Machine Learning and Physics
Winter Semester 2022/2023

Exercise Sheet #4

Due: 21.11.22 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_sheet04`. 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 Bayes: Signal or Noise?

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 Bayes Classifiers

In the lecture, we derived the Bayes classifier for the 0-1 loss. In this exercise, you will find the optimal classifier for two other loss functions.

(a) Consider binary classification with an asymmetric loss matrix:

$L(y, \hat{y})$	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	0	1
$y = 1$	10	0

Derive the optimal Bayes classifier and interpret the result. When would you use such an asymmetric loss matrix?

(4 pts)

(b) Consider classification with k classes in the ground-truth $y \in \{1, \dots, k\}$, but adding 0 as an additional "reject class" to the prediction $\hat{y} \in 0, 1, \dots, k$. For a fixed $\alpha \in (0, 1)$, consider a loss function L with

$$
\begin{aligned}
L(y, \hat{y}) &= 1 - \delta_{y\hat{y}} && \text{for} \quad y, \hat{y} = 1, \dots, k \\
L(y, 0) &= \alpha && \text{for} \quad y = 1, \dots, k\,.
\end{aligned}
$$

Derive the optimal Bayes Classifier in this setting. What is the influence of α? When would you prefer this classifier over the one discussed in the lecture? (4 pts)

3 K-Nearest Neighbors: Cross-Validation

In this exercise you will apply K-Nearest Neighbors from `scikit-learn` to the jet-tagging data and find the optimal value for k using cross-validation, for three different dataset sizes. For n-fold cross-validation, the data is first randomly partitioned into n equally sized chunks. Then, for the i-th of the n folds (i.e. training and validation runs), the i-th chunk is used for validation while the rest is used for training.

(a) Implement the splits for n-fold cross-validation. Write a function that takes a feature matrix and a labels vectors, and returns a list of training and validation sets to be used in n-fold cross validation as described above. (4 pts)

(b) Evaluate the 10-fold cross validation error for k-nn classification on the jet-tagging data for different values of k between 1 and 30. Of these, which is the optimal value for k? (3 pts)

(c) Repeat the above for the two reduced datasets containing subsets of the data (as defined in the jupyter notebook). Does the value for the optimal k change? Why is that? For each dataset, make a plot of the cross validation error (including it's standard deviation over the folds as an error bar) versus $-k$ and interpret what you observe. (3 pts)

4 Bonus: Alternative K-Nearest Neighbors

(a) Consider the following decision rule for non-parametric two-class classification:

(i) Find the k-th nearest neighbor of the query point for each of the two classes.

(ii) Predict that class whose k-th nearest neighbor is closer to the query.

What is the relation of this procedure to k-NN classification? (2 pts)

(b) Generalize the procedure to more than two classes. Is the relation to k-NN classification the same? (1 pts)