Artificial Intelligence (COMP111)

Exercise 9

To answer the questions, please watch the videos under *Learning*.

Your answer to Question 1 should be submitted on canvas for assignment *Exercise 9* either as a text entry, a text file (txt), a pdf file, or a photo of the handwritten solution. The deadline is Monday, 14th of December, at 6pm. You should also attempt to answer the other questions before your tutorial (but not submit them).

You obtain 2 points (2 percent of the final mark) if you make a reasonable attempt to answer Question 1 *and* actively participate in your tutorial in the week starting Monday 14th of December.

We would like to encourage you to discuss the questions with your fellow students, but do not copy your answer from anybody else.

1. Your friend has a strong opinion on cars: he either likes a car or does not like it at all. We have collected the following training data about his preferences using the features

   - *Color* taking values: Red, Yellow;
   - *Type* taking values: Sports, SUV;
   - *Origin* taking values: Domestic, Imported.

No	*color*	*type*	*origin*	*likes*
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

   Use a naive Bayes' classifier to predict whether your friend likes a Red Domestic SUV. Explain your computation in detail.

2. Compute the Euclidean distance between $(3, 7, 5)$ and $(5, 3, 7)$.

3. Consider the following data about the weather on days on which Peter plays tennis or not. We would like to predict, using a k-Nearest Neighbor classifier, whether Peter plays tennis on a particular day depending on the weather.

Day	Outlook	Temperature	Humidity	Wind	PlaysTennis
D1	Sunny	26	High	Strong	No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain	21	Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8	High	Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D11	Overcast	21	High	Strong	Yes
D12	Rain	26	Normal	Weak	Yes
D13	Rain	24	High	Strong	No
D14	Sunny	23	Normal	Weak	No
D15	Sunny	21	Normal	Weak	Yes

To this end, we first have to define a distance function $d(D, D')$ between days D and D'. There are many options. We could simply count the number of features (from Outlook, Temperature, Humidity, Wind) on which D and D' differ (matching distance). In this case, however, we would not take into account by how much the temperature differs. On the other hand, if we count the number of features from Outlook, Humidity, and Wind on which D and D' differ and add the difference in temperature, then the temperature would dominate the other features. We go for the following compromise: normalise the difference in temperature by dividing by $21 = 29 - 8$ (the highest temperature minus the lowest temperature in the data and the days we want to classify). Then define:

$$d(D, D') = n + \frac{|\text{temp}(D) - \text{temp}(D')|}{21},$$

where n is the number of features from Outlook, Humidity, and Wind for which D and D' differ. For example,

$$d(D1, D2) = \frac{2}{21}, \quad d(D1, D3) = 2\frac{3}{21}$$

The questions are now as follows:

1. Determine the prediction of the k-nearest neighbor classifier for D16 given by the following values
   - Outlook = Sunny, Temperature = 23, Humidity= High, Wind=Strong

   for all k between 1 and 12. If the number of k-nearest neighbors on which Peter plays tennis and on which he does not play tennis coincide, then the prediction is undecided.
2. Argue that d defines a metric on the days D1, D2,..., D15. That is, it satisfies the following for all x, y, z:
   - $d(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles);
   - $d(x, y) = d(y, x)$ (symmetry);
   - $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality).