KNN

SYRACUSE UNIVERSITY
School of Information Studies

REVIEW DISTANCE MEASURE

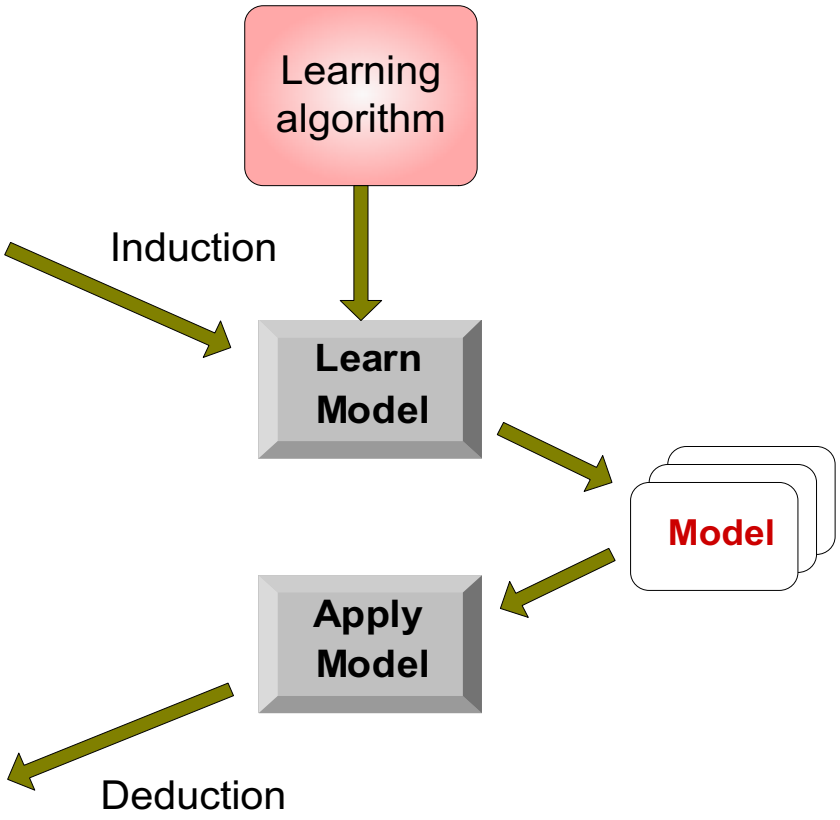
CLASSIFICATION PROCESS

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

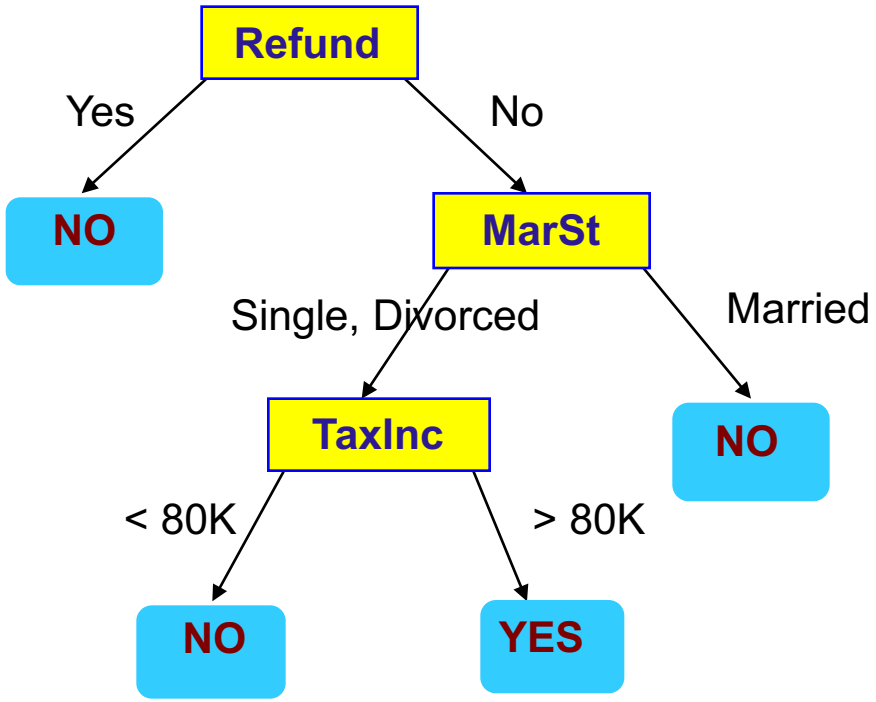
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



MACHINE LEARNING ALGORITHMS

Algorithms like decision tree and naive Bayes will construct a learning model from training examples and then apply the model for prediction on new test examples.



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) = 2/3
P(Marital Status=Divorced|Yes)=1/3
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

INSTANCE-BASED LEARNING

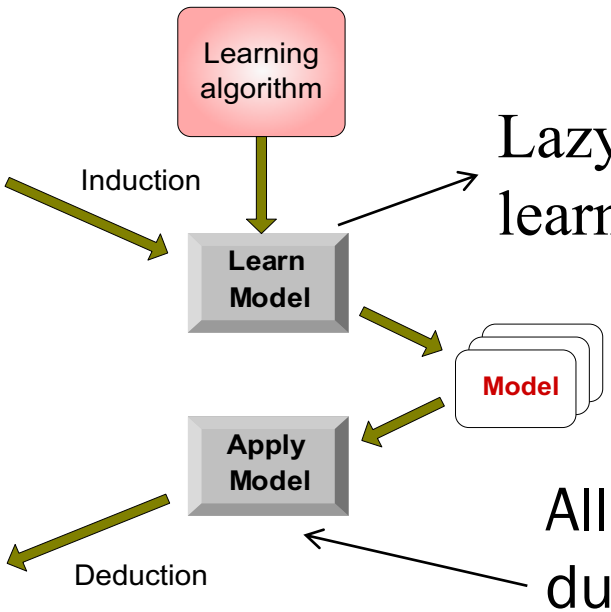
In contrast, instance-based learning methods simply store the training examples without doing any calculations during training process, and classification and prediction are delayed until new examples are given.

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Lazy learner

All calculations happen during prediction.

K-NEAREST NEIGHBOR (K-NN)

Training process:

Read in all training examples.

Classification process:

Given a test example, compare the similarity between the test example and all training examples. Choose the majority-voted category label in the k-nearest training examples.

NEAREST NEIGHBOR CLASSIFICATION

Choosing the value of k:

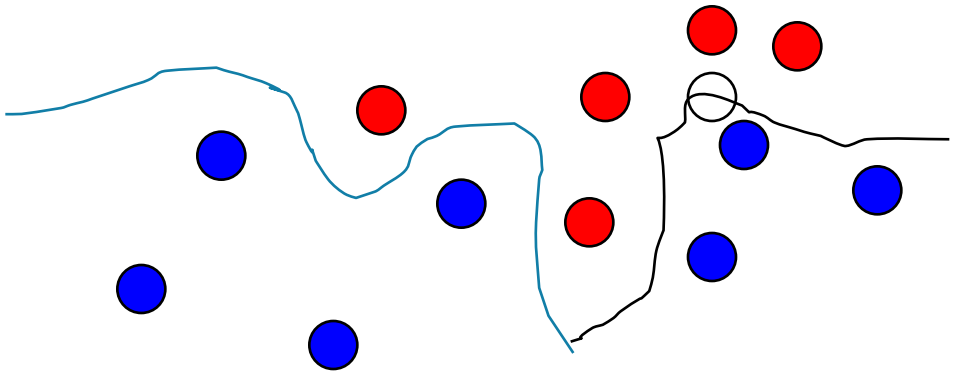
- If k is too small, sensitive to noise points.
- If k is too large, neighborhood may include points from other classes.

