Instance-based Classifiers

AW

Lecture Overview

1. Nearest Neighbor Classifier

2. Pebbls

Instance-Based Classifiers

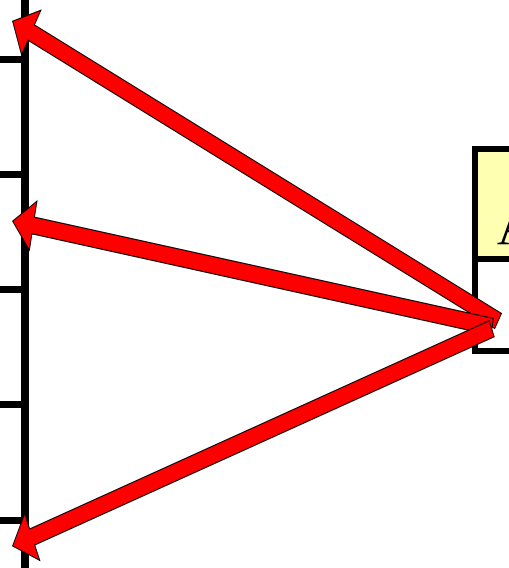
- Store the training records
- Use training records to predict the class label of unseen cases

Set of Stored Cases

Atr1	.....	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

Unseen Case

Atr1	.....	AtrN

