Memory-based reasoning: nearest neighbors

Lecture 05

by Marina Barsky

Classification example: bankruptcy dataset

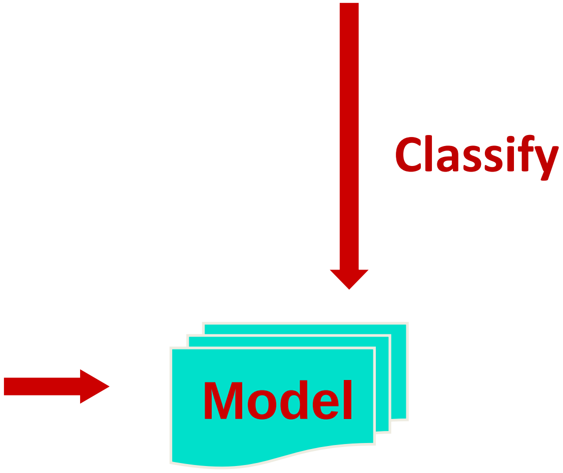
Training set

Late payments, L	Spending ratio, R	Bankruptcy
3	0.2	No
1	0.3	No
4	0.5	No
2	0.7	No
0	1.0	No
1	1.2	No
1	1.7	No
6	0.2	Yes
7	0.3	Yes
6	0.7	Yes
3	1.1	Yes
2	1.5	Yes
4	1.7	Yes
2	1.9	Yes

Class labels

New customer

L	R	B
2	0.3	?



Classify

Model

L: #late payments / year
R: expenses / income ratio

Memory-based reasoning

Seems poisonous

Amanita muscaria

Classification by similarity

“ If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck. ”

New classifier: Nearest Neighbor

- Remember the entire labeled training set
- When a new sample comes:
  - Find the most similar sample in the labeled collection (**the nearest neighbor**)
  - Return the class label associated with it

Classification: eager classifier

(for example logistic regression or decision tree)

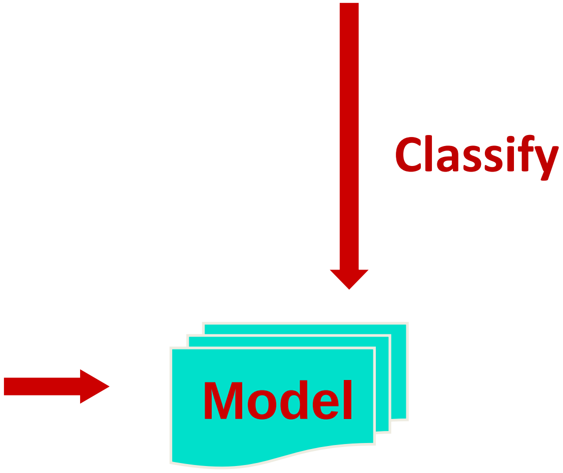
Training set

Late payments, L	Spending ratio, R	Bankruptcy
3	0.2	No
1	0.3	No
4	0.5	No
2	0.7	No
0	1.0	No
1	1.2	No
1	1.7	No
6	0.2	Yes
7	0.3	Yes
6	0.7	Yes
3	1.1	Yes
2	1.5	Yes
4	1.7	Yes
2	1.9	Yes

Class labels

New customer

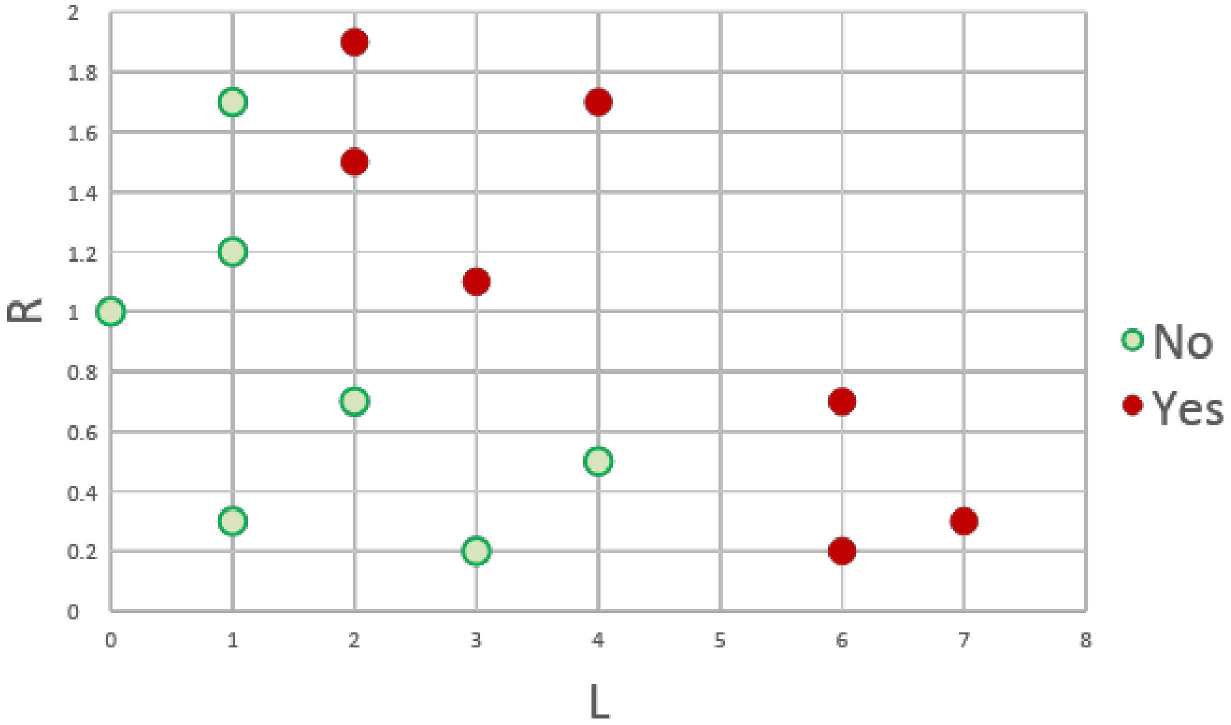
L	R	B
2	0.3	?



L: #late payments / year
R: expenses / income ratio

Different approach: lazy classifier

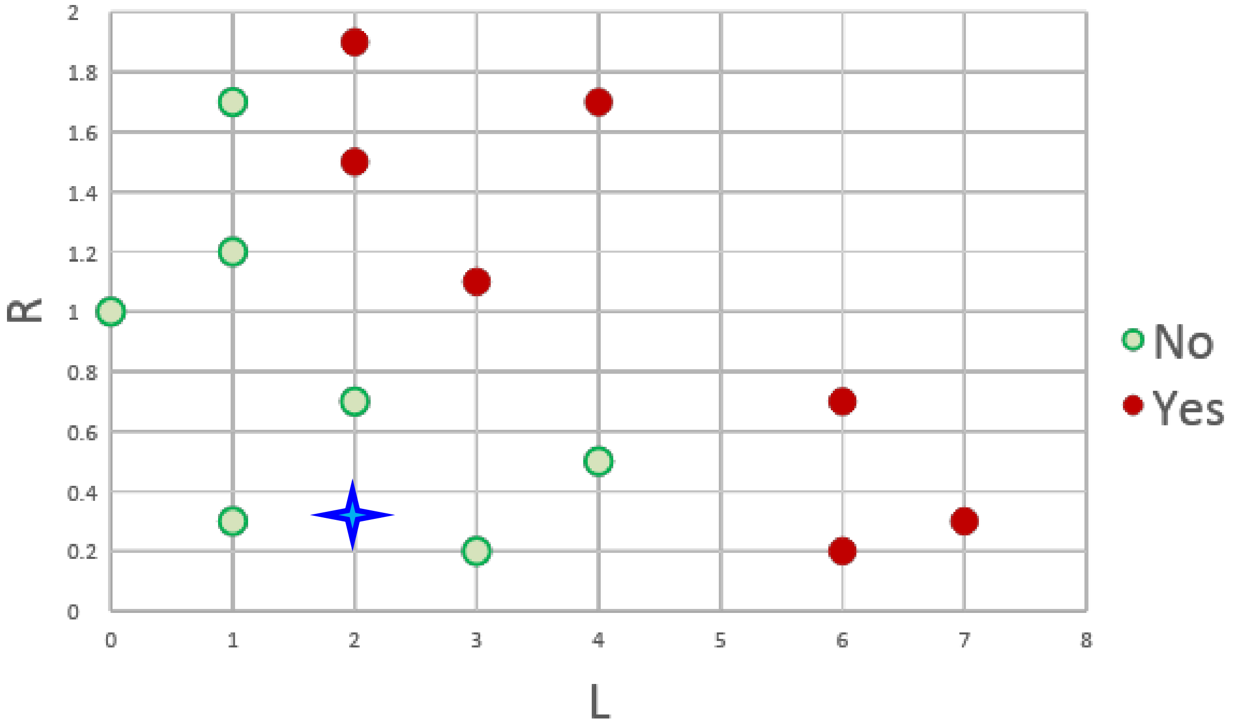
L	R	B
3	0.2	No
1	0.3	No
4	0.5	No
2	0.7	No
0	1	No
1	1.2	No
1	1.7	No
6	0.2	Yes
7	0.3	Yes
6	0.7	Yes
3	1.1	Yes
2	1.5	Yes
4	1.7	Yes
2	1.9	Yes



