Assignment 1

Martin Pawelczyk
martin.pawelczyk@uni-tuebingen.de
Gjergji Kasneci
DSAR
University of Tuebingen

October 2019

General instructions. Please submit your assignments in groups of 2-3 as a jupyter notebook file to ilias. In the beginning of your notebook, please state all group member names and corresponding matriculation numbers. Please name your file `a1_name1_name2_name3.ipynb`. Assignment 1 is due **November 18th, 12:00 pm.**

Task 1: PDFs and CDFs

1. Look up the t-distribution and plot the histogram of 100 observations from the t-distribution with 1 degree of freedom.

2. Compute the **mean** and **variance**. Plot the estimated PDF and its corresponding empirical CDF. Report your findings for **mean** and **variance** and comment on them, if necessary. Compare your results with those from a standard normal distribution, $\mathcal{N}(0, 1)$.

Task 2: Maximum likelihood and beyond

We suspect that our data was generated by a 1-dimensional Gaussian mixture model (GMM) with 2 different mixture components.

1. What are the parameters belonging to the Gaussian mixture?

2. Name a commonly used inference algorithm that allows you to make inference on the parameters from the GMM. Explain the algorithm in your own words. In particular, what is the **E-step** and what is the **M-step**? On a conceptual level, what is the difference between the two steps?

3. Explain the underlying assumptions used to make inference. In particular, how do they differ from classical Maximum Likelihood Estimation (MLE)? Which assumptions do both algorithms have in common?

Task 3: Useful statistical concepts

Using your own words, describe;

1. what **unbiasedness** of an estimator means?;

2. the concept of **consistency** of an estimator? How do **unbiasedness** and **consistency** differ? Can you think of a consistent, but biased estimator? Explain why your chosen estimator is consistent, but biased.

3. the **Central Limit Theorem** (CLT).

Now, we will make use of the CLT.

1. Pick a pdf that was presented in the lecture (not Gaussian/Normal or Bernoulli pdf) and conduct a simulation experiment that illustrates how the CLT works. Do your theoretical predictions differ from those of your experiment? How does this difference depend on the number of observations n?

Task 4: T-tests applied to data

Download the **wine** data set from the **UCI Machine Learning Repository**. We are interested in finding out whether alcohol levels differ from a chosen (hypothesized) value.

1. For white wine, test whether the mean alcohol level is different from 10 *percent alcohol by volume*. Clearly state your null hypothesis. State the alternative. What conclusions do you draw at the 5% significance level? At the 10 percent significance level? Repeat the same procedure for red wine.

2. Test whether the mean alcohol levels differ across wine types (red vs. white). State your null hypothesis. State the alternative. What conclusions do you draw at the 1% significance level? What conclusions do you draw at the 5% significance level? At the 10 percent significance level?

3. Finally, you are in charge of finding out a *statistically significant difference* between red and white wines (other than the mean alcohol level). You can use any feature/variable from the **wine** data set to formulate your own hypothesis test. When conducting the hypothesis test, repeat the procedure as suggested in **2.**

Task 5: Information theory (optional)

You can reach full marks without completing this task. In this task, we will investigate how certain concepts are being used at the research frontier. Using your own words, describe (2 out of 3);

1. Entropy;

2. the KL-divergence;

3. the Jensen-Shannon Divergence.

From the list of list of NeurIPS or ICML papers identify a paper that uses one of the three above concepts. Explain what it is used for: do the authors use it as regularization term to an objective function? Do they find a way to compute these quantities fast in practice? Do they have a different application? Use 2-3 sentences to describe the paper's objective and use 1-2 sentences to describe the intuition behind the used information theoretic concepts.