Artificial Intelligence I 2022/2023

Week 8 Tutorial and Additional Exercises

k-Nearest Neighbours

School of Computer Science

April 21, 2023

In this tutorial...

In this tutorial we will be covering

- Distance metrics.
- Normalisation.
- k-nearest neighbours.
- Optional theoretical exercises.

Parametric and non-parametric models

- *Parametric models* are learning models that summarise data with a set of parameters.
- Linear and logistic regression are examples of parametric models.
- *Non-parametric models* are learning models that do not assume any parameters.
- In this tutorial, we will study a non-parametric model.

Distance metrics

- A *distance metric* is a way to quantify the similarity or dissimilarity between instances.
- There are many available distance metrics. We need to choose the one that best fits the problem at hand.
- A distance metric takes two vectors as inputs and outputs a non-negative number.
- Different notions of distance metrics are used for vectors of numerical variables and for vectors of categorical variables.

Distance metrics (continued)

- For numerical variables, we will use the *Minkowski distance*.
- Given a number $p \geq 1$ and two vectors with d numerical variables

$$\mathbf{x}^{(1)} = (x_1^{(1)}, \ldots, x_d^{(1)}) \quad \text{and} \quad \mathbf{x}^{(2)} = (x_1^{(2)}, \ldots, x_d^{(2)})$$

their Minkowski distance (or L^p-norm) is defined as

$$L^p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sqrt[p]{\sum_{j=1}^{d} \left| x_j^{(1)} - x_j^{(2)} \right|^p}.$$

- For $p = 2$ we obtain the *Euclidean distance*.
- For $p = 1$ we obtain the *Manhattan distance*.

Exercise 1

- Consider the following vectors with 3 numerical variables.

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0 \\ 3 \\ -1 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} -2 \\ 3 \\ -1 \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

- Compute the Euclidean and Manhattan distance matrices for these vectors.
- Hint: You need to compute 6 distances on total.

Exercise 1: Solution

- The Euclidean distance matrix is the following:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$
$\mathbf{x}^{(1)}$	0	2.000	4.583	3.162
$\mathbf{x}^{(2)}$	2.000	0	5.385	3.742
$\mathbf{x}^{(3)}$	4.583	5.385	0	1.732
$\mathbf{x}^{(4)}$	3.162	3.742	1.732	0

- The Manhattan distance matrix is the following:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$
$\mathbf{x}^{(1)}$	0	2	7	4
$\mathbf{x}^{(2)}$	2	0	9	6
$\mathbf{x}^{(3)}$	7	9	0	3
$\mathbf{x}^{(4)}$	4	6	3	0

- What do you notice when comparing these matrices?

Distance metrics (continued)

- For categorical variables, we will use the *Hamming distance*.
- Given two vectors with d categorical variables

$$\mathbf{x}^{(1)} = (x_1^{(1)}, \ldots, x_d^{(1)}) \quad \text{and} \quad \mathbf{x}^{(2)} = (x_1^{(2)}, \ldots, x_d^{(2)}),$$

their Hamming distance is defined as

$$H(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sum_{j=1}^{d} \mathbf{1}(x_j^{(1)} \neq x_j^{(2)})$$

where $\mathbf{1}$ is the indicator function. For all $j = 1, \ldots, d$, we have

$$\mathbf{1}(x_j^{(1)} \neq x_j^{(2)}) = \begin{cases} 1 \text{ if } x_j^{(1)} \neq x_j^{(2)} \\ 0 \text{ if } x_j^{(1)} = x_j^{(2)}. \end{cases}$$

Exercise 2

- Consider the following vectors with 1 ordinal variable and 3 categorical variables.

$$\mathbf{x}^{(1)} = \begin{bmatrix} yes \\ red \\ FR \\ \triangle \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} yes \\ blue \\ FR \\ \square \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} no \\ green \\ UK \\ \bigcirc \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} yes \\ red \\ DE \\ \triangle \end{bmatrix}.$$

- Find the Hamming distance matrix for these vectors.
- Hint: You need to compute 6 distances on total. Each distance can be at most 4.

Exercise 2: Solution

- The Hamming distance matrix is the following:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$
$\mathbf{x}^{(1)}$	0	2	4	1
$\mathbf{x}^{(2)}$	2	0	4	3
$\mathbf{x}^{(3)}$	4	4	0	4
$\mathbf{x}^{(4)}$	1	3	4	0

- Notice that each distance is at most 4.

Normalisation

- *Normalisation* is used to restrict numerical variables in $[0, 1]$.
- Given a set of n vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, with d numerical variables, for all $j = 1, \dots, d$, we write

$$\min_j = \min\{x_j^{(1)}, \dots, x_j^{(n)}\} \text{ and } \max_j = \max\{x_j^{(1)}, \dots, x_j^{(n)}\}.$$

- Then, the j-th variable of the i-th vector is normalised as

$$\mathit{normalise}(x_j^{(i)}) = \frac{x_j^{(i)} - \min_j}{\max_j - \min_j}.$$

- We calculate the above formula for all $i = 1, \dots, n$ and for all $j = 1, \dots, d$ and normalise all variables in all vectors.

Exercise 3

- Consider the following vectors with 3 numerical variables.

$$\mathbf{x}^{(1)} = \begin{bmatrix} -2 \\ 3 \\ 300 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 2 \\ 1 \\ -100 \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} 0 \\ 2 \\ 100 \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} 1 \\ 2 \\ -200 \end{bmatrix}.$$

- Normalise all variables in all vectors, using the methodology we just described.
- Hint: First compute $\min_j$ and $\max_j$ for all $j = 1, 2, 3$. Then use the normalisation formula.

Exercise 3: Solution

- We first find that

$$\min_1 = -2, \qquad \min_2 = 1, \qquad \min_3 = -200$$

and

$$\max_1 = 2, \qquad \max_2 = 3, \qquad \max_3 = 300.$$

- We then normalise all variables in all vectors as follows:

$$\tilde{\mathbf{x}}^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \tilde{\mathbf{x}}^{(2)} = \begin{bmatrix} 1 \\ 0 \\ 0.2 \end{bmatrix}, \tilde{\mathbf{x}}^{(3)} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.6 \end{bmatrix}, \tilde{\mathbf{x}}^{(4)} = \begin{bmatrix} 0.75 \\ 0.5 \\ 0 \end{bmatrix}.$$

- Notice that all numerical variables are now in $[0, 1]$.

Mixed distance

- When vectors have both numerical and categorical variables, we need to use some combination of distance metrics.
- One way is to use the *mixed distance*.
- Given a number $p \geq 1$ and two vectors with d variables

$$\mathbf{x}^{(1)} = (x_1^{(1)}, \ldots, x_d^{(1)}) \quad \text{and} \quad \mathbf{x}^{(2)} = (x_1^{(2)}, \ldots, x_d^{(2)}),$$

their mixed distance is defined as

$$D^p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sqrt[p]{\sum_{j=1}^{d} |\bar{x}_j|^p}$$

where

$$\bar{x}_j = \begin{cases} normalise(x_j^{(1)}) - normalise(x_j^{(2)}), & \text{if } j \text{ is numerical} \\ \mathbf{1}(x_j^{(1)} \neq x_j^{(2)}), & \text{if } j \text{ is categorical.} \end{cases}$$

k-nearest neighbours

- *k-nearest neighbours* (k-NN) is one of the most popular classification algorithms.
- Given a labeled training set, we predict the labels of future instances depending on their distance from the labeled ones.
- k determines how many of the closest labeled instances to consider. We need to choose its value ourselves.
- Different notions of distance can be used, depending on the types of variables.
- In our examples we will use *mixed Euclidean distance*, which is mixed distance for $p = 2$.

Exercise 4

- Consider the following data set.

	gen	age	bmi	city	ill
$\mathbf{x}^{(1)}$	male	33	28.8	Bristol	no
$\mathbf{x}^{(2)}$	female	45	23.8	London	no
$\mathbf{x}^{(3)}$	female	68	21.3	Edinburgh	yes
$\mathbf{x}^{(4)}$	male	21	22.6	London	yes
$\mathbf{x}^{(5)}$	male	71	18.3	Birmingham	no
$\mathbf{x}^{(6)}$	female	27	28	Birmingham	yes
$\mathbf{x}^{(new)}$	female	26	20	Birmingham	?

- Use k-NN to find the missing value of $\mathbf{x}^{(new)}$.
- Use $k = 3$ and the mixed Euclidean distance. Use the majority vote to determine the label for $\mathbf{x}^{(new)}$.

Exercise 4: Solution

- We first normalise all numerical variables of all instances.
- The new table is the following:

	gen	age	bmi	city	ill
$\mathbf{x}^{(1)}$	male	0.24	1	Bristol	no
$\mathbf{x}^{(2)}$	female	0.48	0.524	London	no
$\mathbf{x}^{(3)}$	female	0.94	0.286	Edinburgh	yes
$\mathbf{x}^{(4)}$	male	0	0.410	London	yes
$\mathbf{x}^{(5)}$	male	1	0	Birmingham	no
$\mathbf{x}^{(6)}$	female	0.12	0.924	Birmingham	yes
$\mathbf{x}^{(new)}$	female	0.1	0.162	Birmingham	?

- We next find the squared distances of $\mathbf{x}^{(new)}$ with all other instances, for each variable separately.