L: #late payments / year
R: expenses / income ratio

Predicting bankruptcy: nearest neighbor

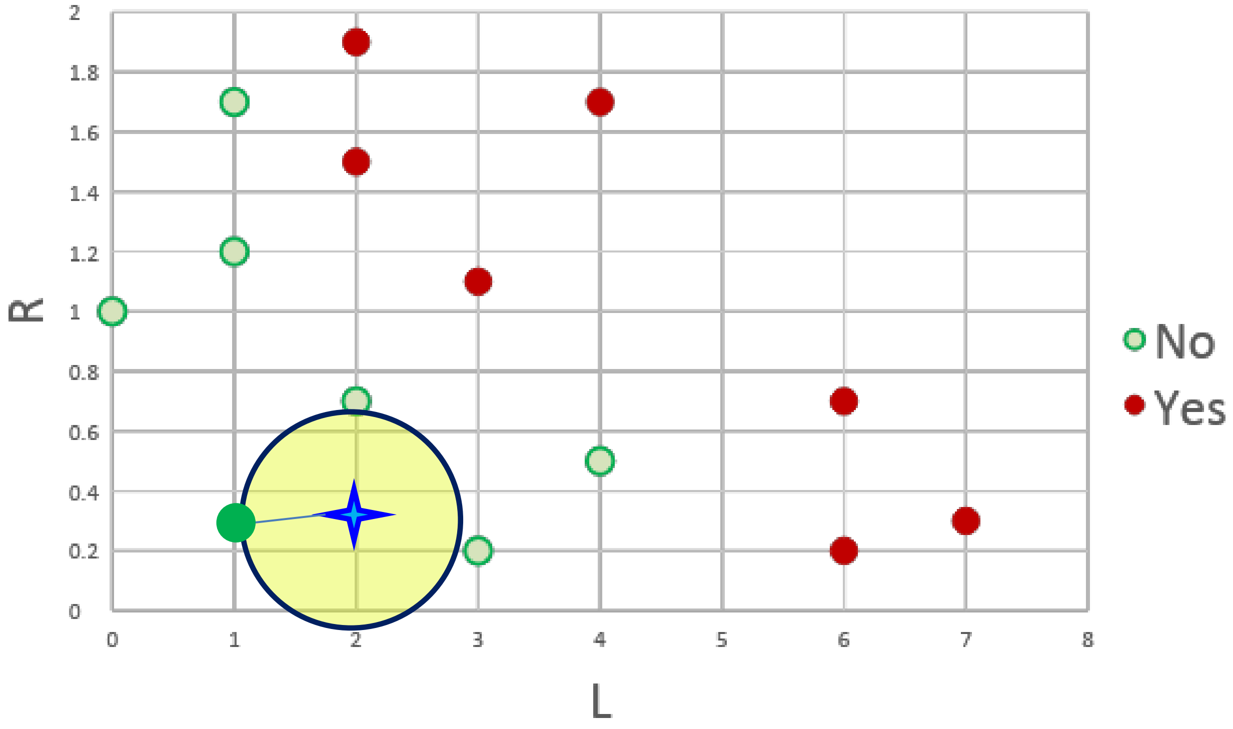
L	R
2	0.3



L: #late payments / year
R: expenses / income ratio

Predicting bankruptcy: nearest neighbor

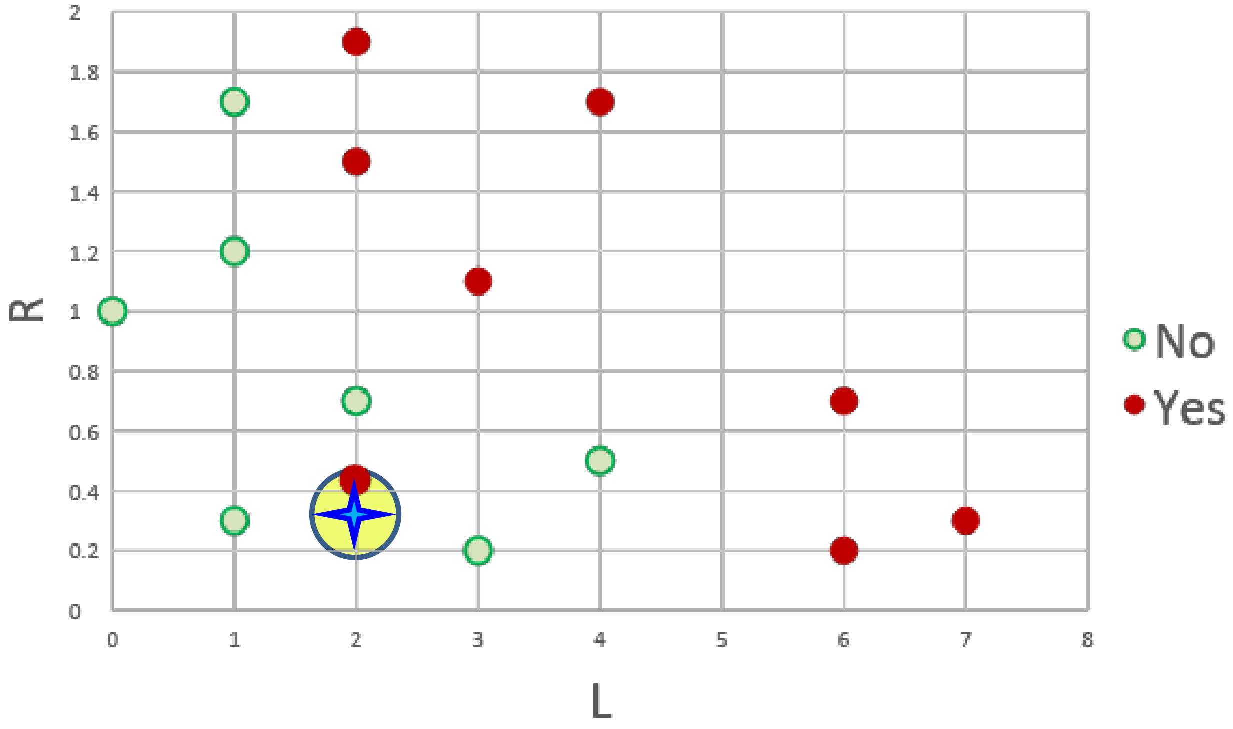
L	R
2	0.3



L: #late payments / year
R: expenses / income ratio

Predicting bankruptcy: noise

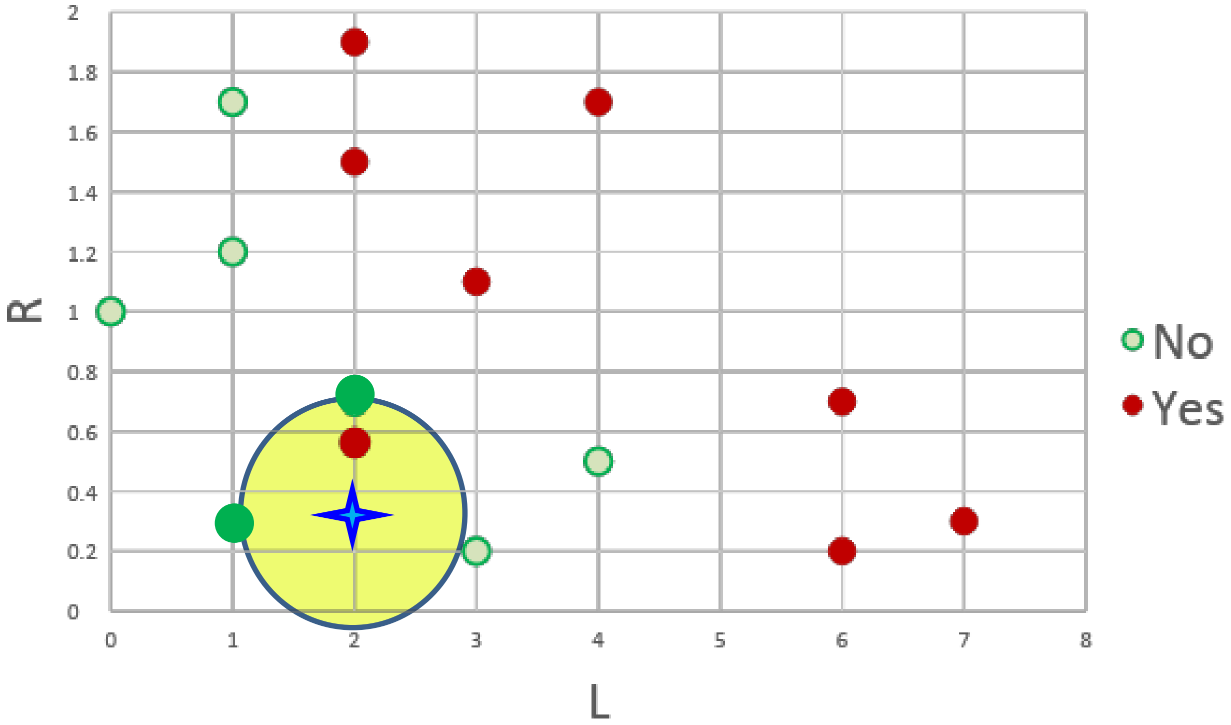
L	R
2	0.3



L: #late payments / year
R: expenses / income ratio

Predicting bankruptcy: *K* neighbors

L	R
2	0.3



L: #late payments / year
R: expenses / income ratio

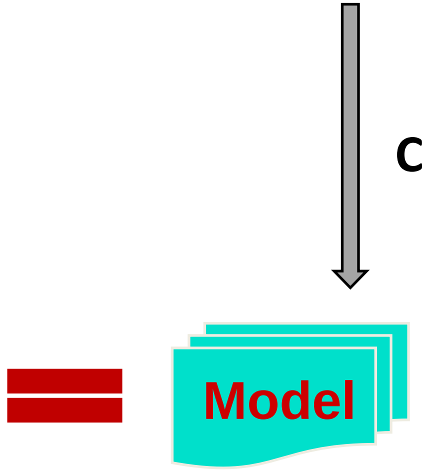
K-NN classifier: lazy classifier