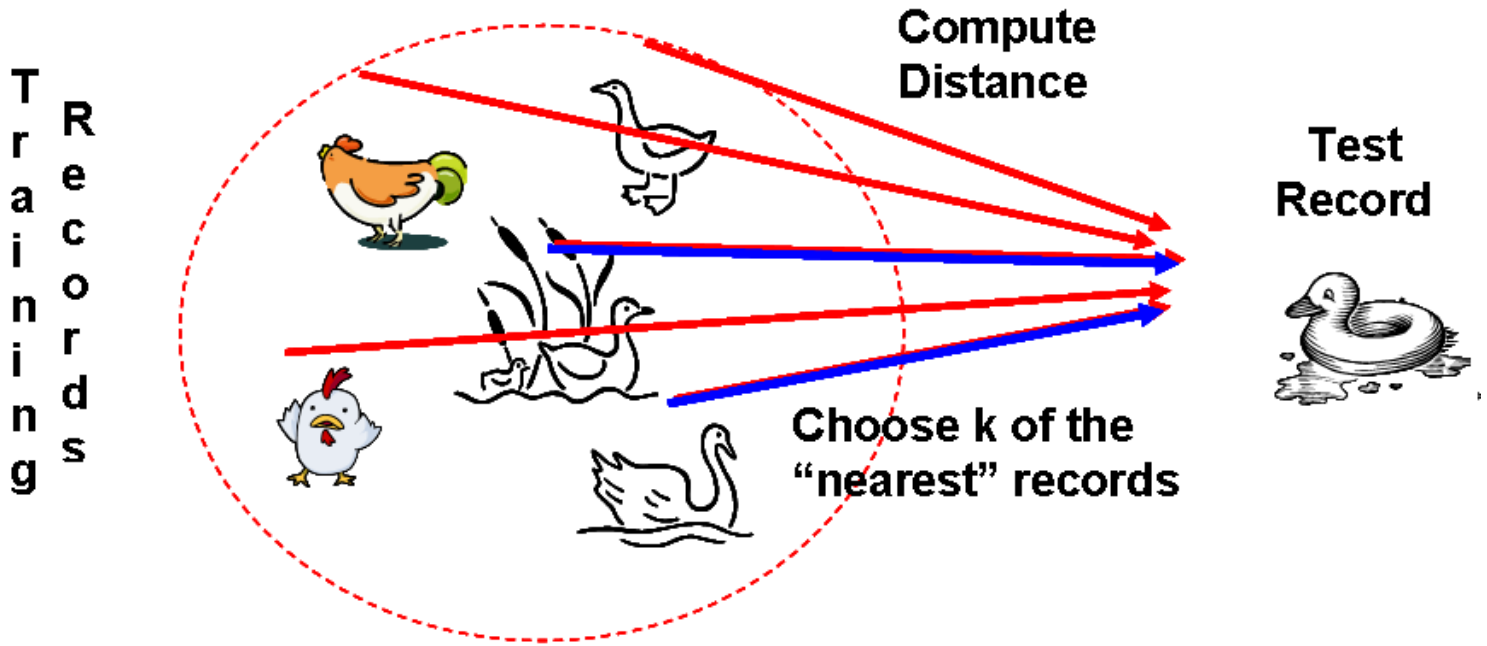


Instance Based Classifiers

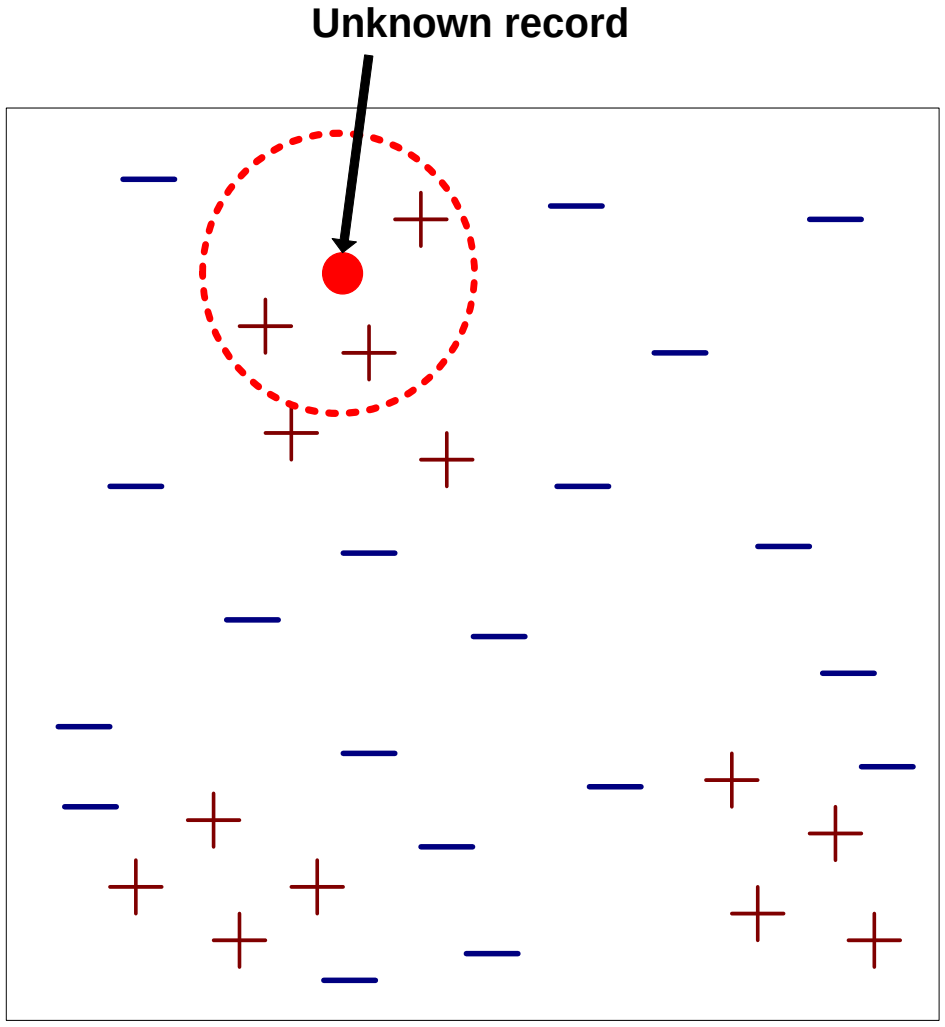
- Examples:
  - RoTe-learner
    - Memorizes entire training data and performs classification only if attributes of record match one of the training examples exactly
  - Nearest neighbor
    - Uses k "closest" points (nearest neighbors) for performing classification

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



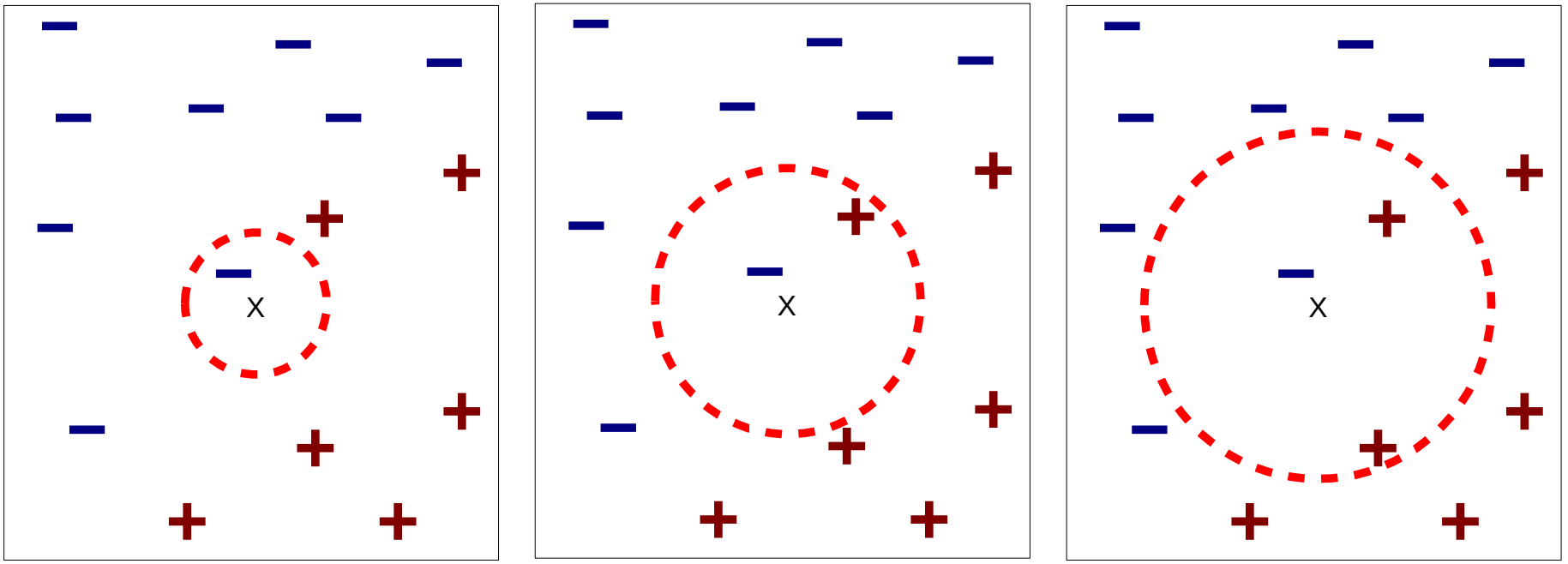
Nearest-Neighbor Classifiers



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

Definition of Nearest Neighbor

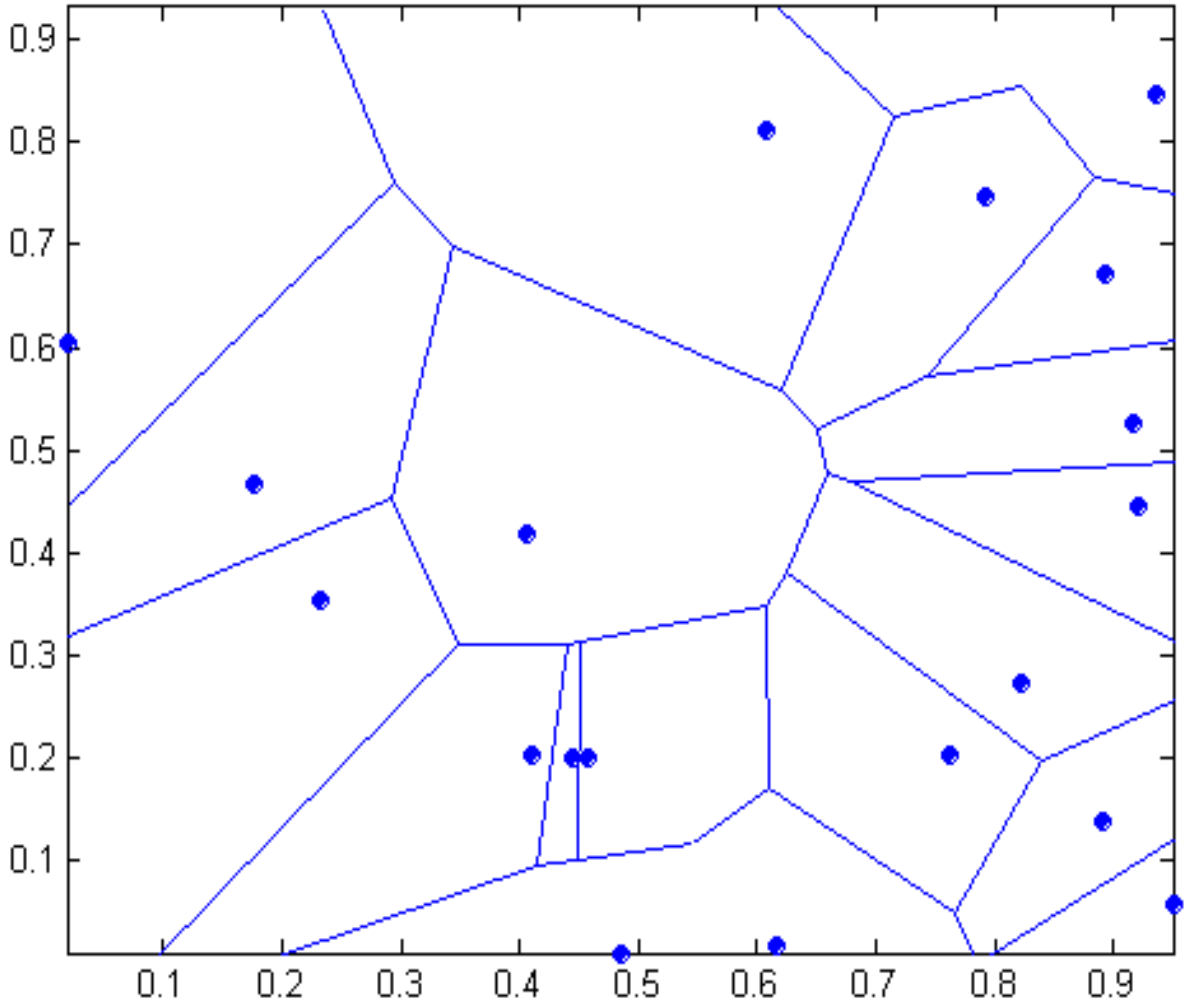


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

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

1 nearest-neighbor

Voronoi Diagram

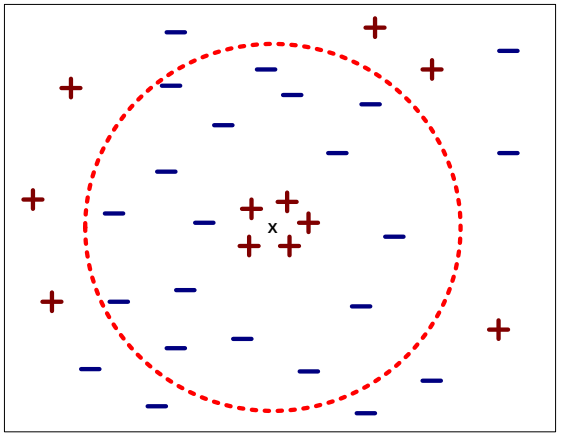


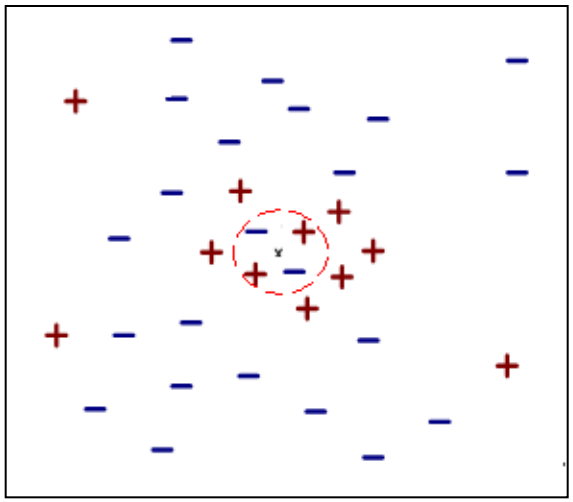
Nearest Neighbor Classification

- 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, $w = 1/d^2$

Nearest Neighbor Classification...

- Choosing the value of k:
  - If k is too small - high variance ⇒ sensitive to noise points
  - If k is too large - biased data (neighborhood may include points from other classes) ⇒ biased misclassification





Nearest Neighbor Classification...

- Scaling issues
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from \$10K to \$1M

Nearest Neighbor Classification...

- Problem with Euclidean measure:
  - High dimensional data
    - curse of dimensionality
