    - Uses k “closest” points (nearest neighbors) for performing classification

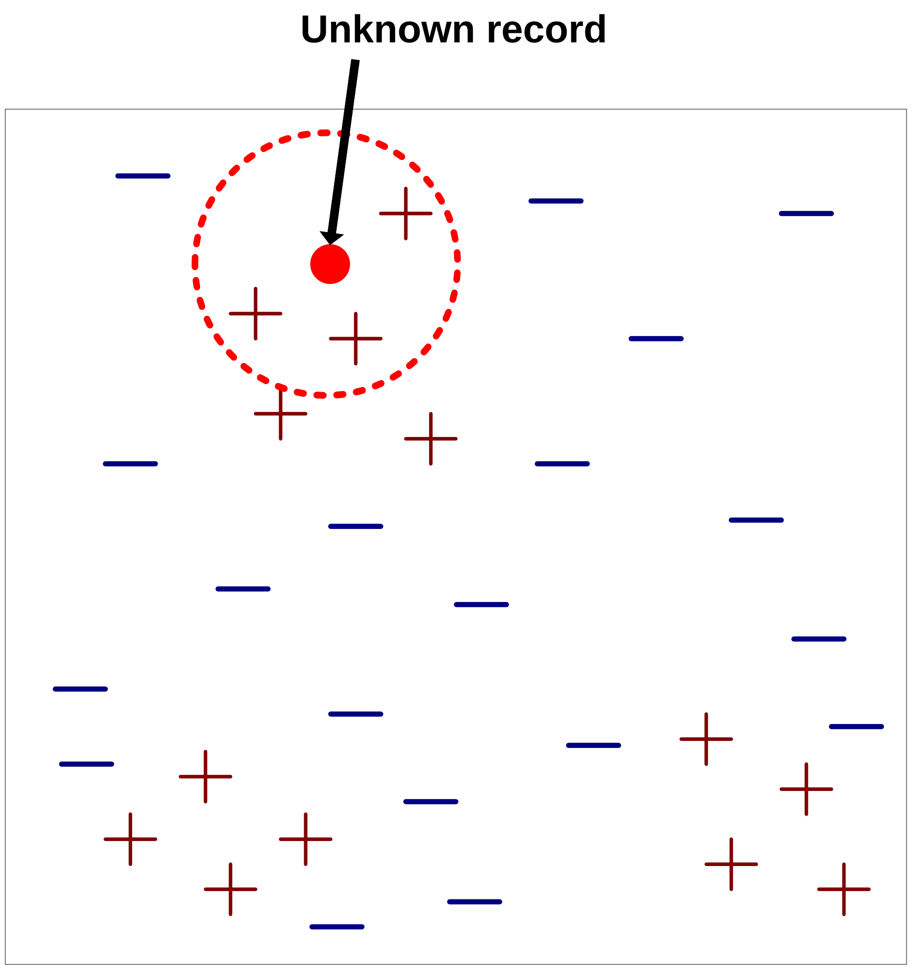
Nearest Neighbor Classifiers

- Basic idea:

《导论》P122



Nearest-Neighbor Classifiers



Requires three things

- The set of stored records
- Distance Metric to compute distance between records
- The value of *k*, the number of nearest neighbors to retrieve

To classify an unknown record:

1. Compute distance to other training records
2. Identify *k* nearest neighbors
3. Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

