Machine Learning and Physics
Winter Semester 2022/2023

Exercise Sheet #3

Due: 14.11.22 11h00

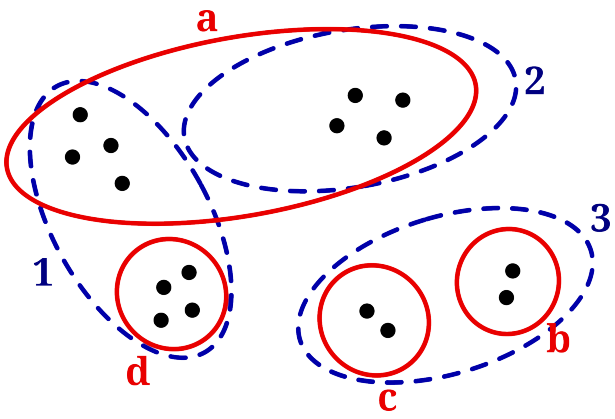
Roman Remme

General Regulations.

- Please hand in your solutions in groups of three people. A mix of attendees from Monday and Tuesday tutorials is fine.
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at https://github.com/hci-unihd/mlph_sheet03. Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter. Please hand in both the notebook (`.ipynb`), as well as an exported pdf.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of three.

1 Rand Index and Variation of Information

In this exercise, you will compute some of the measures from the lecture on a toy example to get a better feeling for them. You will use the the red clustering R (unbroken lines, cluster labels a, b, c, d) and blue clustering B (dashed lines, clusters labels 1, 2, 3) from the figure below.



(a) Create a contingency table for the two clusterings (use the cluster names in the figure to label the rows and columns). (1 pt)

(b) Count the number of pairs of points on which the clusterings agree, i.e. both place them in same or different clusters. Compute the Rand index. (1 pt)

To each clustering C, one can assign a random variable (which we also call C): It describes which cluster a point randomly drawn from the data belongs to. This allows us to define the entropy of a clustering and derived quantities, such as mutual information between clusterings.

(c) Compute the entropy of both clusterings separately. (1 pt)

(d) For the set of points from the figure above, how could you cluster them such as to achieve minimal and maximal entropy of the respective partitions. (1 pt)

(e) Compute the probability distribution of the joint clustering, and its entropy. (1 pt)

(f) Show that $H(X,Y) = H(X) + H(Y)$ if X, Y are independent and use this to make a statement about the independence of C_1 and C_2. (1 pt)

(g) What is the Mutual Information (MI) of the clusterings? (1 pt)

(h) What is the Variation Of Information (VOI) between the clusterings? (1 pt)

(i) As a metric to compare clusterings, VOI is favoured over MI. Why is that? Hint: if MI was the criterion, and if the ground truth clustering was available - how would a clustering look like that is guaranteed to maximize the MI? (1 pt)

2 Similarity Measures on Jet-Tagging Data

On the last sheet, you partitioned the jet-tagging dataset using the k-means algorithm. In this exercise, you will evaluate the quality of the clusterings using different measures.

(a) For each of the clusterings saved in the rows of `dijet_clusters.npy`, create a contingency table comparing it to the partition corresponding to the ground-truth labels. Visualize them using Hinton plots (you can use the provided method or create your own) and interpret what you see. (2 pts)

(b) Compute the Rand score, the adjusted Rand score, and the variation of information of the clusterings comparing them to the ground-truth clustering. Which clustering would you deem the best, and why? (3 pts)

3 Mutual Information for Image Matching

The two most common "weighting" schemes for used in "Magnetic Resonance Imaging" (MRI), T1 and T2, lead to different tissues being highlighted. When registering images acquired with different weighting, this has to be taken into account, as it renders simple correlation between pixel values a poor measure of how well the images match. Hence, in this exercise, you will use the mutual information score to find out which of a list of T2 proposal images best matches a T1 reference image.

(a) Load the reference image and the proposals and create a scatter plot of pixel intensities of the reference vs the proposals. Which qualitative differences do you observe? (2 pts)

(b) For each pairing of reference and proposal, compute and display a 2D histogram over the joint distribution of pixel intensities of the two images. Interpret each histogram as a contingency matrix to compute the mutual information score. Sort the proposals by this score and plot them in order. What do you observe? (4 pts)

4 Bonus: Variation of Information as a Metric

Show that the variation of information is a true metric, in particular that it obeys the triangle inequality. (5 pts)