ADVANTAGES OF K-NN

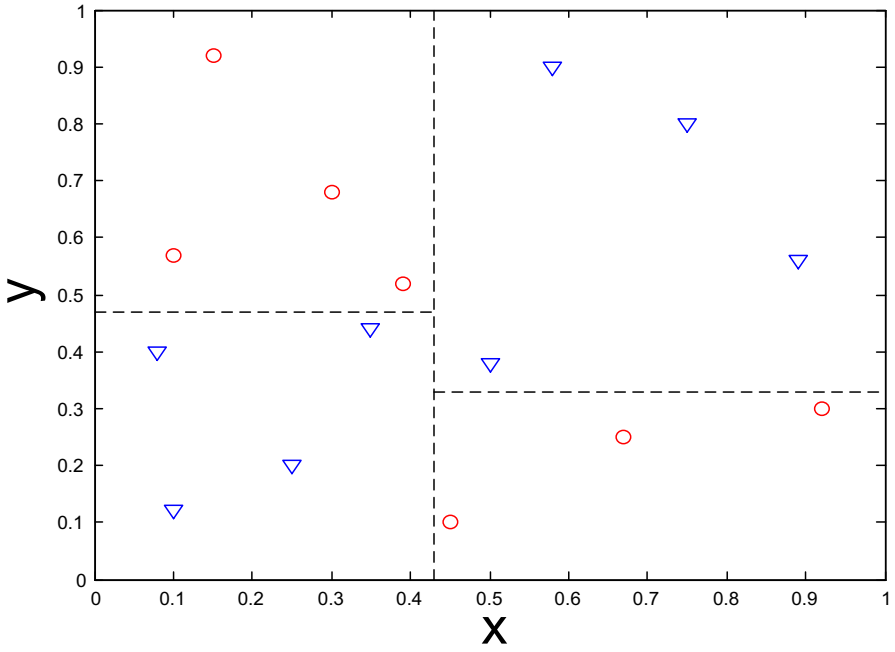
No assumptions made

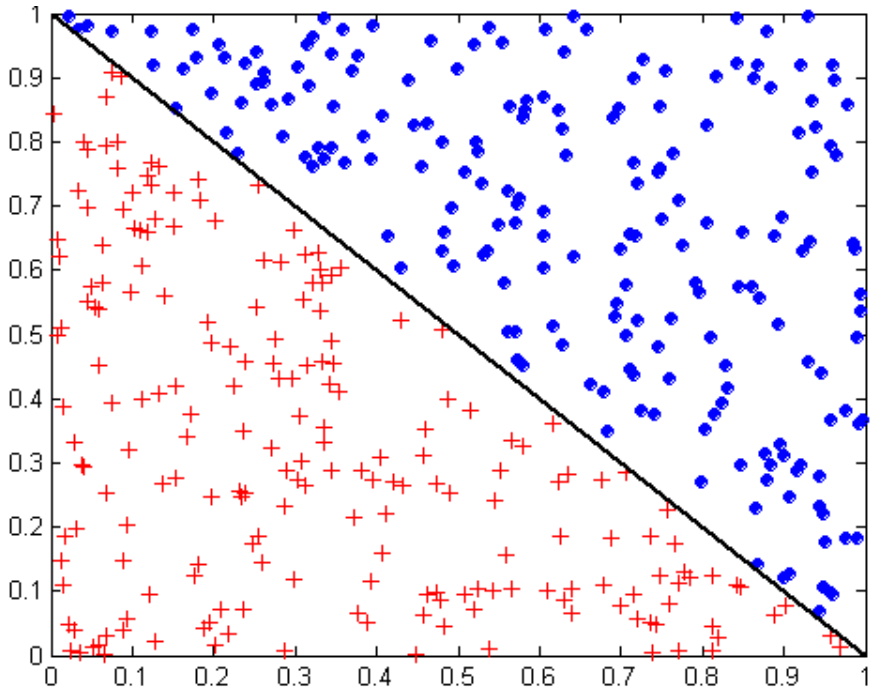
Remember the independence assumption in naive Bayes.

Works well when the decision function to be learned is very complex

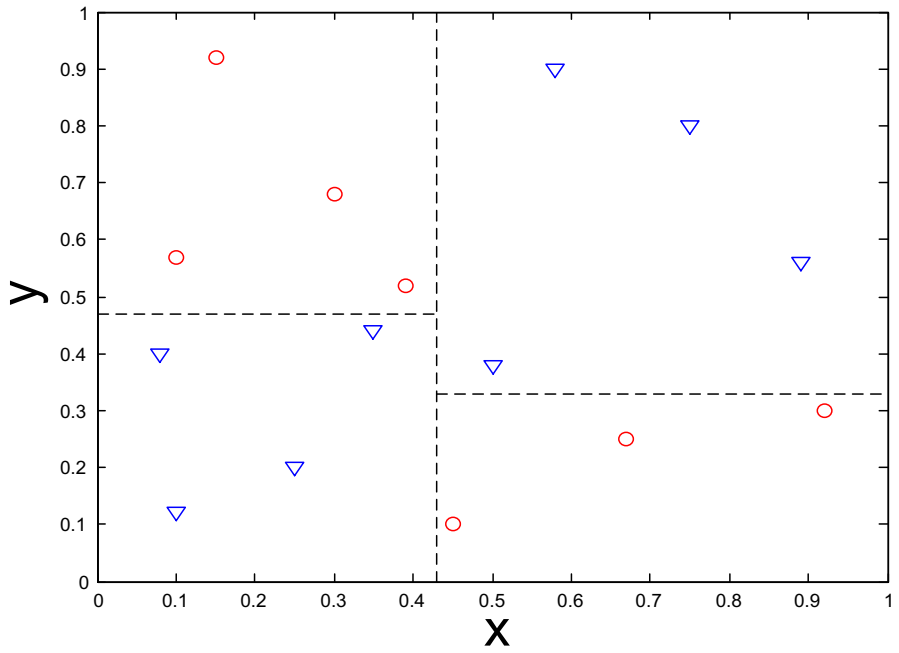


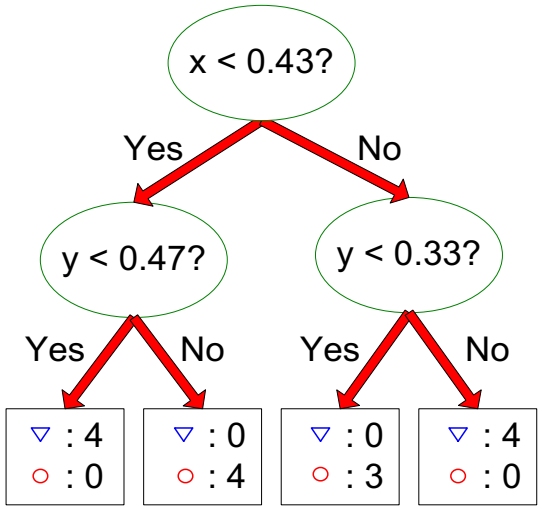
THE SHAPE OF DECISION BOUNDARY MATTERS





DECISION BOUNDARY OF DECISION TREE MODELS





DECISION BOUNDARY OF LINEAR MODELS

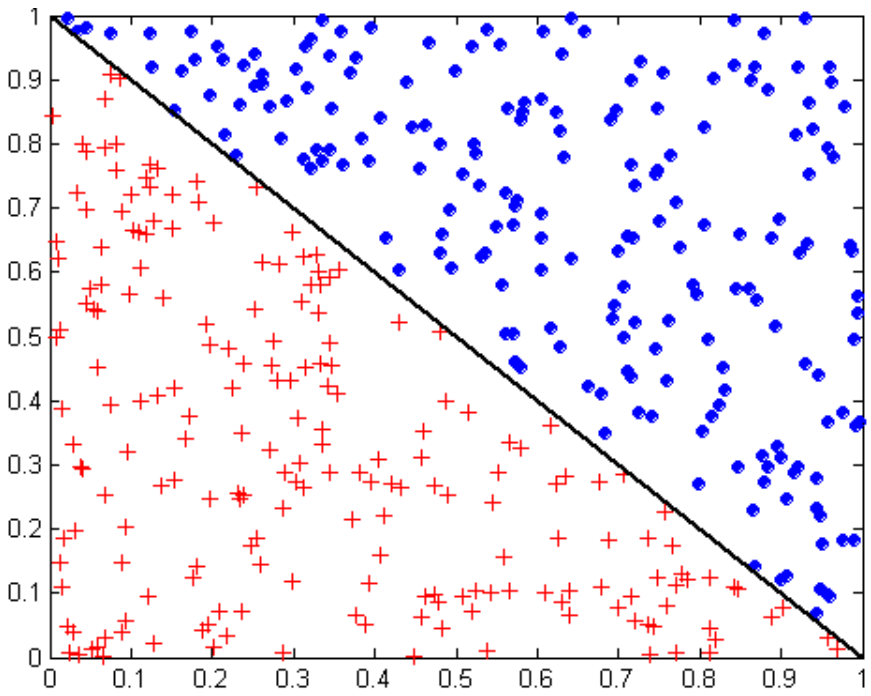
Linear: Naive Bayes, SVM

