Exercise Sheet #4

Due: 18.11.24 14h00

Peter Lippmann

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group). Submissions by a single person alone will not be corrected.
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes please make sure that they are legible and not too blurry.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at `https://github.com/sciai-lab/mlph_w24`. Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter. Please hand in your notebook (`.ipynb`), as well as an exported pdf-version of it.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of two. Specify all names of your group in the submission.

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

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

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 QDA

In this exercise, we apply QDA to a one-dimensional binary classification problem.

(a) For the 1D binary classification problem, with data `data1d.npy` and `labels1d.npy`, fit a normal distribution to each class by computing the mean and standard deviation of the points belonging to it. (2 pts)

(b) QDA: In the range $[-10, 10]$, compute and plot the two Gaussian class densities as well as the posterior $p(y = 0|x)$ assuming equal prior probabilities, i.e. $p(y = 0) = p(y = 1)$. What do you observe? Then, change the prior probabilities to $p(y = 0) = 2p(y = 1)$. What has changed? (4 pts)

4 Trees and Random Forests

(a) Consider a two class classification problem ($C = 2$). At the current node there are $N = 400$ data points of each class (denoted by $(400, 400)$). Evaluate two possible splits:

- Split A: Create two nodes with $(300, 100)$ and $(100, 300)$ data points respectively.
- Split B: Create two nodes with $(200, 0)$ and $(200, 400)$ data points respectively.

Calculate the misclassification rate for each split as well as the Gini impurity and the entropy. Which split would each criterion prefer? Remember

$$\text{Gini impurity:} \quad H = 1 - \sum_{c=1}^{C} p(y = c)^2 \quad \text{and} \quad \text{Entropy:} \quad H = -\sum_{c=1}^{C} p(y = c) \log p(y = c).$$

(3 pts)

(b) 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. (For formulae, see a) and https://en.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity.) (3 pts)

(c) 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) Use the following combination of parameters on the train set and evaluate the resulting 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 in {Gini, 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.

(4 pts)

5 Bonus: The Multivariate Normal

Marginal and conditional distributions of normal distributions are again normal (as we will show below for the special case of 2D). Gaussian processes are a useful tool in statistical modeling which make use of this fact (see e.g. https://distill.pub/2019/visual-exploration-gaussian-processes/ for a gentle introduction).

Consider a two-dimensional Normal distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \\ &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}),\end{aligned}$$

formulated once with the covariance matrix $\boldsymbol{\Sigma}$ and once with the precision matrix $\boldsymbol{\Lambda}$, where $\mathbf{x} = (x_1, x_2)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$,

$$\boldsymbol{\Sigma} = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix} \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{pmatrix}\Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22}\end{pmatrix}.$$

(a) Derive that $p(x_1|x_2 = c) = \mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2})$ and give the expressions for $\mu_{1|2}$ and $\Sigma_{1|2}$. To get from $p(\mathbf{x}) = p(x_1, x_2)$ to the conditional we can just fix x_2 to the observed value c and normalize the expression. In order to do this go through the following steps:

1. Consider $p(\mathbf{x})$ and, ignoring the normalization constant, expand the square in the exponential sorting it into terms depending on x_1 and those independent of it. Do this in the form of the $\boldsymbol{\Lambda}$ instead of $\boldsymbol{\Sigma}$ for simplicity.
2. The resulting term is again quadratic, i.e. has the form of a Gaussian and you only need to find $\mu_{1|2}$ and $\Sigma_{1|2}$. Do this by comparing the form you get via 1. with the expanded exponent of a general Gaussian, comparing the relevant coefficients in each term. This allows you to write $\mu_{1|2}$ and $\Sigma_{1|2}$ in terms of $x_2, \mu_1, \mu_2, \Lambda_{11}, \Lambda_{12}$.
3. It can be shown that

$$\begin{aligned}\Lambda_{11} &= \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} \\ \Lambda_{12} &= -\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}.\end{aligned}$$

Use these results to finally formulate $\mu_{1|2}$ and $\Sigma_{1|2}$ in terms of $x_2, \mu_1, \mu_2, \Sigma_{11}, \Sigma_{12}, \Sigma_{21}$.

(3 pts)

(b) Derive $p(x_1) = \int p(x_1, x_2)\mathrm{d}x_2 = \mathcal{N}(x_1|\tilde{\mu}_1, \tilde{\Sigma}_1)$ showing that it is again a normal distribution, and give the expressions for $\tilde{\mu}_1, \tilde{\Sigma}_1$. In order to do this go through the following steps:

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

1. As in **i)** just focus on the quadratic in the exponential ignoring the normalization for now and work with the precision matrix. Expand it collecting all the terms depending on x_2 and form a new quadratic from them. Then, integrate out x_2 analytically.
2. Reorder the remaining terms in the exponential to get the expressions for $\tilde{\mu}_1, \tilde{\Sigma}_1$ in terms of $\mu_1, \Lambda_{11}, \Lambda_{12}, \Lambda_{21}, \Lambda_{22}$.
3. Using the result that
$$\Sigma_{11} = \left(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\right),$$
simplify your expression further.

(3 pts)