Faculty of Computing and Information Technology/Department of Computer Science and Mathematics

BACS3013 Data Science

Tutorial 10 (Supervised Learning - part 1)

Q1. Suppose we want to classify potential bank customers as good creditors or bad creditors for loan applications. We have a training dataset describing past customers using the following attributes:

Marital status {married, single, divorced}, Gender {male, female}, Age {[18..30[, [30..50[, [50..65[, [65+]}, Income {[10K..25K[, [25K..50K[, [50K..65K[, [65K..100K[, [100K+]}.

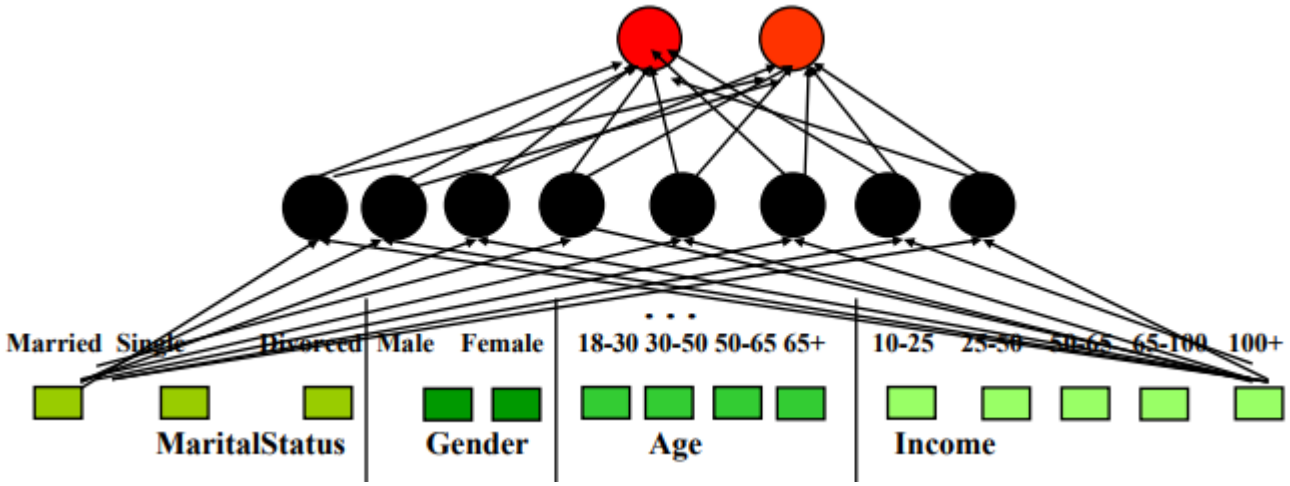
Design a neural network that could be trained to predict the credit rating of an applicant. A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Ans:

We have 2 classes, good creditor and bad creditor. This means we would need two nodes in the output layer.

There are 4 variables: Marital Status, Gender, Age and Income. However, since we have 3 values for Marital status, 2 values for Gender, 4 intervals for Age and 5 intervals for Income, we would have 14 neuron units in the input layer.

In the hidden layer, we can have (14+2)/2=8 neurons (*no fixed rule)
The architecture of the neural networks could look like this:



BACS3013 Data Science

Q2. Given the training data in the table below (Buy Computer data), predict the class of the following example using Naïve Bayes classification: age<=30, income=medium, student=yes, credit-rating=fair

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Ans:

E = age<=30, income=medium, student=yes, credit-rating=fair
E1 is age<=30, E2 is income=medium, student=yes, E4 is credit-rating=fair
We need to compute P(yes | E) and P(no | E) and compare them.

$$P(yes \mid E) = \frac{P(E_1 \mid yes)\,P(E_2 \mid yes)\,P(E_3 \mid yes)\,P(E_4 \mid yes)\,P(yes)}{P(E)}$$

P(yes)=9/14=0.643 P(no)=5/14=0.357

P(E1|yes)=2/9=0.222 P(E1|no)=3/5=0.6
P(E2|yes)=4/9=0.444 P(E2|no)=2/5=0.4
P(E3|yes)=6/9=0.667 P(E3|no)=1/5=0.2
P(E4|yes)=6/9=0.667 P(E4|no)=2/5=0.4

$$P(yes \mid E) = \frac{0.222\ 0.444\ 0.667\ 0.668\ 0.443}{P(E)} = \frac{0.028}{P(E)} \qquad P(no \mid E) = \frac{0.6\ 0.4\ 0.2\ 0.4\ 0.357}{P(E)} = \frac{0.007}{P(E)}$$

Hence, the Naïve Bayes classifier predicts buys_computer=yes for the new example.

BACS3013 Data Science

Q3. Given the training data in Q2, predict the class of the following new example using k-Nearest Neighbour for k=5: age<=30, income=medium, student=yes, credit-rating=fair. For similarity measure use a simple match of attribute values:

Similarity(A,B)= $\sum_{i=1}^{4} w_i * \partial(a_i, b_i) \Big/ 4$

where $\partial(a_i, b_i)$ is 1 if a_i equals b_i and 0 otherwise. a_i and b_i are either age, income, student or credit_rating. Weights are all 1 except for income it is 2.

Ans:

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

RID	Class	Distance to New
1	No	(1+0+0+1)/4=0.5
2	No	(1+0+0+0)/4=0.25
3	Yes	(0+0+0+1)/4=0.25
4	**Yes**	**(0+2+0+1)/4=0.75**
5	Yes	(0+0+1+1)/4=0.5
6	No	(0+0+1+0)/4=0.25
7	Yes	(0+0+1+0)/4=0.25
8	**No**	**(1+2+0+1)/4=1**
9	**Yes**	**(1+0+1+1)/4=0.75**
10	**Yes**	**(0+2+1+1)/4=1**
11	**Yes**	**(1+2+1+0)/4=1**
12	Yes	(0+2+0+0)/4=0.5
13	Yes	(0+0+1+1)/4=0.5
14	No	(0+2+0+0)/4=0.5

Among the five nearest neighbours four are from class Yes and one from class No. Hence, the k-NN classifier predicts buys_computer=yes for the new example.