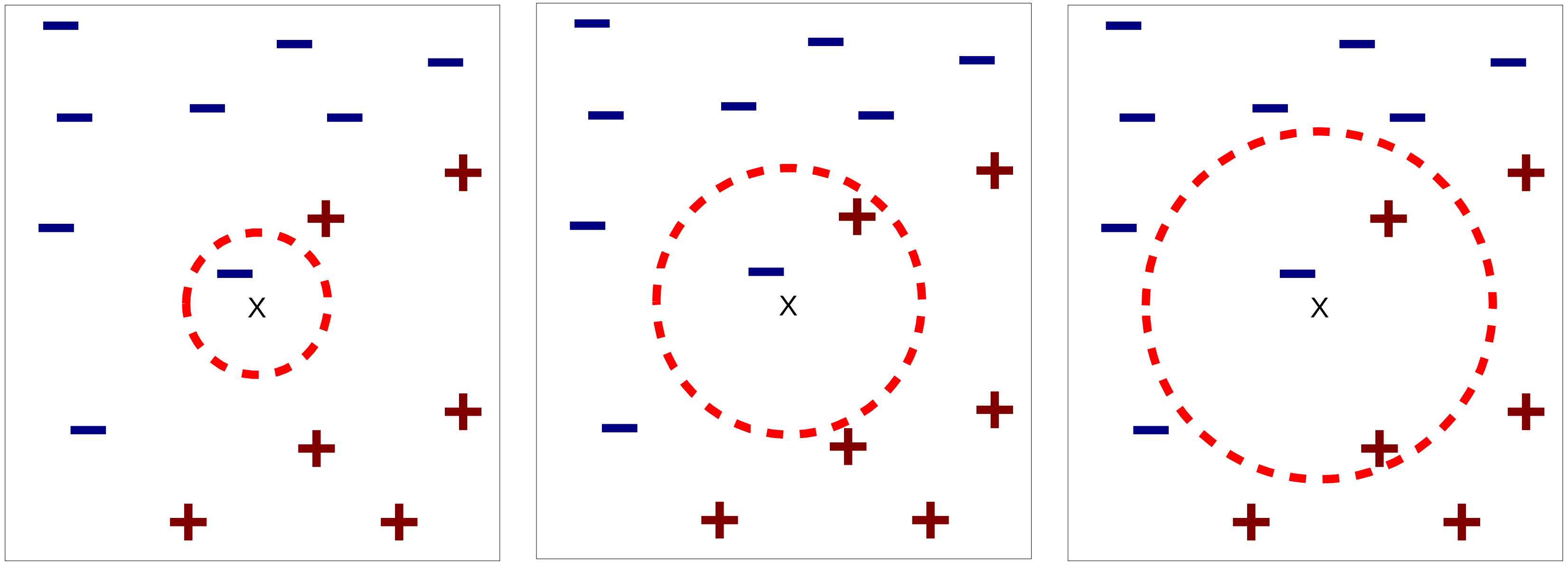
Nearest Neighbor Classification

- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, $w = 1/d^2$

Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

Nearest Neighbor Classification…

- Problem with Euclidean measure:
  - High dimensional data
    - curse of dimensionality　　“距离在高维空间失效”
  - Can produce counter-intuitive results

1 1 1 1 1 1 1 1 1 1 1 0	vs	1 0 0 0 0 0 0 0 0 0 0 0	
0 1 1 1 1 1 1 1 1 1 1 1		0 0 0 0 0 0 0 0 0 0 0 1	
d = 1.4142		d = 1.4142	

◆ Solution: Normalize the vectors to unit length

Nearest neighbor Classification…

- k-NN classifiers are lazy learners
  - It does not build models explicitly
  - Unlike eager learners such as decision trees
- Classifying unknown records is relatively expensive
  - Naïve algorithm: O(n)
  - Need for structures to retrieve nearest neighbors fast.
    - The Nearest Neighbor Search problem.
    - Also, Approximate Nearest Neighbor Search

KD-tree, very popular when dealing with geo locations

- Issues with distance in very high-dimensional spaces

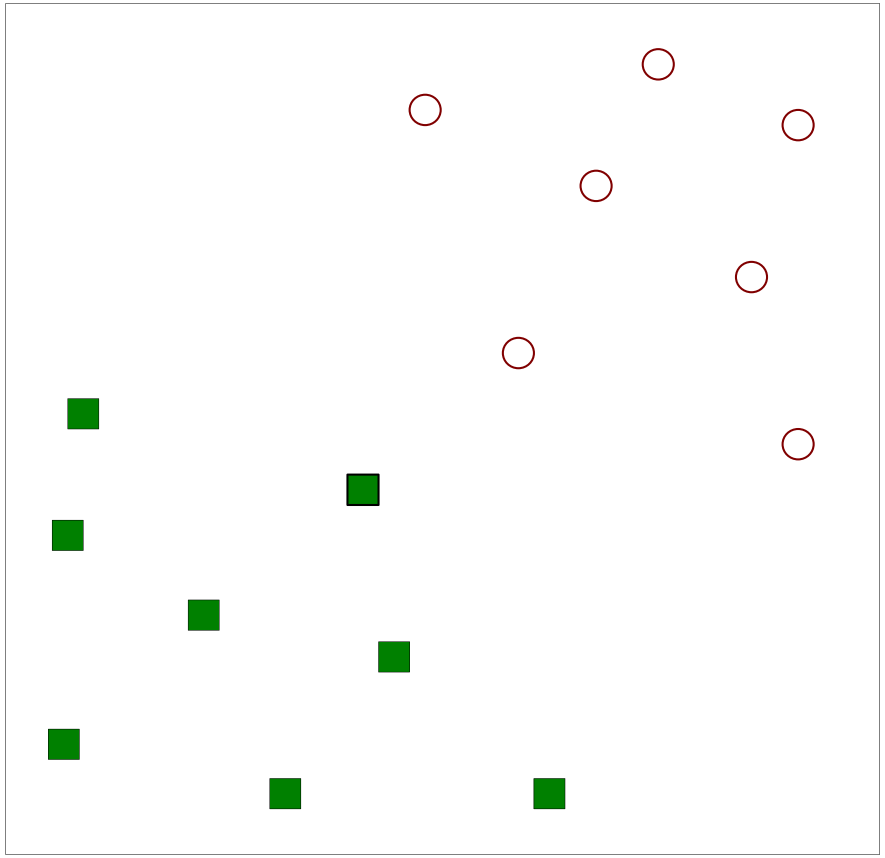
SUPPORT VECTOR MACHINES

Linear classifiers

- SVMs are part of a family of classifiers that assumes that the classes are linearly separable
- That is, there is a hyperplane that separates (approximately, or exactly) the instances of the two classes.
- The goal is to find this hyperplane

