Exercise Sheet #2

Due: 30.10.2023 11h00

Jan Pawlowski, Tilman Plehn
Jonas Spinner

General Regulations.

- Please hand in your solutions in groups of three people. A mix of attendees from different tutorials is fine.
- Your solutions to theoretical exercises can be either handwritten notes (scanned), typeset using LaTeX, or directly in the jupyter notebook using Markdown.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at https://github.com/heidelberg-hepml/mlph2023-Exercises. Always provide the (commented) python code as well as the output, and don't forget to explain/interpret the latter, we do not give points for code that does not run. Please hand in both the notebook (`.ipynb`) and an exported pdf.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of three.

1 Bayes Theorem

Imagine you are operating an imaging atmospheric Cherenkov telescope, such as the H.E.S.S. telescope in Namibia (https://www.mpi-hd.mpg.de/hfm/HESS/). Let's say you assume a priori that 10% of the detections are gamma rays from the observation target and the rest is background (e.g. cosmic rays), i.e.

$$p(\text{gamma ray}) = 0.1 \qquad p(\text{background}) = 0.9.$$

To distinguish the gamma rays from the background, you analyze the image from the telescope to deduce the approximate direction of the original particle and compare it with the direction of your target. Assume that

$$p(\text{target direction}|\text{gamma ray}) = 0.95 \qquad p(\text{target direction}|\text{background}) = 0.1,$$

and that your algorithms tell you that the particle came from the direction of the target. Compute the posterior probability that the detection is a gamma ray from the observation target, i.e. compute $p(\text{gamma ray}|\text{target direction})$.

(2 pts)

2 Trees and Random Forests

(a) Calculate optimal splits: For the provided (`data1d.npy`, `labels1d.npy`) one-dimensional binary classification problem, consider all splits where the smallest $i = 1, \dots, N-1$ data points are grouped into one node and the remaining $N - i$ points into the other. For each of these splits, compute the Gini impurity, entropy and misclassifcation rate, and visualize the split that each of these methods would choose. (2 pts)

(b) Use the implementation of random forests in sklearn[1] to classify the jet tagging data. Perform the following steps:

i) Load the data and split it into train, validation and test set. Validation and test set should each contain $N = 200$ data points with the rest belonging to the training set.

ii) Train the following combination of parameters on the train set and evaluate the learned model on the validation set.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

- Number of trees in $\{5, 10, 20, 100\}$
- Split criterion either Gini or Entropy.
- Depth of the individual trees in $\{2, 5, 10, \text{pure}\}$[2]

iii) Finally choose your preferred set of hyperparameters and evaluate the performance on the test set.

(2 pts)

3 Fits

Johannes measured the pressure p of a gas for molar volumes V_m at temperature $T = 293\text{K}$ and saved the results in the file `gas.npy`. For the pressure measurement he assumes a gaussian uncertainty, yielding the likelihood

$$p(\text{data}|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (p_i - p(V_{m,i}, \theta))^2 \Big), \tag{1}$$

(a) Calculate the negative log-likelihood. Do you recognize this expression? (1 pts)

(b) Following his friend Benoît, Johannes first assumes the ideal gas law and wants to extract the ideal gas constant R ($\theta = R$). Use modern fitting tools (`scipy.optimize.minimize`) to help Johannes with minimizing the negative log-likelihood, and extract the profile likelihood estimator for R. Hint: Constant terms do not affect the minimization. (2 pts)

(c) Johannes comes up with an extended equation of state to improve the fit

$$p(V_m) = \frac{RT}{V_m - b} - \frac{a}{V_m^2}. \tag{2}$$

Use your estimator for R from part (b) to obtain estimates for the new parameters $\theta = \{a, b\}$ following the same procedure. Compare the resulting negative log-likelihood with your result from (b). (2 pts)

[2] where pure refers to growing each tree until each leaf is pure