DATA MINING 2

Exercises – KNN, Naïve Bayes, Lift Chart

Riccardo Guidotti, Salvatore Citraro

a.a. 2019/2020

UNIVERSITÀ DI PISA

K-NN

k-Nearest Neighbor Classifier

A medical expert is going to build up a case-based reasoning system for diagnosis tasks. Cases correspond to individual persons where the case problem parts are made up of a number of features describing possible symptoms and the solution parts represent the diagnosis (classification of disease). The case base contains the seven cases provided in the table below.

Training	Fever	Vomiting	Diarrhea	Shivering	Classification
c_1	no	no	no	no	healty (H)
c_2	average	no	no	no	influenza (I)
c_3	high	no	no	yes	influenza (I)
c_4	high	yes	yes	no	salmonella poisoning (S)
c_5	average	no	yes	no	salmonella poisoning (S)
c_6	no	yes	yes	no	bowel inflammation (B)
c_7	average	yes	yes	no	bowel inflammation (B)

Similarity provided by an expert

sim_F

q \ c	no	avg	high
no	1.0	0.7	0.2
avg	0.5	1.0	0.8
high	0.0	0.3	1.0

$sim_V = sim_D = sim_{Sh}$

q \ c	yes	no
yes	1.0	0.0
no	0.2	1.0

Weights

$w_F = 0.3$

$w_V = 0.2$

$W_D = 0.2$

$w_{Sh} = 0.3$

Classify the new instance q = (high; no; no; no) by applying the KNN algorithm with K=1,2,3

Calculate the similarity between all cases from the case base and the new instance q = (high; no; no; no)

c1 = (no; no; no; no):

Sim(q; c1) = 0.3*0.0 + 0.2 *1.0 + 0.2*1.0 + 0.3* 1.0 = 0.70

c2 = (average; no; no; no):

Sim(q; c2) = 0.3* 0.3 + 0.2 *1.0 + 0.2*1.0 + 0.3*1.0 = 0.79

c3 = (high; no; no; yes)

Sim(q; c3) = 0.3*1.0 + 0.2*1.0 + 0.2*1.0 + 0.3*0.2 = 0.76

c4 = (high; yes; yes; no):

Sim(q; c4) = 0.3*1.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.68

c5 = (average; no; yes; no):

Sim(q; c5) = 0.3*0.3 + 0.2*1.0 + 0.2*0.2 + 0.3*1.0 = 0.63

c6 = (no; yes; yes; no):

Sim(q; c6) = 0.3*0.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.28

c7 = (average; yes; yes; no):

Sim(q; c7) = 0.3*0.3 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.47

sim_F

q \ c	no	avg	high
no	1.0	0.7	0.2
avg	0.5	1.0	0.8
high	0.0	0.3	1.0

$sim_V = sim_D = sim_{Sh}$

q \ c	yes	no
yes	1.0	0.0
no	0.2	1.0

Weights

$w_F = 0.3$

$w_V = 0.2$

$W_D = 0.2$

$w_{Sh} = 0.3$

KNN Classification for K=1

c1 = (no; no; no; no):

Sim(q; c1) = 0.3*0.0 + 0.2 *1.0 + 0.2*1.0 + 0.3* 1.0 = 0.70

c2 = (average; no; no; no):

Sim(q; c2) = 0.3* 0.3 + 0.2 *1.0 + 0.2*1.0 + 0.3*1.0 = 0.79

c3 = (high; no; no; yes)

Sim(q; c3) = 0.3*1.0 + 0.2*1.0 + 0.2*1.0 + 0.3*0.2 = 0.76

c4 = (high; yes; yes; no):

Sim(q; c4) = 0.3*1.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.68

c5 = (average; no; yes; no):

Sim(q; c5) = 0.3*0.3 + 0.2*1.0 + 0.2*0.2 + 0.3*1.0 = 0.63

c6 = (no; yes; yes; no):

Sim(q; c6) = 0.3*0.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.28

c7 = (average; yes; yes; no):

Sim(q; c7) = 0.3*0.3 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.47

sim_F

q \ c	no	avg	high
no	1.0	0.7	0.2
avg	0.5	1.0	0.8
high	0.0	0.3	1.0

Weights

$w_F=0.3$

$w_V=0.2$

$W_D=0.2$

$w_{Sh}=0.3$

Class: Influenza

KNN Classification for K=2

c1 = (no; no; no; no):