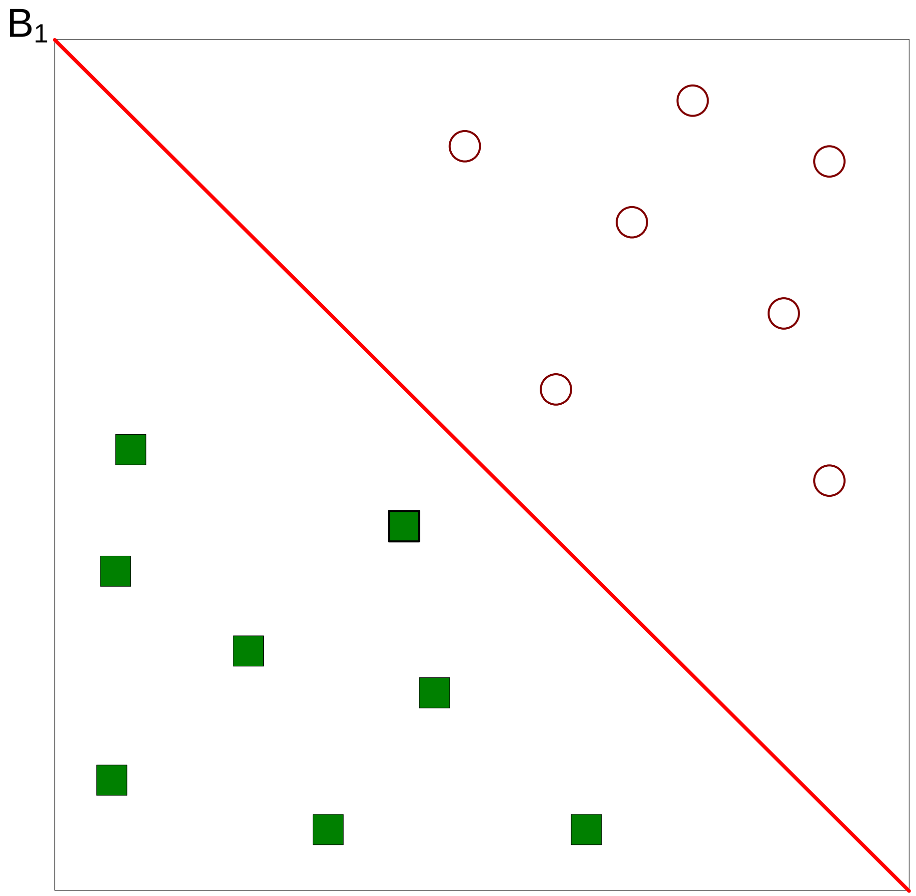
《导论》P158

Support Vector Machines



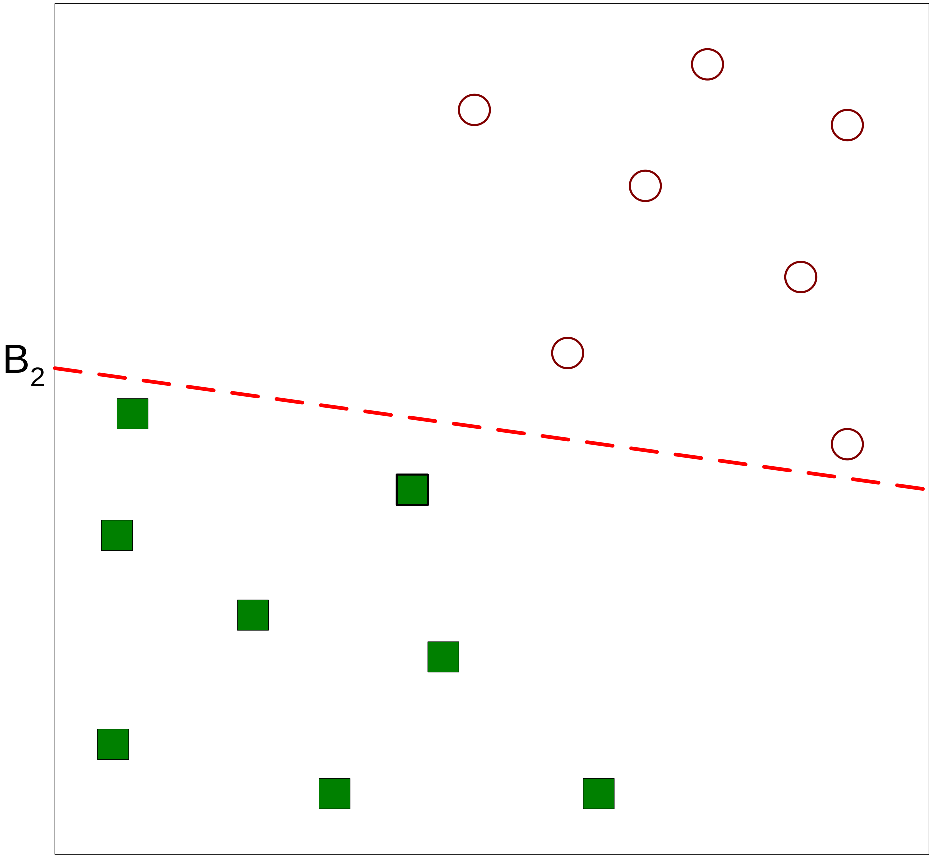
- Find a linear hyperplane (decision boundary) that will separate the data