Exercise 4: Solution (continued)

- The squared distances for each variable are:

	gen	age	bmi	city	$D^2(\cdot)$	$y^{(i)}$
$Dist(\mathbf{x}^{(new)},\mathbf{x}^{(1)})$	1	0.0196	0.70	1	1.65	no
$Dist(\mathbf{x}^{(new)},\mathbf{x}^{(2)})$	0	0.1444	0.13	1	1.13	no
$Dist(\mathbf{x}^{(new)},\mathbf{x}^{(3)})$	0	0.7056	0.015	1	1.31	yes
$Dist(\mathbf{x}^{(new)},\mathbf{x}^{(4)})$	1	0.01	0.06	1	1.44	yes
$Dist(\mathbf{x}^{(new)},\mathbf{x}^{(5)})$	1	0.81	0.026	0	1.36	no
$Dist(\mathbf{x}^{(new)}, \mathbf{x}^{(6)})$	0	0.0004	0.58	0	0.76	yes

- The 3-nearest neighbours of $\mathbf{x}^{(new)}$ are $\mathbf{x}^{(2)}$, $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(6)}$.
- We finally classify $\mathbf{x}^{(new)}$ using the distances we found.
- Therefore, $y^{(new)}$ is predicted to be yes.

Up next. . .

Optional Material

Weighted k-NN

- In k-NN, ties can occur that need to be broken somehow.
- One way is to use a version of k-NN called *weighted k-NN*.
- First, we convert each label as $\{\text{no}, \text{yes}\} \to \{0, 1\}$.
- Then, each point $\mathbf{x}^{(i)}$ is given a weight w_i using a function called *kernel function*. We then calculate a weighted sum:

$$S = \frac{1}{k} \sum_{i \in \mathcal{N}_k} w_i y^{(i)}$$

  where $\mathcal{N}_k$ is the set of indices of the k closest points to $\mathbf{x}^{(new)}$.
- The label $y^{(new)}$ is then predicted using the formula:

$$\hat{y}^{(new)} = \begin{cases} \text{yes, if } S > 0.5 \\ \text{no, if } S \leq 0.5. \end{cases}$$

- The most straightforward kernel function is the inverse of the mixed Euclidean distance, that is, $w_i = 1/D^2(\mathbf{x}^{(new)}, \mathbf{x}^{(i)})$.

Optional Exercise 1

- Reconsider this distance table with $y^{(i)}$ converted to $\{0, 1\}$.

	gen	age	bmi	city	$D^2(\cdot)$	$y^{(i)}$
$Dist(\mathbf{x}^{(new)}, \mathbf{x}^{(1)})$	1	0.0196	0.70	1	1.65	0
$Dist(\mathbf{x}^{(new)}, \mathbf{x}^{(2)})$	0	0.1444	0.13	1	1.13	0
$Dist(\mathbf{x}^{(new)}, \mathbf{x}^{(3)})$	0	0.7056	0.015	1	1.31	1
$Dist(\mathbf{x}^{(new)}, \mathbf{x}^{(4)})$	1	0.01	0.06	1	1.44	1
$Dist(\mathbf{x}^{(new)}, \mathbf{x}^{(5)})$	1	0.81	0.026	0	1.36	0
$Dist(\mathbf{x}^{(new)}, \mathbf{x}^{(6)})$	0	0.0004	0.58	0	0.76	1

- What will weighted k-NN predict for $y^{(new)}$ with $k = 4$ and the mixed Euclidean distance as the kernel function?
- Hint: Calculate $S = \frac{1}{k} \sum_{i \in \mathcal{N}_k} \frac{y^{(i)}}{D^2(\mathbf{x}^{(new)}, \mathbf{x}^{(i)})}$ where $k = 4$ and $\mathcal{N}_4 = \{2, 3, 5, 6\}$. Predict yes if $S > 0.5$ or no if $S \leq 0.5$.

Optional Exercise 1: Solution

- We calculate the weighted sum S as follows:

$$
\begin{aligned}
S &= \frac{1}{4} \sum_{i \in \{2,3,5,6\}} \frac{y^{(i)}}{D^2(\mathbf{x}^{(new)}, \mathbf{x}^{(i)})} \\
&= \frac{1}{4} \left(\frac{0}{1.13} + \frac{1}{1.31} + \frac{0}{1.36} + \frac{1}{0.76} \right) \\
&\approx 0.5198 > 0.5.
\end{aligned}
$$

- Since $S > 0.5$, we predict that $\hat{y}^{(new)} = 1$.
- Try using the weighted k-NN rule for $k = 2$ and $k = 6$.
- Weighted k-NN can also be used when k is odd, in place of the majority vote. Try to use it for $k = 3$ and $k = 5$.

Optional Exercise 2

Definition 1 (Distance metric)

A function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *distance metric*, if and only if, for all vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$, the following hold:

1. $f(\mathbf{x}, \mathbf{y}) = 0$, if and only if, $\mathbf{x} = \mathbf{y}$;
2. $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$; and
3. $f(\mathbf{x}, \mathbf{z}) \leq f(\mathbf{x}, \mathbf{y}) + f(\mathbf{y}, \mathbf{z})$.

- Show that the Minkowski distance, L^p, (for any $p \geq 1$) and the Hamming distance, H, are distance metrics.
- Hint: Use *Minkowski's inequality*: for all $a_1, a_2 \dots, a_d \in \mathbb{R}$ and $b_1, b_2 \dots, b_d \in \mathbb{R}$ and $p \geq 1$, we have

$$\sqrt[p]{\sum_{j=1}^{d} |a_j + b_j|^p} \leq \sqrt[p]{\sum_{j=1}^{d} |a_j|^p} + \sqrt[p]{\sum_{j=1}^{d} |b_j|^p}.$$

Optional Exercise 2: Solution

- For the Minkowski distance, let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ be arbitrary vectors with d numerical variables, and $p \geq 1$. We have
  1. If $\mathbf{x} = \mathbf{y}$, then $L^p(\mathbf{x}, \mathbf{y}) = \sqrt[p]{\sum_{j=1}^{d} |x_j - y_j|^p} = \sqrt[p]{\sum_{j=1}^{d} 0} = 0$.
     If $L^p(\mathbf{x}, \mathbf{y}) = 0$, then
     $\sqrt[p]{\sum_{j=1}^{d} |x_j - y_j|^p} = 0 \Rightarrow x_1 = y_1, \ldots, x_d = y_d \Rightarrow \mathbf{x} = \mathbf{y}$.
  2. $L^p(\mathbf{x}, \mathbf{y}) = \sqrt[p]{\sum_{j=1}^{d} |x_j - y_j|^p} = \sqrt[p]{\sum_{j=1}^{d} |y_j - x_j|^p} = L^p(\mathbf{y}, \mathbf{x})$.
  3. $L^p(\mathbf{x}, \mathbf{z}) = \sqrt[p]{\sum_{j=1}^{d} |x_j - z_j|^p} = \sqrt[p]{\sum_{j=1}^{d} |x_j - y_j + y_j - z_j|^p} \leq$
     $\sqrt[p]{\sum_{j=1}^{d} |x_j - y_j|^p} + \sqrt[p]{\sum_{j=1}^{d} |y_j - z_j|^p} = L^p(\mathbf{x}, \mathbf{y}) + L^p(\mathbf{y}, \mathbf{z})$.
- Therefore, Minkowski distance is a distance metric.

Optional Exercise 2: Solution (continued)

- For the Hamming distance, let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ be arbitrary vectors with d variables. We have
  1. If $\mathbf{x} = \mathbf{y}$, then $H(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} \mathbf{1}(x_j \neq y_j) = \sum_{j=1}^{d} 0 = 0$.
     If $H(\mathbf{x}, \mathbf{y}) = 0$, then
     $\sum_{j=1}^{d} \mathbf{1}(x_j \neq y_j) = 0 \Rightarrow x_1 = y_1, \ldots, x_d = y_d \Rightarrow \mathbf{x} = \mathbf{y}$.
  2. $H(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} \mathbf{1}(x_j \neq y_j) = \sum_{j=1}^{d} \mathbf{1}(y_j \neq x_j) = H(\mathbf{y}, \mathbf{x})$.
  3. $H(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{d} \mathbf{1}(x_j \neq z_j) \leq \sum_{j=1}^{d} (\mathbf{1}(x_j \neq y_j) + \mathbf{1}(y_j \neq z_j)) =$
     $\sum_{j=1}^{d} \mathbf{1}(x_j \neq y_j) + \sum_{j=1}^{d} \mathbf{1}(y_j \neq z_j) = H(\mathbf{x}, \mathbf{y}) + H(\mathbf{y}, \mathbf{z})$.
- Therefore, Hamming distance is a distance metric.

Q&A

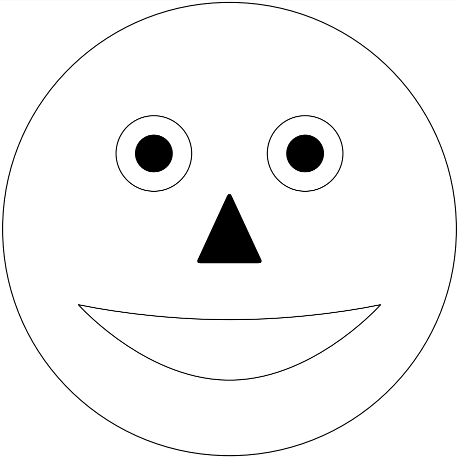
Any questions?

Some closing words. . .

—People glorify all sorts of bravery except the bravery they might show on behalf of their nearest neighbours.

George Eliot

Until the next time...



Thank you for your attention!