Training set

Late payments, L	Spending ratio, R	Bankruptcy
3	Very low	No
1	Very low	No
4	Low	No
2	Low	No
0	Normal	No
1	Medium	No
1	High	No
6	Very low	Yes
7	Very low	Yes
6	Low	Yes
3	Normal	Yes
2	Medium	Yes
4	High	Yes
2	High	Yes

New sample

L	R	B
2	Low	?



L: #late payments / year
R: expenses / income ratio

K-NN classification algorithm

```
Input:
    set $T$ of $N$ labeled records,
    $K$,
    instance $A$ to classify

Classification:
    for $i$ from 1 to $N$
        compute distance $d\ (A,T_i)$
    sort $T$ asc by $d\ (A,T_i)$ into $T_{sorted}$
    from top $K$ records in $T_{sorted}$
        extract class labels $L_{1...K}$

Output:
    return combination $(L_{1...K})$
```

K-NN classification algorithm

Input:

set ***T*** of *N* labeled records,
K,
instance ***A*** to classify

Classification:

```
for $i$ from 1 to $N$
    compute distance $d(A, T_i)$
sort $T$ asc by $d(A, T_i)$ into $T_{sorted}$
from top $K$ records in $T_{sorted}$
    extract class labels $L_{1...K}$
```

Output:

```
return combination ($L_{1...K}$)
```

We need to discuss:

- How many neighbors: choice of K
- Distance/similarity function
- Combining neighbor class labels

At this point we just need to know that K should be odd

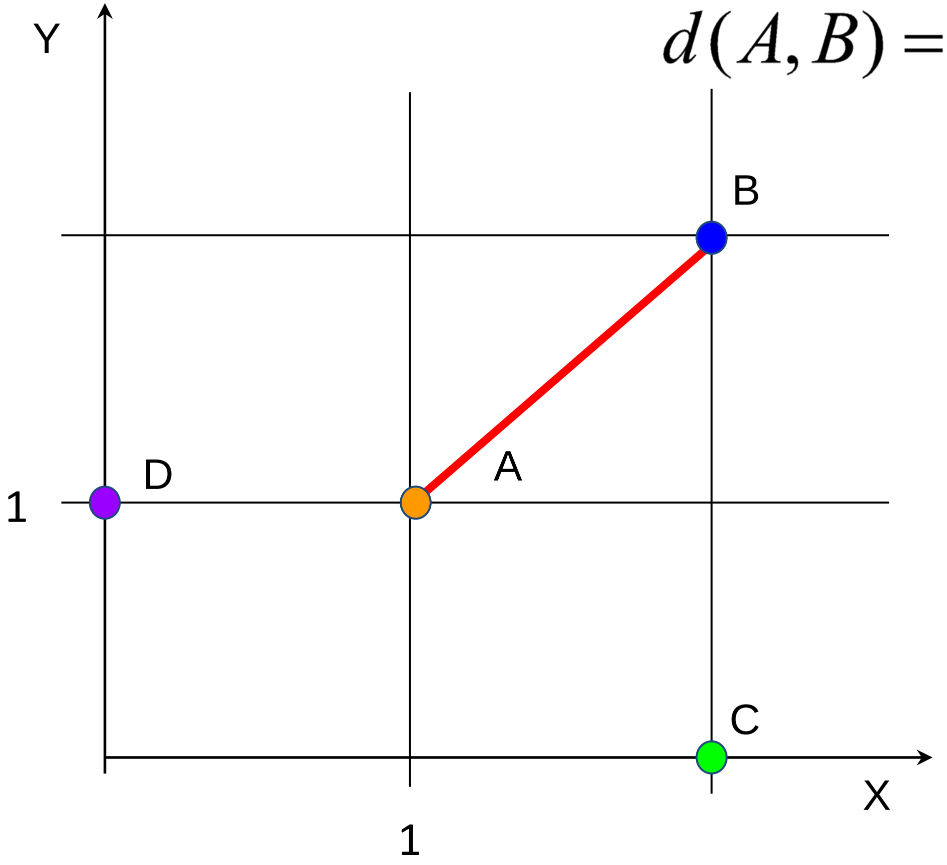
We need to discuss:

- How many neighbors: choice of K
- Distance/similarity function
- Combining neighbor class labels

If attributes are numeric: Simple distance function

Geometry: Euclidean distance

$$d(A,B) = \sqrt{| A_X - B_X |^2 + | A_Y - B_Y |^2}$$



$d(A,B)=?$

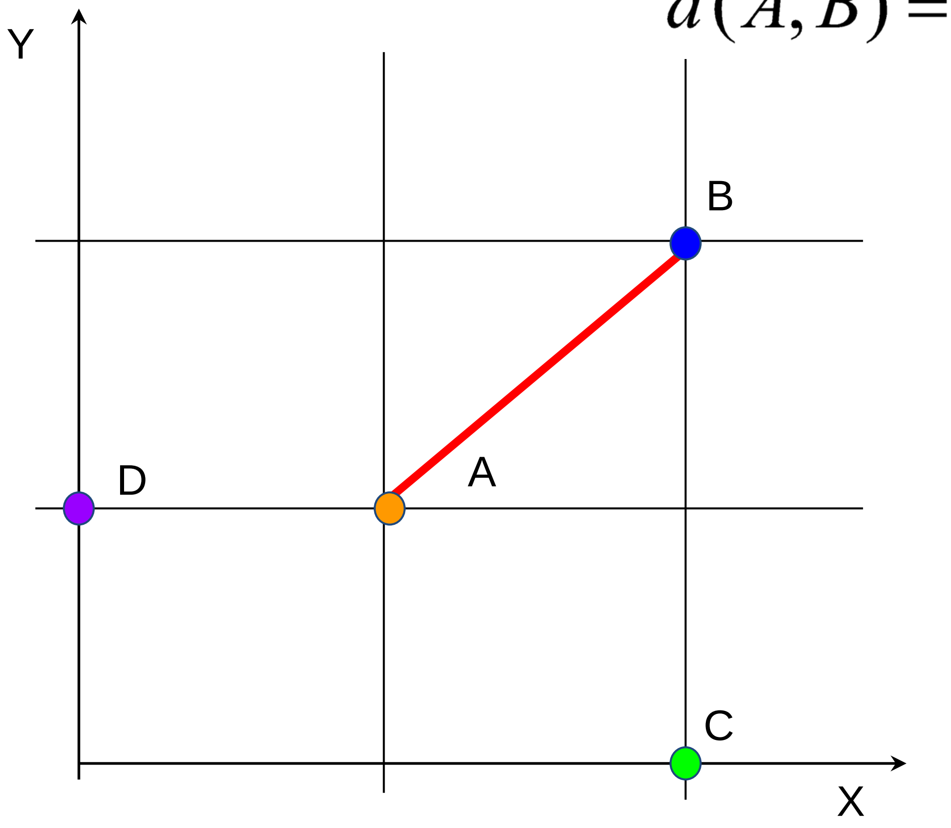
$d\ (B,C)=?$

$d(C,D)=?$

Simple distance function

Geometry: Euclidean distance

$$d(A,B) = \sqrt{|A_X - B_X|^2 + |A_Y - B_Y|^2}$$



For *N* dimensions:

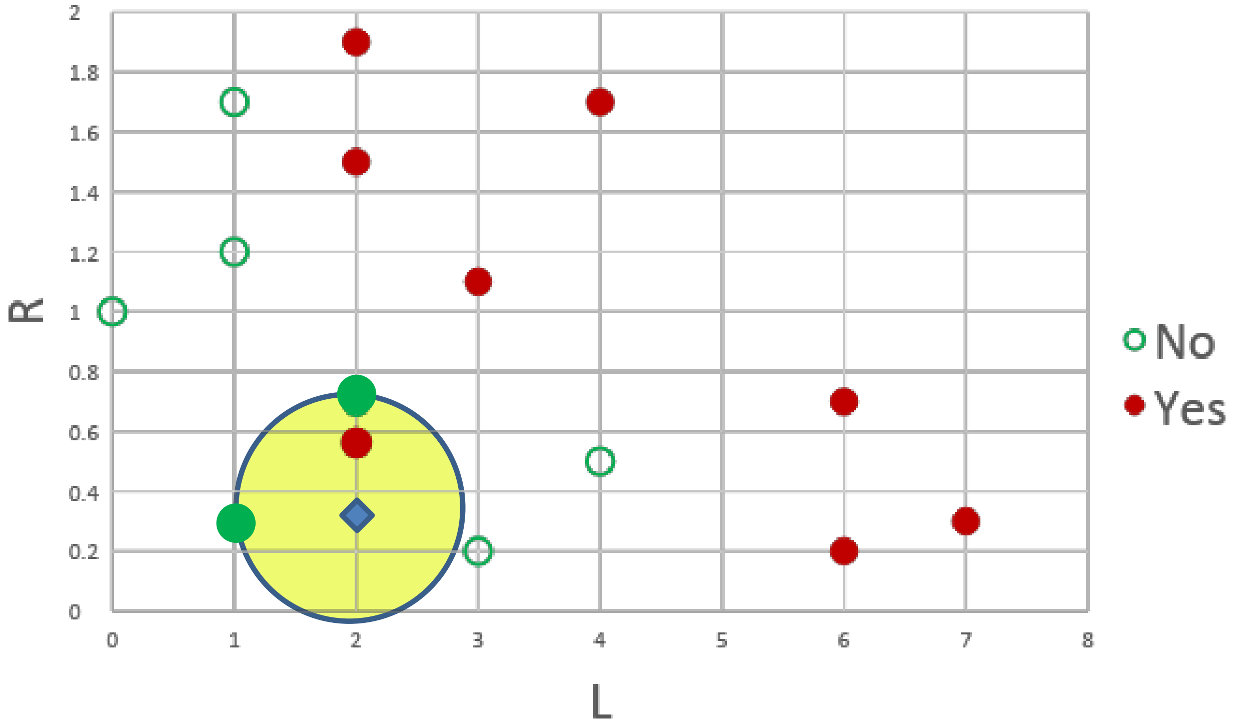
$$d(A,B) = \sqrt{\sum_{i=1}^{N} |A_i - B_i|^2}$$

We need to discuss:

- How many neighbors: choice of K
- Distance/similarity function
- Combining neighbor class labels

