Machine Learning and Physics
Winter Semester 2025/26

Exercise Sheet 4

Due: 17.11.25 14h00

Christof Gehrig, Gerrit Gerhartz

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group).
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes, please ensure they are legible and not blurry.
- For the practical exercises, always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter.
- Please hand in a **single PDF** that includes both the exported notebook and your solutions to the theoretical exercises. Submit the PDF to the Übungsgruppenverwaltung once per group, making sure to include the names of both group members in the submission.
- You can find all the data in the GitHub Repository.

1 Posterior form in GDA

In Gaussian discriminant analysis, each class-conditional density is modeled as

$$p(\mathbf{x} \mid C = k) = \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k), \quad k \in \{0, 1\}$$

with class priors $\pi_0, \pi_1 > 0$ and $\pi_0 + \pi_1 = 1$.

(a) Using Bayes' rule, derive the expression for the posterior probability $p(C = 1 \mid \mathbf{x})$ and show that it can be written in the form

$$p(C = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-g(\mathbf{x})}}$$

for some function $g(\mathbf{x})$ that depends on the Gaussian parameters. (3 pts)

(b) Determine the functional form of $g(\mathbf{x})$ and show how it simplifies under the assumptions of equal covariances ($\Sigma_0 = \Sigma_1$) and unequal covariances ($\Sigma_0 \neq \Sigma_1$). (3 pts)

(c) Comment on whether $p(C = 1 \mid \mathbf{x})$ is itself Gaussian in $\mathbf{x}$ in these cases, and justify your answer. (2 pts)

2 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

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

(4 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 misclassification 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.) (4 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.

- 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)

3 Bonus: Underdetermined linear regression

Consider an underdetermined linear regression problem. Given training data $\{(\mathbf{x}_i, y_i)\}$, we want to fit a model $y_i = [1, \mathbf{x}_i^T]\mathbf{w}$. The perfect solution space is defined as the set of weight vectors for which the resulting mean squared error $\mathcal{E}$ is zero, i.e.,

$$\mathcal{S} = \{\mathbf{w} \in \mathbb{R}^p \mid \mathcal{E}(\mathbf{w}) = 0\}.$$

Assume that the observations $\{\mathbf{x}_i\}$ are in general position.

(a) Describe the nature of the perfect solution space $\mathcal{S}$ (i.e., its geometry and dimension) for $n = 5$ and $p \in \{4, 5, 6, 7\}$. (2 pts)

(b) A family of solutions can be found by

(i) randomly initializing the weights according to a multivariate Gaussian distribution with mean 0 and covariance matrix $\mathbf{I}$.

(ii) performing non-stochastic gradient descent ($\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{E}(\mathbf{w}_t)$, learning rate $\eta \in \mathbb{R}^+$) on the mean squared error criterion.

If this were repeated multiple times, prove that the set of solutions found follows a Gaussian distribution on the solution space. (3 pts)

(c) Characterize the mean of that distribution (hint: You can argue in many ways, including geometrically or using the Moore-Penrose pseudoinverse). (2 pts)

(d) Characterize the covariance of that distribution. (2 pts)

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

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

① $p(\vec{x}\,|\,C=k) = \mathcal{N}(\vec{x}\,|\,\mu_k, \Sigma_k)$, $k \in \{0,1\}$

class priors $\pi_0, \pi_1 > 0$, $\pi_0 + \pi_1 = 1$

use multivariate normal distribution !

$\sqrt{(2\pi)^d|\Sigma|}$ ← det

a) $$p(C=1\,|\,\vec{x}) = \frac{p(\vec{x}\,|\,C=1)\,p(C=1)}{p(\vec{x})} = \frac{\mathcal{N}(\vec{x}\,|\,\vec{\mu}_1, \Sigma_1)\,\pi_1}{\pi_0\mathcal{N}(\vec{x}\,|\,\vec{\mu}_0, \Sigma_0) + \pi_1\mathcal{N}(\vec{x}\,|\,\vec{\mu}_1, \Sigma_1)}$$

$$= \frac{1}{\frac{\pi_0\mathcal{N}(\vec{x}\,|\,\vec{\mu}_0, \Sigma_0)}{\pi_1\mathcal{N}(\vec{x}\,|\,\vec{\mu}_1, \Sigma_1)} + 1} \checkmark \quad = \left[1 + \frac{\pi_0}{\pi_1}\frac{\sqrt{2\pi\Sigma_0^{\cancel{2}}}}{\sqrt{2\pi\Sigma_1^{\cancel{2}}}}\,e^{-\frac{1}{2}\left(\frac{\vec{x}-\vec{\mu}_0}{\Sigma_0}\right)^2 + \frac{1}{2}\left(\frac{\vec{x}-\vec{\mu}_1}{\Sigma_1}\right)^2}\right]^{-1}$$

$$= \left\{1 + \frac{\pi_0}{\pi_1}\frac{\Sigma_0}{\Sigma_1}\exp\left[\frac{1}{2}\left(\left(\frac{\vec{x}-\vec{\mu}_1}{\Sigma_1}\right)^2 - \left(\frac{\vec{x}-\vec{\mu}_0}{\Sigma_0}\right)^2\right)\right]\right\}^{-1}$$