Sim(q; c1) = 0.3*0.0 + 0.2 *1.0 + 0.2*1.0 + 0.3* 1.0 = 0.70

c2 = (average; no; no; no):

Sim(q; c2) = 0.3* 0.3 + 0.2 *1.0 + 0.2*1.0 + 0.3*1.0 = 0.79

c3 = (high; no; no; yes):

Sim(q; c3) = 0.3*1.0 + 0.2*1.0 + 0.2*1.0 + 0.3*0.2 = 0.76

c4 = (high; yes; yes; no):

Sim(q; c4) = 0.3*1.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.68

c5 = (average; no; yes; no):

Sim(q; c5) = 0.3*0.3 + 0.2*1.0 + 0.2*0.2 + 0.3*1.0 = 0.63

c6 = (no; yes; yes; no):

Sim(q; c6) = 0.3*0.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.28

c7 = (average; yes; yes; no):

Sim(q; c7) = 0.3*0.3 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.47

sim_F

q \ c	no	avg	high
no	1.0	0.7	0.2
avg	0.5	1.0	0.8
high	0.0	0.3	1.0

Weights

w_F=0.3

w_V=0.2

W_D=0.2

w_{Sh}=0.3

C2: Influenza

C3: Influenza

↓

Class: Influenza

KNN Classification for K=3

c1 = (no; no; no; no):
Sim(q; c1) = 0.3*0.0 + 0.2 *1.0 + 0.2*1.0 + 0.3* 1.0 = 0.70

c2 = (average; no; no; no):
Sim(q; c2) = 0.3* 0.3 + 0.2 *1.0 + 0.2*1.0 + 0.3*1.0 = 0.79

c3 = (high; no; no; yes):
Sim(q; c3) = 0.3*1.0 + 0.2*1.0 + 0.2*1.0 + 0.3*0.2 = 0.76

c4 = (high; yes; yes; no):
Sim(q; c4) = 0.3*1.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.68

c5 = (average; no; yes; no):
Sim(q; c5) = 0.3*0.3 + 0.2*1.0 + 0.2*0.2 + 0.3*1.0 = 0.63

c6 = (no; yes; yes; no):
Sim(q; c6) = 0.3*0.0 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.28

c7 = (average; yes; yes; no):
Sim(q; c7) = 0.3*0.3 + 0.2*0.2 + 0.2*0.2 + 0.3*1.0 = 0.47