Simple combination function: **majority voting**

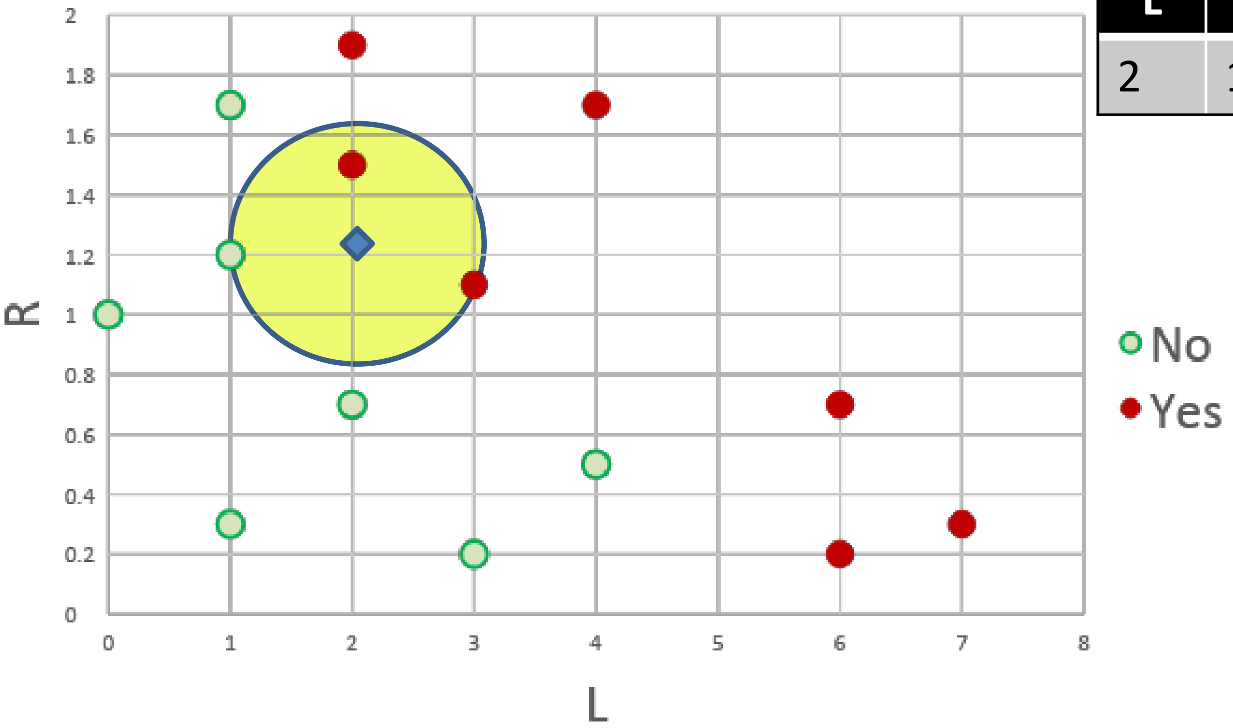
L	R
2	0.3



Classified as non-bankrupt

K-NN regressor:
simple combination function: **average**

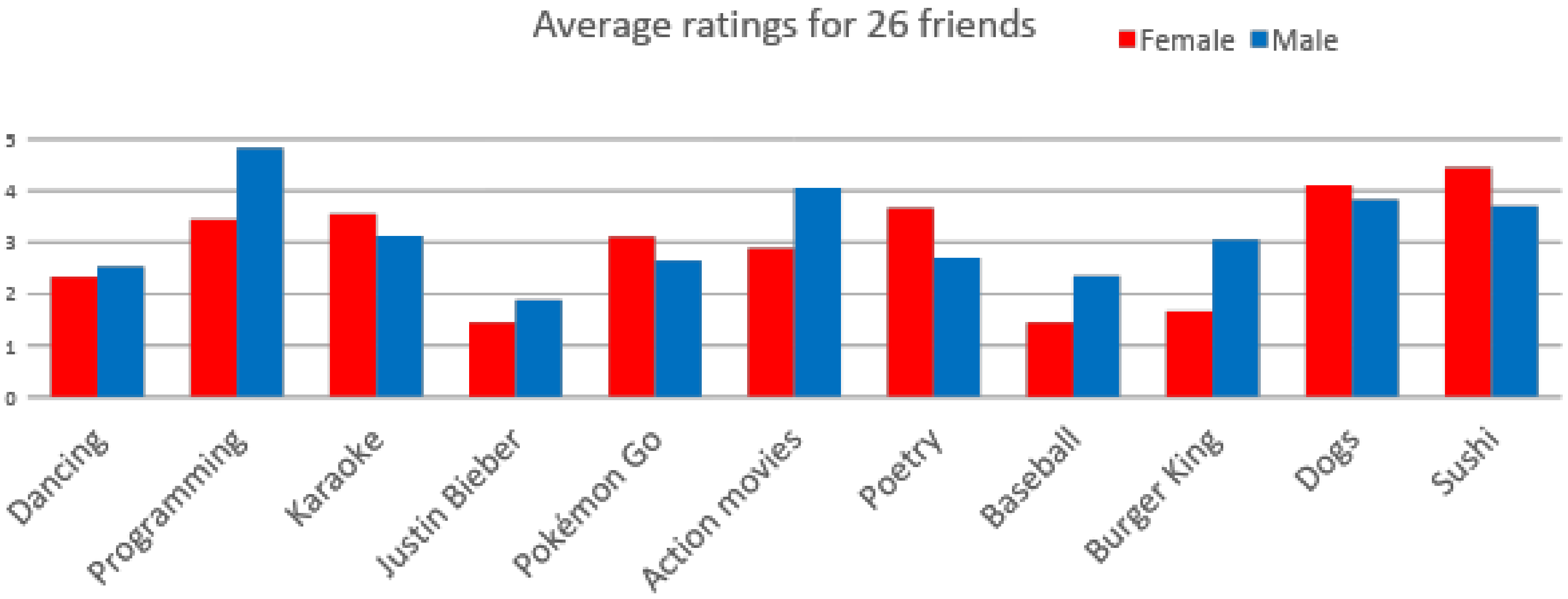
L	R	D
3	0.2	0
1	0.3	0
4	0.5	0
2	0.7	0
0	1	0
1	1.2	0
1	1.7	0
6	0.2	50K
7	0.3	100K
6	0.7	500K
3	1.1	25K
2	1.5	30K
4	1.7	150K
2	1.9	40K



L	R
2	1.2

Predicted default:
(0+30+25)/3=18K

My friends dataset



https://hope.simons-rock.edu/~mbarsky/intro19/lectures/data/predict/