Support Vector Machines



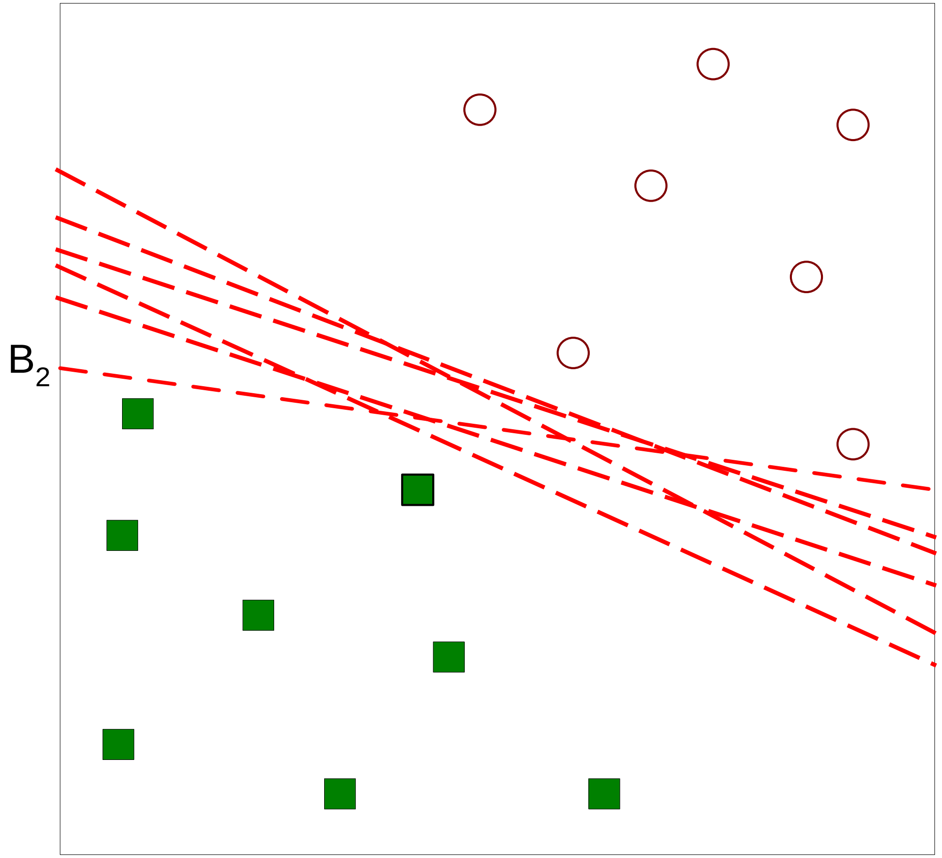
- One Possible Solution

Support Vector Machines



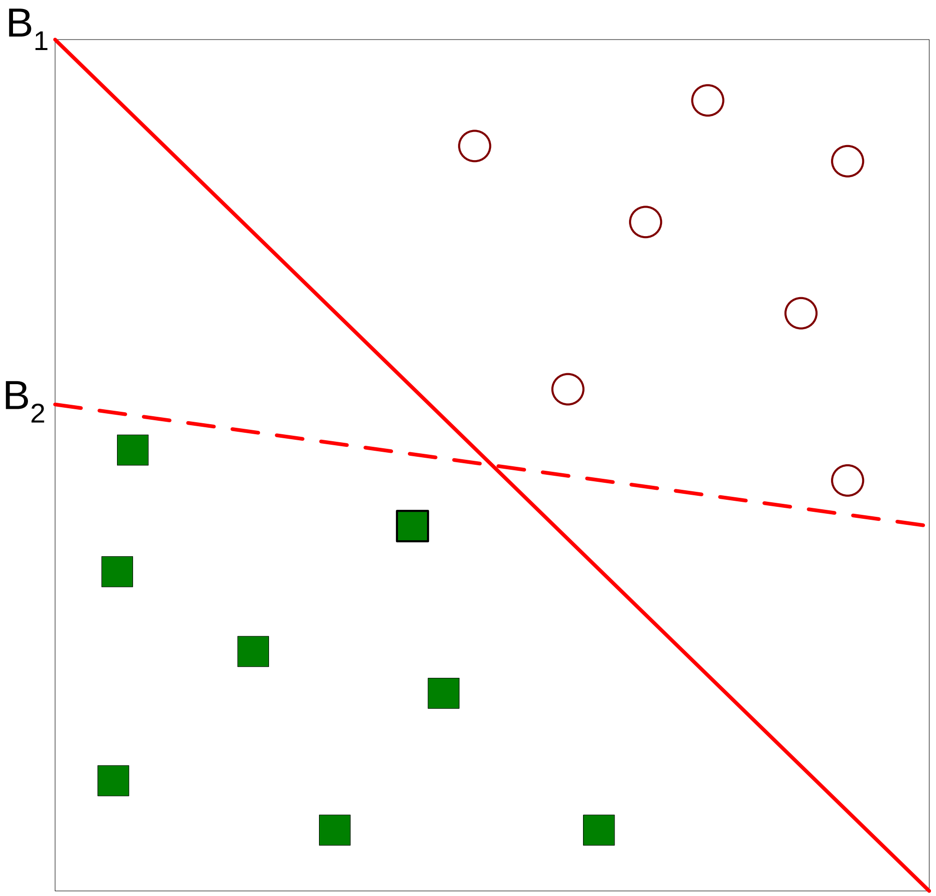
- Another possible solution

Support Vector Machines



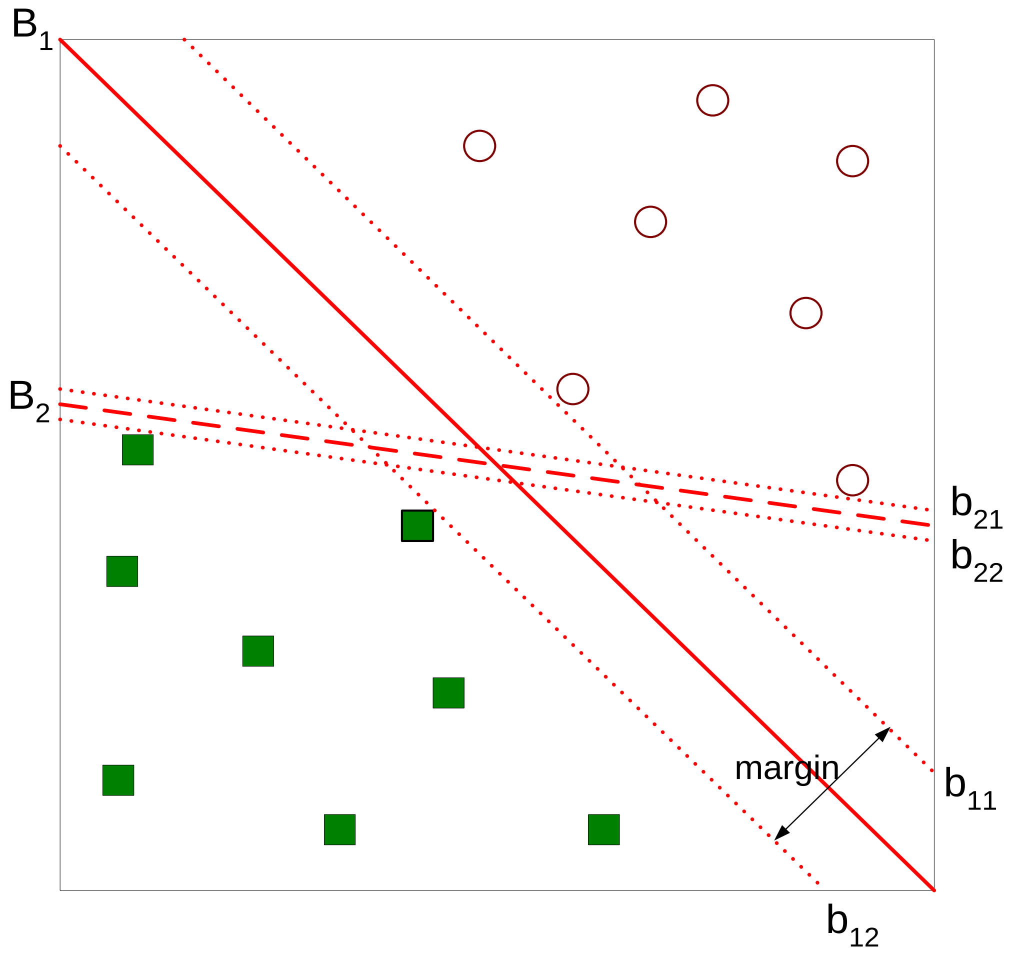
- Other possible solutions

Support Vector Machines



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

Support Vector Machines



《导论》P158

- Find hyperplane maximizes the margin : B1 is better than B2

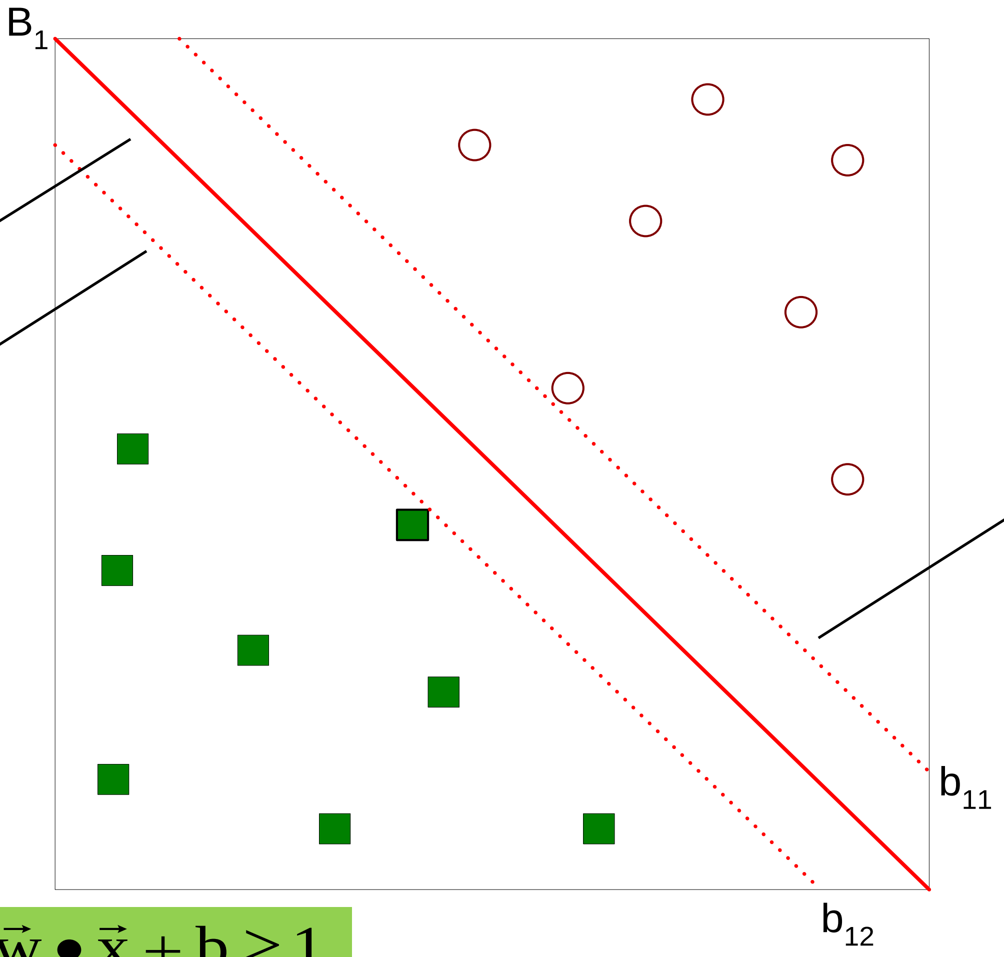
Support Vector Machines

Instance: (x_i, y_i)
x_i is the feature vector of an instance,
y_i is its label, $y_i \in \{-1,1\}$

（注：$\vec{w}$为列向量）

$$\vec{w} \bullet \vec{x} + b = 0$$

$$\vec{w} \bullet \vec{x} + b = -1$$



$$\vec{w} \bullet \vec{x} + b = +1$$

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

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

Support Vector Machines

- We want to maximize: $Margin = \frac{2}{\|\vec{w}\|}$
- Which is equivalent to minimizing: $L(\vec{w}) = \frac{\|\vec{w}\|}{2}$
- But subjected to the following constraints:

$$\vec{w} \cdot \vec{x_i} + b \geq 1 \text{ if } y_i = 1$$
$$\vec{w} \cdot \vec{x_i} + b \leq -1 \text{ if } y_i = -1$$

Concisely:
$y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1$

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

《导论》P160

NAÏVE BAYES CLASSIFIER

Bayes Classifier

- A probabilistic framework for solving classification problems
- A, C random variables
- Joint probability: P(A=a,C=c) 《导论》P124
- Conditional probability: P(C=c | A=a);
- $\mathrm{P}(C|A) = \mathrm{P}(C,A)/P(A)$ $\mathrm{P}(\mathrm{A}|C) = \mathrm{P}(C,A)/P(C)$
- Relationship between joint and conditional probability distributions

$$\mathrm{P}(C,A) = \mathrm{P}(C|A)\,P(A) = P(A|C)P(C)$$

- Bayes Theorem:

$$\underbrace{P(C|A)}_{\text{Posterior probability}} = \frac{P(A|C)P(C)}{P(A)}$$

Bayesian Classifiers

- How to classify the new record X = (‘Yes’, ‘Single’, 80K)

Tid	Refund	Marital Status	Taxable Income	Evade
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

Find the class with the highest probability given the vector values.

Maximum Posterior Probability estimate:

- Find the value c for class C that maximizes P(C=c| X)

- How do we estimate P(C|X) for the different values of C?
- We want to estimate
  - P(C=Yes| X)
  - P(C=No| X)

Bayesian Classifiers

- In order for probabilities to be well defined:
  - Consider each attribute and the class label as random variables
  - Probabilities are determined from the data

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
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

Evade C
Event space: {Yes, No}
$$P(C) = (0.3, 0.7)$$

Refund A_1
Event space: {Yes, No}
$$P(A_1) = (0.3, 0.7)$$

Martial Status A_2
Event space: {Single, Married, Divorced}
$$P(A_2) = (0.4, 0.4, 0.2)$$

Taxable Income A_3
Event space: R
$$P(A_3) \sim Normal(\mu, \sigma^2)$$
$\mu = 104$:sample mean, $\sigma^2 = 1874$:sample variance

Bayesian Classifiers

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

$$P(C|A_1, A_2, \ldots, A_n) = \frac{P(A_1, A_2, \ldots, A_n|C)P(C)}{P(A_1, A_2, \ldots, A_n)}$$

  - Maximizing

$$P(C \mid A_1, A_2, \ldots, A_n)$$

  is equivalent to maximizing

$$P(A_1, A_2, \ldots, A_n|C)\, P(C)$$

    - The value $P(A_1, \ldots, A_n)$ is the same for all values of C.