sim_F

q \ c	no	avg	high
no	1.0	0.7	0.2
avg	0.5	1.0	0.8
high	0.0	0.3	1.0

Weights
w_F=0.3
w_V=0.2
W_D=0.2
w_{Sh}=0.3

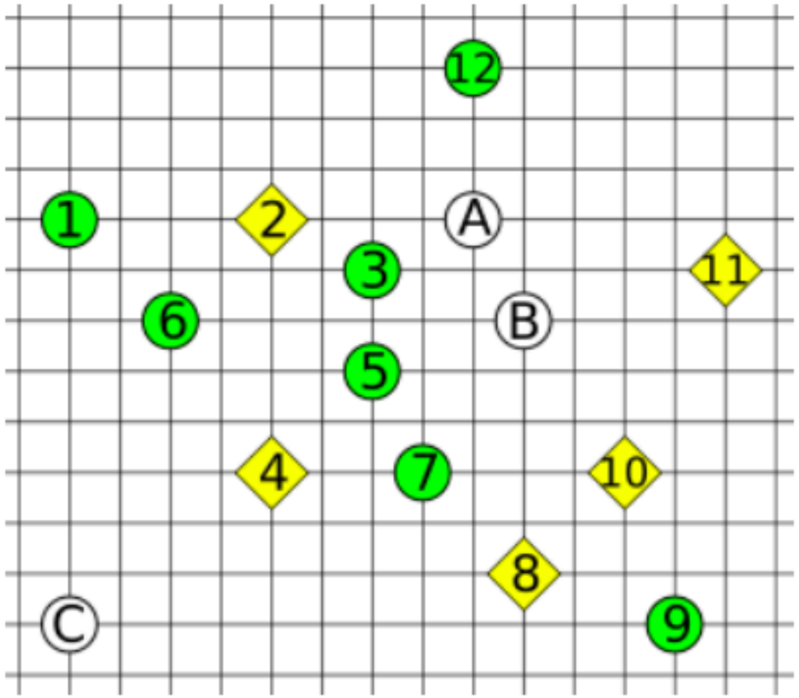
C1: healty
C2: Influenza
C3: Influenza

↓

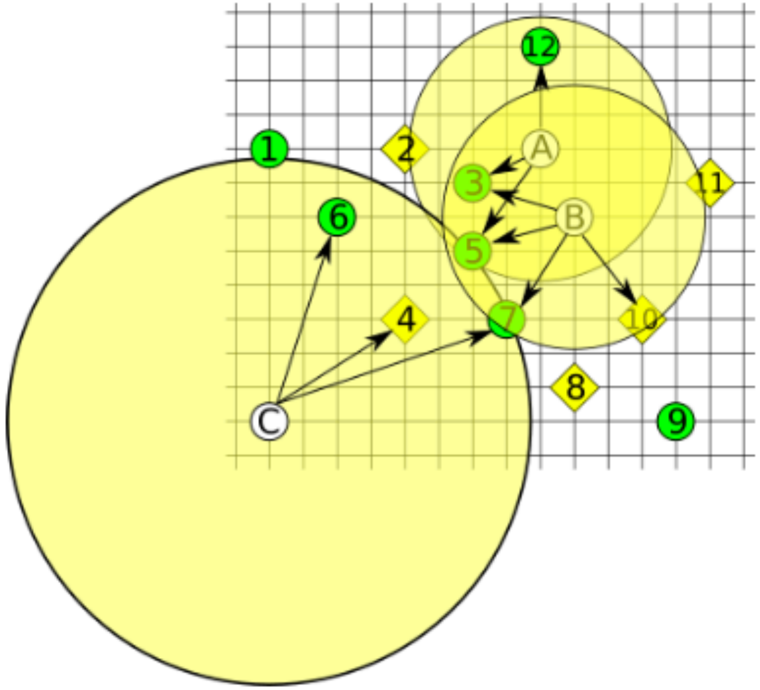
Class: Influenza

b) k-NN (**3 points**)

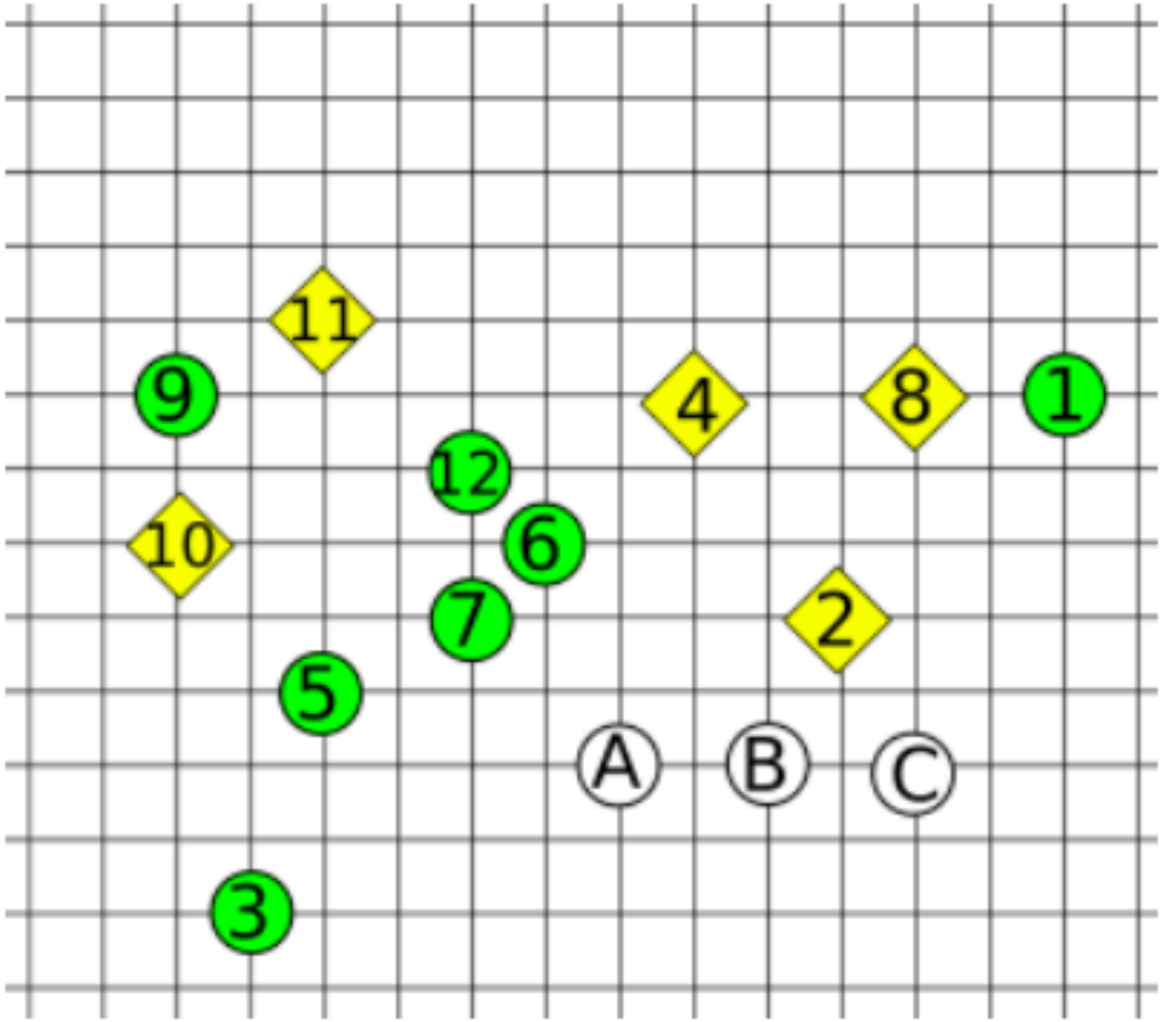
Given the training set on the right, composed of elements numbered from 1 to 12, and labelled as circles and diamonds, use it to classify the remaining 3 elements (letters A, B and C) using a k-NN classifier with k=3. For each point to classify, list the points of the dataset that belong to its k-NN set.
Notice: A, B and C belong to the test set, not to the training set. Also, the Euclidean distance should be used.



Answer:
kNN(A) = {3, 5, 12} → CIRCLE
kNN(B) = { 3, 5, 7, 10 } → CIRCLE
kNN(C) = { 4, 6, 7 } → CIRCLE



Given the training set on the right, composed of elements numbered from 1 to 12, and labelled as circles and diamonds, use it to classify the remaining 3 elements (letters A, B and C) using a k-NN classifier with k=3.
For each point to classify, list the points of the dataset that belong to its k-NN set.
Notice: A, B and C belong to the test set, not to the training set. Also, the Euclidean distance should be used.



Naïve Bayes

Play-tennis example. estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14	
P(n) = 5/14	

outlook	
P(sunny|p) =	P(sunny|n) =
P(overcast|p) =	P(overcast|n) =
P(rain|p) =	P(rain|n) =
temperature	
P(hot|p) =	P(hot|n) =
P(mild|p) =	P(mild|n) =
P(cool|p) =	P(cool|n) =
humidity	
P(high|p) =	P(high|n) =
P(normal|p) =	P(normal|n) =
windy	
P(true|p) =	P(true|n) =
P(false|p) =	P(false|n) =

Play-tennis example. estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14	
P(n) = 5/14	

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 1/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

Play-tennis example. estimating $P(x_i|C)$

$P(p) = 9/14$	
$P(n) = 5/14$	

Outlook	Temeprature	Humidity	Windy	Class
rain	hot	high	false	?

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 1/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

$P(X|p) \cdot P(p) =$

$P(X|n) \cdot P(n) =$

Play-tennis example. estimating $P(x_i|C)$

P(p) = 9/14
P(n) = 5/14

Outlook	Temeprature	Humidity	Windy	Class
rain	hot	high	false	N

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 1/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

$P(X|p)\cdot P(p)$ = $P(rain|p)\cdot P(hot|p)\cdot P(high|p)\cdot P(false|p)\cdot P(p) = 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$

$P(X|n)\cdot P(n)$ = $P(rain|n)\cdot P(hot|n)\cdot P(high|n)\cdot P(false|n)\cdot P(n) = 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 =$ 0.018286

Example of Naïve Bayes Classifier

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

A: attributes

M: mammals

N: non-mammals

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

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

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

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

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

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

=> Mammals

a) Naive Bayes (**3 points**)
Given the training set below, build a Naive Bayes classification model (i.e. the corresponding table of probabilities) using (i) the normal formula and (ii) using Laplace formula. What are the main effects of Laplace on the models?