$$= \left\{1 + \exp\left[\ln\frac{\pi_0}{\pi_1} + \ln\frac{\Sigma_0}{\Sigma_1} + \frac{1}{2}\left(\left(\frac{\vec{x}-\vec{\mu}_1}{\Sigma_1}\right)^2 - \left(\frac{\vec{x}-\vec{\mu}_0}{\Sigma_0}\right)^2\right)\right]\right\}^{-1}$$

b) $$g(x) = \frac{1}{2}\left[(\vec{x}-\vec{\mu}_0)^T\Sigma_0^{-1}(\vec{x}-\vec{\mu}_0) - (\vec{x}-\vec{\mu}_1)^T\Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)\right] - \ln\frac{\pi_0}{\pi_1} \overset{\frac{1}{2}}{-} \ln\frac{\Sigma_0}{\Sigma_1}$$

<u>Case 1: $\Sigma_0 = \Sigma_1$:</u>

$$g(x) = \frac{1}{2}(\vec{x}-\vec{\mu}_0)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_0) - \frac{1}{2}(\vec{x}-\vec{\mu}_1)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_1) - \ln\frac{\pi_0}{\pi_1} - \cancel{\ln\frac{\Sigma}{\Sigma}}$$

$$= \frac{1}{2}\left(\cancel{\vec{x}^T\Sigma^{-1}}\vec{x} - \vec{x}^T\Sigma^{-1}\vec{\mu}_0 - \vec{\mu}_0^T\Sigma^{-1}\vec{x} + \vec{\mu}_0^T\Sigma^{-1}\vec{\mu}_0\right)$$

$$- \frac{1}{2}\left(\cancel{\vec{x}^T\Sigma^{-1}}\vec{x} - \vec{x}^T\Sigma^{-1}\vec{\mu}_1 - \vec{\mu}_1^T\Sigma^{-1}\vec{x} + \vec{\mu}_1^T\Sigma^{-1}\vec{\mu}_1\right) - \ln\frac{\pi_0}{\pi_1}$$

$$= \frac{1}{2}\left[\vec{x}^T\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0) + (\vec{\mu}_1 - \vec{\mu}_0)^T\Sigma^{-1}\vec{x} + \vec{\mu}_0^T\Sigma^{-1}\vec{\mu}_0 - \vec{\mu}_1^T\Sigma^{-1}\vec{\mu}_1\right] - \ln\frac{\pi_0}{\pi_1}$$

Due to symmetry of covaicnce matrices (in general):

$$= x^T\Sigma^{-1}(\vec{\mu}_1 - \mu_0) + \frac{1}{2}\left(\vec{\mu}_0^T\Sigma^{-1}\vec{\mu}_0 - \vec{\mu}_1^T\Sigma^{-1}\vec{\mu}_1\right) - \ln\frac{\pi_0}{\pi_1} \checkmark$$

→ Linear in $\vec{x}$

<u>Case 2: $\Sigma_0 \neq \Sigma_1$:</u>

$$g(x) = \frac{1}{2}\left(\vec{x}^T\Sigma_0^{-1}\vec{x} - \vec{x}^T\Sigma_0^{-1}\vec{\mu}_0 - \vec{\mu}_0^T\Sigma_0^{-1}\vec{x} + \vec{\mu}_0^T\Sigma_0^{-1}\vec{\mu}_0\right)$$

$$- \frac{1}{2}\left(\vec{x}^T\Sigma_1^{-1}\vec{x} - \vec{x}^T\Sigma_1^{-1}\vec{\mu}_1 - \vec{\mu}_1^T\Sigma_1^{-1}\vec{x} + \vec{\mu}_1^T\Sigma_1^{-1}\vec{\mu}_1\right) - \ln\frac{\pi_0}{\pi_1} - \ln\frac{\Sigma_0}{\Sigma_1}$$

$$= \frac{1}{2}\left(\vec{x}^T\Sigma_0^{-1}\vec{x} - 2\vec{x}^T\Sigma_0^{-1}\vec{\mu}_0 + \vec{\mu}_0^T\Sigma_0^{-1}\vec{\mu}_0 - \vec{x}^T\Sigma_1^{-1}\vec{x} + 2\vec{x}^T\Sigma_1^{-1}\vec{\mu}_1 - \vec{\mu}_1^T\Sigma_1^{-1}\vec{\mu}_1\right)$$

$$- \ln\frac{\pi_0}{\pi_1} - \ln\frac{\Sigma_0}{\Sigma_1}$$

$$= \frac{1}{2}\left(\vec{x}^T\Sigma_0^{-1}\vec{x} + \vec{\mu}_0^T\Sigma_0^{-1}\vec{\mu}_0 - \vec{x}^T\Sigma_1^{-1}\vec{x} - \vec{\mu}_1^T\Sigma_1^{-1}\vec{\mu}_1\right) - \vec{x}^T\Sigma_0^{-1}\vec{\mu}_0 + \vec{x}^T\Sigma_1^{-1}\vec{\mu}_1$$

