COMP5318/COMP4318 Machine Learning and Data Mining

Week 2 Tutorial exercises
K-Nearest Neighbor.

Welcome to your first COMP5318/COMP4318 tutorial! Please note the following:

- **For most of the weeks there will be 2 documents with tutorial exercises:**
  1) **theoretical (as this one), involving paper-based exercises and calculations, testing your understanding of the algorithms**
  2) **practical using Python and its machine learning and neural network libraries (available in 2 formats: ipynb (Jupyter Notebook) and pdf – see Canvas)**
- **Theoretical: We will do some of these exercises at the lecture (usually the first exercise). The rest should be done at your own time. Make sure that you do all theoretical exercises as they are similar in style to the exam questions.**
- **Practical: This will be the main focus of the tutorial. Sometimes it may not be possible to finish all Python exercises during the tutorial. Please do this at home as this part is important for your assignments. We have prepared very detailed notes for the practical part, we hope you will find them useful.**
- **The solutions for both type of exercises will be provided on Friday evening after the last tutorial – please see the Unit Outline (detailed) document of Canvas.**

<u>Exercise 1.</u> Nearest Neighbor

The dataset below consists of 4 examples described with 3 numeric features (a1, a2 and a3); the class has 2 values: yes and no.

What will be the prediction of 1-Nearest Neighbor (1-NN) and 3-Nearest Neighbor (3-NN) with Euclidian distance for the following new example: a1=2, a2=4, a3=2?

Assume that all attributes are measured on the same scale – no need for normalization.

	a1	**a2**	**a3**	**class**
1	1	3	1	yes
2	3	5	2	yes
3	3	2	2	no
4	5	2	3	no

Exercise adapted from M. Kubat, Introduction to Machine Learning, Springer, 2017

<u>Exercise 2.</u> Nearest neighbor with nominal features

Consider the *iPhone* dataset given below. There are 4 nominal attributes (age, income, student, and credit_rating) and the class is buys_iPhone with 2 values: yes and no.

What would be the prediction of 1-NN and 3-NN for the following new example:
age<=30, income=medium, student=yes, credit-rating=fair

If there are ties, make random selection.

Tip: As the examples are described with nominal attributes, when calculating the distance use the following rule:

difference=1 between 2 values that are not the same
difference=0 between 2 values that are the same
e.g. D(1, new)=sqrt(0+1+1+0=)=sqrt(2)

	age	**income**	**student**	**credit rating**	**buy iPhone**
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	excellent	no
6	[31,40]	low	yes	excellent	yes
7	<=30	medium	no	fair	no
8	[31,40]	medium	no	excellent	yes
9	>40	medium	no	excellent	no

Dataset adapted from J. Han and M. Kamber, Data Mining, Concepts and Techniques, Morgan Kaufmann.