A	B	class
no	green	N
no	red	Y
yes	green	N
no	red	N
no	red	Y
no	green	Y
yes	green	N

Answer:

Normal

	Y	N		Y	N
	3	4		0.43	0.57
	A | Y	A | N		A | Y	A | N
yes	0	2	yes	0.00	0.50
no	3	2	no	1.00	0.50
	B | Y	B | N		B | Y	B | N
green	1	3	green	0.33	0.75
red	2	1	red	0.67	0.25

Laplace

	Y	N		Y	N
	3	4		0.43	0.57
	A | Y	A | N		A | Y	A | N
yes	0	2	yes	0.20	0.50
no	3	2	no	0.80	0.50
	B | Y	B | N		B | Y	B | N
green	1	3	green	0.40	0.67
red	2	1	red	0.60	0.33

a) Naive Bayes (**3 points**)
Given the training set on the left, build a Naive Bayes classification model and apply it to the test set on the right.

SCORE	FIRST-TRY	FACULTY	class
good	no	science	Y
medium	yes	science	N
bad	yes	science	N
bad	yes	humanities	Y
good	no	humanities	N
good	no	science	Y
medium	no	humanities	Y

SCORE	FIRST-TRY	FACULTY	class
bad	no	humanities	
good	yes	science	
medium	yes	humanities	

Lift Chart

Exercise on Lift charts

- We are given a test set with the real labels

X	Y	Z	Class
7	8	45	Yes
30	8	40	No
13	23	21	No
47	43	34	No
37	10	29	Yes
19	49	31	No
20	13	8	Yes
33	44	16	Yes
47	12	41	No
49	21	3	Yes

- Our model provides the following predictions and associated confidences
- Plot the corresponding Lift chart

X	Y	Z	Class	Predicted Class	Confidence of Prediction
7	8	45	Yes	No	0.5618863452
30	8	40	No	Yes	0.6701976614
13	23	21	No	No	0.6816996196
47	43	34	No	No	0.8795983369
37	10	29	Yes	Yes	0.8245785853
19	49	31	No	Yes	0.8194210517
20	13	8	Yes	Yes	0.5079911998
33	44	16	Yes	No	0.5736005213
47	12	41	No	No	0.8702045378
49	21	3	Yes	Yes	0.7856012356

- Step 1: focus on relevant information

Class	Predicted Class	Confidence of Prediction
Yes	No	0.5618863452
No	Yes	0.6701976614
No	No	0.6816996196
No	No	0.8795983369
Yes	Yes	0.8245785853
No	Yes	0.8194210517
Yes	Yes	0.5079911998
Yes	No	0.5736005213
No	No	0.8702045378
Yes	Yes	0.7856012356

- Step 2: from prediction and confidence, derive the score
  - Score = probability of having positive

Class	Predicted Class	Confidence of Prediction	Score
Yes	No	0.5618863452	0.4381136548
No	Yes	0.6701976614	0.6701976614
No	No	0.6816996196	0.3183003804
No	No	0.8795983369	0.1204016631
Yes	Yes	0.8245785853	0.8245785853
No	Yes	0.8194210517	0.8194210517
Yes	Yes	0.5079911998	0.5079911998
Yes	No	0.5736005213	0.4263994787
No	No	0.8702045378	0.1297954622
Yes	Yes	0.7856012356	0.7856012356

- Sort records according to score
  - Descending order (most likely positives first)

Sorted Class	Sorted Scores
Yes	0.8245785853
No	0.8194210517
Yes	0.7856012356
No	0.6701976614
Yes	0.5079911998
Yes	0.4381136548
Yes	0.4263994787
No	0.3183003804
No	0.1297954622
No	0.1204016631

- Evaluate, for ech possibile threshold, how many true positives we captures

Sorted Class	Sorted Scores	True Positives
		0
Yes	0.8245785853	
No	0.8194210517	
Yes	0.7856012356	
No	0.6701976614	
Yes	0.5079911998	
Yes	0.4381136548	
Yes	0.4263994787	
No	0.3183003804	
No	0.1297954622	
No	0.1204016631	

- Evaluate, for ech possibile threshold, how many true positives we captures

Sorted Class	Sorted Scores	True Positives
		0
Yes	0.8245785853	1
No	0.8194210517	1
Yes	0.7856012356	2
No	0.6701976614	2
Yes	0.5079911998	3
Yes	0.4381136548	4
Yes	0.4263994787	5
No	0.3183003804	5
No	0.1297954622	5
No	0.1204016631	5

- Plot

Sorted Class	Sorted Scores	True Positives
		0
Yes	0.8245785853	1
No	0.8194210517	1
Yes	0.7856012356	2
No	0.6701976614	2
Yes	0.5079911998	3
Yes	0.4381136548	4
Yes	0.4263994787	5
No	0.3183003804	5
No	0.1297954622	5
No	0.1204016631	5