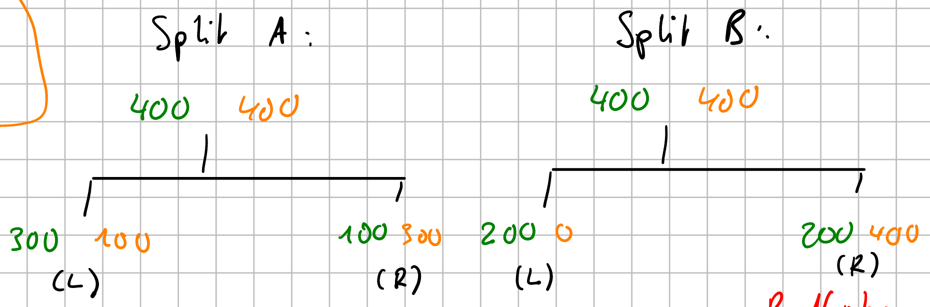
$$- \ln\frac{\pi_0}{\pi_1} - \ln\frac{\Sigma_0}{\Sigma_1}$$

→ Quadratic in $\vec{x}$

c) In both cases the posterior is a logistic sigmoid, not a Gaussian. A Gaussian has a different shape, not $\frac{1}{1+e^{-s(x)}}$.

② a)

Careful, wrong numerical results!!!



R: Number of points in set

Split A:

Misclassification rate: $1-\max_c \hat{\pi}_{ic} = 1-\max_c \frac{1}{|R_i|}\sum_{n\in R_i} I(y_n=c)$

i: node

c: class

In this case:

$$1-\max_c \hat{\pi}_{ic} = 1-$$

$$= 1-\left(\frac{1}{400}\frac{300}{400}+\frac{1}{400}\frac{300}{400}\right) = 1-\frac{3}{1600} = \frac{1597}{1600}$$

GINI impurity:

$$G_L = 1-\sum_{c=1}^{C}\hat{\pi}_{Lc}^2 = 1-\left(\hat{\pi}_{L1}^2+\hat{\pi}_{L2}^2\right) = 1-\left[\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right] = 1-\frac{10}{16} = \frac{3}{8}$$

$$G_R = G_L = \frac{3}{8}$$

$$G_{tot} = \frac{1}{2}G_R+\frac{1}{2}G_L = \frac{3}{8} = 0.375$$

Entropy:

$$H_L = -\sum_{c=1}^{C}\hat{\pi}_{Lc}\log\hat{\pi}_{Lc} = -\left(\hat{\pi}_{L1}\log\hat{\pi}_{L1}+\hat{\pi}_{L2}\log\hat{\pi}_{L2}\right) = -\frac{1}{4}\left(3\log\frac{3}{4}+\log\frac{1}{4}\right)$$

$$= -\frac{1}{4}(3\log 3-3\log 4+\underbrace{\log 1}_{=0}-\log 4$$

$$= -\frac{3}{4}\log 3+\log 4$$

$$H_R = H_L$$

$$H_{tot} = \log 4-\frac{3}{4}\log 3 \simeq 0.56$$

Split B:

Misclassification rate:

$$1-\max_c \hat{\pi}_{ic} = 1-\left(\frac{1}{200}\cdot 1+\frac{1}{600}\frac{4}{6}\right) = 1-\left(\frac{1}{200}+\frac{2}{1800}\right)$$

$$= 1-\left(\frac{11}{1800}\right) = \frac{1789}{1800}$$

GINI impurity :

$$G_L = 1 - \sum_{c=1}^{C} \hat{\pi}_{Lc}^{2} = 1 - (1^2 + 0^2) = 0$$

$$G_R = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 1 - \frac{5}{9} = \frac{4}{9}$$

$$G_{tot} = \frac{200}{800} \cdot 0 + \frac{600}{800} \cdot \frac{4}{9} = \frac{3}{4} \cdot \frac{4}{9} = \frac{1}{3}$$

Entropy :

$$H_L = - \sum_{c=1}^{C} \hat{\pi}_{Lc} \log \hat{\pi}_{Lc} = -(1 \log 1 + \underbrace{0 \cdot \log 0}) = 1$$

Use L'HOPITAL's rule for

$$\lim_{\hat{\pi} \to 0^+} \frac{\log \hat{\pi}}{\frac{1}{\hat{\pi}}} = \lim_{\hat{\pi} \to 0} \frac{\frac{1}{\hat{\pi}}}{-\frac{1}{\hat{\pi}^2}} = -\hat{\pi} = 0$$

$$H_R = -\frac{1}{3}\left(\log \frac{1}{3} + 2 \log \frac{2}{3}\right)$$

$$= -\frac{1}{3}\left(\underbrace{\log 1}_{=0} - 3 \log 3 + 2 \log 2\right)$$

$$= \log 3 - \frac{2}{3} \log 2 \approx 0.64$$

$$H_{tot} = \frac{200}{800} \cdot 1 + \frac{600}{800} \cdot 0.64 = \frac{1}{4} + \frac{2}{3} \cdot 0.64 \approx 0.67$$

① a)