Tutorial Week 12

The following **training data** (Table 1) shows whether the bank approves a credit card application based on the information of the applicant's job status (**nominal feature**), marital status (**nominal feature**) and annual income (**numerical feature**).

Table 1: Training set

Permanent Job	Marital Status	Annual Income	Approved?
Yes	Single	130K	Yes
No	Married	80K	No
No	Single	100K	Yes
Yes	Divorced	90K	Yes
No	Single	60K	No
Yes	Married	120K	Yes
Yes	Single	85K	Yes
No	Divorced	110K	No
Yes	Married	95K	Yes
No	Married	125K	Yes

We also have the **test data** (Table 2) as follows

Table 2: Test set

Permanent Job	Marital Status	Annual Income	Approved?
No	Single	60K	No
Yes	Married	100K	Yes
Yes	Single	90K	Yes
No	Divorced	95K	No
No	Married	85K	No

If we use a k-NN classifier with **Euclidean distance** to predict the test data based on the **training data**, what is the best k among {1, 3, 5} if we use **accuracy** as the performance measurement?

1. transform the data into numerical value (One hot)
2. normalise the data x' = (x-u)/ sigma
3. calculate the distance
4. find the nearest neighbours
5. predict by majority vote

Answers

For the original data, we first transfer it into numerical form (One-hot):

Permanent Job	Single?	Married?	Divorced?	Annual Income
1	1	0	0	130
0	0	1	0	80
0	1	0	0	100
1	0	0	1	90
0	1	0	0	60
1	0	1	0	120
1	1	0	0	85
0	0	0	1	110
1	0	1	0	95
0	0	1	0	125

Standardize/ Normalise the data by $x' = \frac{x-\mu}{\sigma}$, please note that the variance is unbiased variance (divided by n-1).

	Permanent Job	Single?	Married?	Divorced?	Annual Income
μ	0.5	0.4	0.4	0.2	99.5
σ	0.53	0.52	0.52	0.42	22.04

Permanent Job	Single?	Married?	Divorced?	Annual Income
0.94	1.15	-0.77	-0.48	1.38
-0.94	-0.77	1.15	-0.48	-0.88
-0.94	1.15	-0.77	-0.48	0.02
0.94	-0.77	-0.77	1.90	-0.43
-0.94	1.15	-0.77	-0.48	-1.79
0.94	-0.77	1.15	-0.48	0.93
0.94	1.15	-0.77	-0.48	-0.66
-0.94	-0.77	-0.77	1.90	0.48
0.94	-0.77	1.15	-0.48	-0.20
-0.94	-0.77	1.15	-0.48	1.16

Transform and normalise test data:

Permanent Job	Single?	Married?	Divorced?	Annual Income
0	1	0	0	60
1	0	1	0	100
1	1	0	0	90
0	0	0	1	95
0	0	1	0	85

Use same parameters derived from training set for normalisation:

	Permanent Job	Single?	Married?	Divorced?	Annual Income
u	0.5	0.4	0.4	0.2	99.5
σ	0.53	0.52	0.52	0.42	22.04

Permanent Job	Single?	Married?	Divorced?	Annual Income
-0.94	1.15	-0.77	-0.48	-1.79
0.94	-0.77	1.15	-0.48	0.02
0.94	1.15	-0.77	-0.48	-0.43
-0.94	-0.77	-0.77	1.90	-0.20
-0.94	-0.77	1.15	-0.48	-0.66

The distance matrix is:

	1	2	3	4	5	6	7	8	9	10
1	3.69	2.87	1.81	3.84	0.01	4.28	2.20	3.81	3.67	4.01
2	3.04	2.09	3.31	3.09	3.77	0.91	2.80	3.62	0.22	2.20
3	1.81	3.34	1.94	3.06	2.32	3.04	0.23	3.70	2.73	3.67
4	3.93	3.13	3.07	1.90	3.45	3.77	3.62	0.68	3.59	3.35
5	3.88	0.22	2.80	3.60	2.94	2.46	3.31	3.26	1.94	1.82
	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes

Example: distance between training data 1 and test data 1

0.94	1.15	-0.77	-0.48	1.38

-0.94	1.15	-0.77	-0.48	-1.79

$$\sqrt{(0.94+0.94)^2\ +(1.15\text{-}1.15)^2+(-0.77+0.77)^2+(-0.48+0.48)^2+(1.38+1.79)^2}$$
$$=\ 3.69$$

Permanent Job	**Marital Status**	**Annual Income**	**K = n/ Accuracy**			**GT**
			K=1 100%	**K=3** 40%	**K=5** 40%	
No	Single	60K	No	Yes	Yes	No
Yes	Married	100K	Yes	Yes	Yes	Yes
Yes	Single	90K	Yes	Yes	Yes	Yes
No	Divorced	95K	No	Yes	Yes	No
No	Married	85K	No	Yes	Yes	No

So, we chose K = 1 (highest accuracy).