  - Can produce counter-intuitive results

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
	vs		
0 1 1 1 1 1 1 1 1 1 1 1 1 1 0		0 0 0 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 tree induction and rule-based systems
  - Classifying unknown records are relatively expensive

Lecture Overview

1. Nearest Neighbor Classifier

2. Pebbls

Example: PEBLS

- PEBLS: Parallel Examplar-Based Learning System (Cost & Salzberg)
  - Works with both continuous and nominal features
    - For nominal features, distance between two nominal values is computed using modified value difference metric (MVDM)
  - Each record is assigned a weight factor
  - Number of nearest neighbor, k = 1

$$d(V_1, V_2) = \sum_i \left| \frac{n_{1i}}{n_1} - \frac{n_{2i}}{n_2} \right|$$

Example: PEBLS

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

$$d(V_1, V_2) = \sum_i \left| \frac{n_{1i}}{n_1} - \frac{n_{2i}}{n_2} \right|$$

Distance between nominal attribute values:

d(Single,Married)

$= |\ 2/4 - 0/4\ | + |\ 2/4 - 4/4\ | = 1$

d(Single,Divorced)

$= |\ 2/4 - 1/2\ | + |\ 2/4 - 1/2\ | = 0$

d(Married,Divorced)

$= |\ 0/4 - 1/2\ | + |\ 4/4 - 1/2\ | = 1$

d(Refund=Yes,Refund=No)

$= |\ 0/3 - 3/7\ | + |\ 3/3 - 4/7\ | = 6/7$

Class	Marital Status		
	Single	Married	Divorced
Yes	2	0	1
No	2	4	1

Class	Refund	
	Yes	No
Yes	0	3
No	3	4

Example: PEBLS

Tid	Refund	Marital Status	Taxable Income	Cheat
X	Yes	Single	125K	**No**
Y	No	Married	100K	**No**

Distance between record X and record Y:

$$\Delta(X,Y) = w_X w_Y \sum_{i=1}^{d} d(X_i, Y_i)^2$$

where: $w_X = \dfrac{\text{Number of times X is used for prediction}}{\text{Number of times X predicts correctly}}$

$w_X \cong 1$ if X makes accurate prediction most of the time

$w_X > 1$ if X is not reliable for making predictions

Nearest Neighbor classifiers in R

```
library(FNN);data(iris3)
x<-sample(1:50,25,F)
train <- rbind(iris3[x,,1], iris3[x,,2], iris3[x,,3])
test <- rbind(iris3[-x,,1], iris3[-x,,2], iris3[-x,,3])
cl <- factor(c(rep("s",25), rep("c",25), rep("v",25)))
NN<-knn(train, test, cl, k = 3, prob=TRUE)
summary(NN)
table(NN,cl)
err<-(nrow(test)-sum(diag(table(NN,cl))))/nrow(test);err
NN
```

Reading

TSKK Section 4.3