How many parameters to determine a line in 2D space?

Y = ax + b

Weight

Intercept



WHY IS NAIVE BAYES A LINEAR CLASSIFIER?

The decision function can be rewritten to a linear function.

Original decision function:

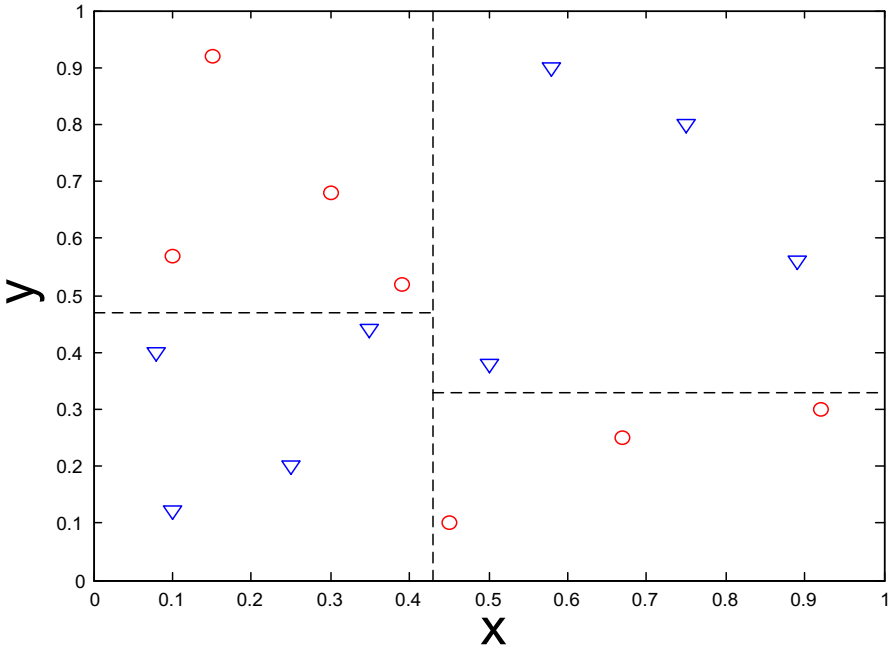
Prob(Ci)*Prob(T1|Ci)*Prob(T2|Ci) * … *Prob(Tm|Ci))

Apply log transformation:

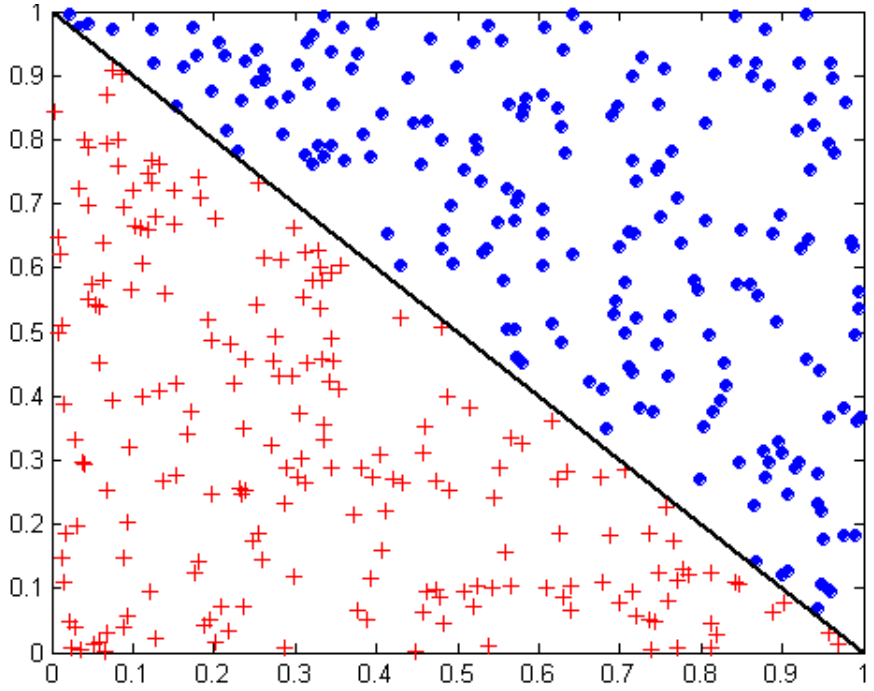
log(Prob(Ci)) + log(Prob(T1|Ci)) + log(Prob(T2|Ci)) + … + log(Prob(Tm|Ci))

http://cs.nyu.edu/faculty/davise/ai/bayesText.html

THE SHAPE OF DECISION BOUNDARY MATTERS



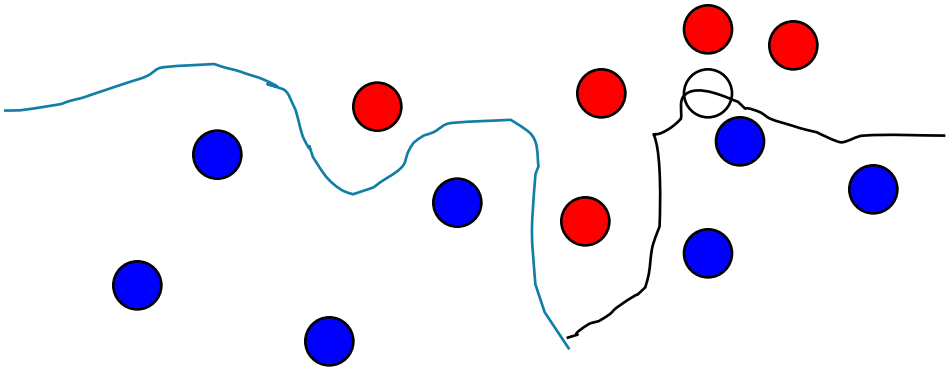
Decision tree model fits;
not the linear model



Linear model fits;
not decision tree model

ADVANTAGE OF K-NN

The decision boundary has no predefined shape.



DISADVANTAGES OF K-NN

Sensitive to noisy training data

All attributes participate in classification.

If only a few relevant attributes are relevant to prediction, the participation of those irrelevant attributes would harm the prediction performance.

DISADVANTAGES OF K-NN

High computational cost

Precomputed models can be quickly applied to test data.

Since there is no training step, nearly all computation takes place in the prediction step.