- How do we estimate $P(A_1, A_2, \ldots, A_n|C\,)$?

Naïve Bayes Classifier

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

  - We can estimate $P(A_i |\, C)$ from the data.

  - New point $X = (A_1 = \alpha_1, \dots, A_n = \alpha_n)$ is classified to class c if

$$P(C = c|X) = P(C = c) \prod_i P(A_i = \alpha_i|c)$$

is maximum over all possible values of C.

Example

- Record

  X = (Refund = Yes, Status = Single, Income =80K)

- For the class C :'Evade', we want to compute:

  P(C = Yes|X) and P(C = No| X)

- We compute:
  - P(C = Yes|X) = P(C = Yes)*P(Refund = Yes |C = Yes)
    *P(Status = Single |C = Yes)
    *P(Income =80K |C= Yes)
  - P(C = No|X) = P(C = No)*P(Refund = Yes |C = No)
    *P(Status = Single |C = No)
    *P(Income =80K |C= No)

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
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 Prior Probability:

$$P(C = c) = \frac{N_c}{N}$$

N_c: Number of records with class c

N = Number of records

P(C = No) = 7/10

P(C = Yes) = 3/10

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

Discrete attributes:

$$P(A_i = a|C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
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

Discrete attributes:

$$P(A_i = a|C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Refund = Yes|No) = 3/7

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
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**

Discrete attributes:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Refund = Yes|Yes) = 0

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

Discrete attributes:

$$P(A_i = a|C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Status=Single|No) = 2/7

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
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**

Discrete attributes:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Status=Single|Yes) = 2/3

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
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

Numerical Attributes:

- Assume a normal distribution for each (A_i, c_j) pair

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

- For Class=Yes and attribute Income
  - sample mean $\mu = 90$
  - sample variance $\sigma^2 = 25$
- For Income = 80

$$P(Income = 80 \mid Yes) = \frac{1}{\sqrt{2\pi}(5)} e^{-\frac{(80-90)^2}{2(25)}} = 0.01$$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

Numerical Attributes:

- Assume a normal distribution for each (A_i, c_j) pair

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

- For Class=No and attribute Income
  - sample mean $\mu = 110$
  - sample variance $\sigma^2 = 2975$
- For Income = 80

$$P(Income = 80 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(80-110)^2}{2(2975)}} = 0.0062$$

Example

- Record

  X = (Refund = Yes, Status = Single, Income =80K)

- We compute:
  - P(C = Yes|X) = P(C = Yes)*P(Refund = Yes |C = Yes)

    *P(Status = Single |C = Yes)

    *P(Income =80K |C= Yes)

    = 3/10* 0 * 2/3 * 0.01 = 0
  - P(C = No|X) = P(C = No)*P(Refund = Yes |C = No)

    *P(Status = Single |C = No)

    *P(Income =80K |C= No)

    = 7/10 * 3/7 * 2/7 * 0.0062 = 0.0005

Example of Naïve Bayes Classifier

- Creating a Naïve Bayes Classifier, essentially means to compute counts:

Total number of records: N = 10

Class No:
Number of records: 7
Attribute Refund:
 Yes: 3
 No: 4
Attribute Marital Status:
 Single: 2
 Divorced: 1
 Married: 4
Attribute Income:
 mean: 110
 variance: 2975

Class Yes:
Number of records: 3
Attribute Refund:
 Yes: 0
 No: 3
Attribute Marital Status:
 Single: 2
 Divorced: 1
 Married: 0
Attribute Income:
 mean: 90
 variance: 25

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

Example of Naïve Bayes Classifier

Given a Test Record:

X = (Refund = Yes, Status = Single, Income =80K)

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|Class=No) = P(Refund=Yes|Class=No)
$\times$ P(Married| Class=No)
$\times$ P(Income=120K| Class=No)
= 3/7 * 2/7 * 0.0062 = 0.00075

P(X|Class=Yes) = P(Refund=No| Class=Yes)
$\times$ P(Married| Class=Yes)
$\times$ P(Income=120K| Class=Yes)
= 0 * 2/3 * 0.01 = 0

- P(No) = 0.3, P(Yes) = 0.7

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

- Laplace Smoothing:

$$P(A_i = a | C = c) = \frac{N_{ac} + 1}{N_c + N_i}$$

  - N_i: number of attribute values for attribute A_i

Example of Naïve Bayes Classifier

- Creating a Naïve Bayes Classifier, essentially means to compute counts:

Total number of records: N = 10

Class No:	Class Yes:
Number of records: 7	Number of records: 3
Attribute Refund:	Attribute Refund:
Yes: 3	Yes: 0
No: 4	No: 3
Attribute Marital Status:	Attribute Marital Status:
Single: 2	Single: 2
Divorced: 1	Divorced: 1
Married: 4	Married: 0
Attribute Income:	Attribute Income:
mean: 110	mean: 90
variance: 2975	variance: 25

Example of Naïve Bayes Classifier

- Creating a Naïve Bayes Classifier, essentially means to compute counts:

With Laplace Smoothing

Total number of records: N = 10

Class No:
Number of records: 7
Attribute Refund:
Yes: 3 +1
No: 4 +1
Attribute Marital Status:
Single: 2 +1
Divorced: 1 +1
Married: 4 +1
Attribute Income:
mean: 110
variance: 2975

Class Yes:
Number of records: 3
Attribute Refund:
Yes: 0 +1
No: 3 +1
Attribute Marital Status:
Single: 2 +1
Divorced: 1 +1
Married: 0 +1
Attribute Income:
mean: 90
variance: 25

naive Bayes Classifier:

P(Refund=Yes|No) = 4/9
P(Refund=No|No) = 5/9
P(Refund=Yes|Yes) = 1/5
P(Refund=No|Yes) = 4/5

P(Marital Status=Single|No) = 3/10
P(Marital Status=Divorced|No)=2/10
P(Marital Status=Married|No) = 5/10
P(Marital Status=Single|Yes) = 3/6
P(Marital Status=Divorced|Yes)=2/6
P(Marital Status=Married|Yes) = 1/6

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

Example of Naïve Bayes Classifier

Given a Test Record:

With Laplace Smoothing

X = (Refund = Yes, Status = Single, Income =80K)

naive Bayes Classifier:

P(Refund=Yes|No) = 4/9
P(Refund=No|No) = 5/9
P(Refund=Yes|Yes) = 1/5
P(Refund=No|Yes) = 4/5

P(Marital Status=Single|No) = 3/10
P(Marital Status=Divorced|No)=2/10
P(Marital Status=Married|No) = 5/10
P(Marital Status=Single|Yes) = 3/6
P(Marital Status=Divorced|Yes)=2/6
P(Marital Status=Married|Yes) = 1/6

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

P(X|Class=No) = P(Refund=No|Class=No)
× P(Married| Class=No)
× P(Income=120K| Class=No)
= 4/9 × 3/10 × 0.0062 = 0.00082

P(X|Class=Yes) = P(Refund=No| Class=Yes)
× P(Married| Class=Yes)
× P(Income=120K| Class=Yes)
= 1/5 × 3/6 × 0.01 = 0.001

- P(No) = 0.7, P(Yes) = 0.3
- P(X|No)P(No) = 0.0005
- P(X|Yes)P(Yes) = 0.0003

=> Class = No

Implementation details

- Computing the conditional probabilities involves multiplication of many very small numbers
  - Numbers get very close to zero, and there is a danger of numeric instability
- We can deal with this by computing the logarithm of the conditional probability

$$\log P(C|A) \sim \log P(A|C) + \log P(C)$$
$$= \sum_i \log P(A_i|C) + \log P(C)$$

Naïve Bayes for Text Classification

- Naïve Bayes is commonly used for text classification
- For a document with k terms $d = (t_1, \dots, t_k)$, its probability of being in class c:

$$P(c|d) = P(c)P(d|c) = P(c)\prod_{t_i \in d} P(t_i|c)$$

Fraction of documents in c

- $P(t_i|c)$= Fraction of terms from all documents in c that are t_i.

Number of times t_i appears in all documents in c

$$P(t_i|c) = \frac{N_{ic} + 1}{N_c + T}$$

Laplace Smoothing

Number of unique words (vocabulary size)

Total number of terms in all documents in c

- Easy to implement and works relatively well
- Limitation: Hard to incorporate additional features (beyond words).
  - E.g., number of adjectives used.

Example

News titles for Politics and Sports

	Politics	Sports
documents	"Obama meets Merkel" "Obama elected again" "Merkel visits Greece again"	"OSFP European basketball champion" "Miami NBA basketball champion" "Greece basketball coach?"
	P(p) = 0.5	P(s) = 0.5
terms Vocabulary size: 14	obama:2, meets:1, merkel:2, elected:1, again:2, visits:1, greece:1	OSFP:1, european:1, basketball:3, champion:2, miami:1, nba:1, greece:1, coach:1
	Total terms: 10	Total terms: 11

New title: X = "Obama likes basketball"

P(Politics|X) ~ P(p)*P(obama|p)*P(likes|p)*P(basketball|p)
= 0.5 * 3/(10+14) *1/(10+14) * 1/(10+14) = 0.000108

P(Sports|X) ~ P(s)*P(obama|s)*P(likes|s)*P(basketball|s)
= 0.5 * 1/(11+14) *1/(11+14) * 4/(11+14) = 0.000128