Pattern Recognition - Sheet 01

Tutor: Sebastian Mueller
Handed out: **Nov 02**
Discussion: **Nov 16**

Ex 1-0: Installing Python

We will be using Python for the practical exercises since it is the the de-facto standard in the open source data science community.

We recommend to use Anaconda to install Python and manage different environments:

Download

Download Anaconda for your platform from https://www.anaconda.com/products/individual and follow the installation instructions.

Set up the environment (Linux)

Either open a terminal and navigate to the root of the LecturePatternRecognition repostiory and execute

```
conda env create -f ./sheet01/patternrec.yml
conda activate patternrec
```

or execute

```
conda create -n patternrec python=3.8 pip
conda activate patternrec
pip install numpy scipy sklearn matplotlib jupyter
conda install ipykernel
```

finally run

```
ipython kernel install --user --name=patternrec_kernel
```

to make the environment available from within your jupyter installation. To check if everything is set up, run

```
jupyter notebook
```

and your browser should open http://localhost:8888/tree. On the right you can find a dropdown menu labeled "New". Click on it and check if it lists patternrec_kernel as an option.

Great! Now you are set up for the exercises.

Ex 1-1: Vocabulary

Answer the following questions in your own words:

1. What is the main difference between a classification and a regression task?
2. What is the relationship between a dataset, classes, samples and features?
3. You are given weather data that was collected by a weather station in Bonn during June and July this year. Each sample consists of 5 features. Would you be able to predict the weather for August? For January? Why (not)?

Ex 1-2: Linear Algebra recap

Given the matrix

$$A = \begin{pmatrix} -5 & 2 & 4 \\ 2 & -8 & 2 \\ 4 & 2 & -5 \end{pmatrix}$$

1. Calculate the Eigenvalues of A
2. Calculate the Eigenvector for the largest Eigenvalue of A
3. What is the rank of A?

Ex 1-3: Similarity Measures

1. State the definition of a metric
2. Order the following L_p norms by the $\leq$ relationship ($\boldsymbol{x} \in \mathbb{R}^n$)
   (a) $||\boldsymbol{x}||_1$ (b) $||\boldsymbol{x}||_2$ (c) $\sqrt{n}||\boldsymbol{x}||_2$ (d) $||\boldsymbol{x}||_\infty$ (e) $n||\boldsymbol{x}||_\infty$
3. The Kullback-Leibler divergence for discrete distributions is a popular similarity measure, it is given by:

   $$D_{\mathrm{KL}}[\mathbb{P}||\mathbb{Q}] = \sum_{x \in \mathcal{X}} \mathbb{P}(x) \log \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$$

   with $\lim_{x \to 0^+} x \log x = 0$, and is defined iff $\forall x(\mathbb{Q}(x) = 0 \Rightarrow \mathbb{P}(x) = 0)$, where $\mathbb{P}$, $\mathbb{Q}$ are probability measures over the same probability space $\mathcal{X}$.

   Why is it not a full metric?