Sheet #3 – Solution
Due:

Fred Hamprecht
Manuel Haußmann

1 Trees and Random Forests (10 pt)

(a) Calculating impurities (4pt). 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

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

Solution:

Missclassification rate The impurity of our node is $H = 0.5$. Split A gives us $H(L) = 0.25 = H(R)$ and the possible reduction in impurity is given as

$$\text{Split A:} \qquad H - H(L)\frac{\#L}{\#L + \#R} - H(R)\frac{\#R}{\#L + \#R} = 0.25.$$

For split B, we have that $H(L) = 0$ and $H(R) = \frac{1}{3}$. The possible reduction in impurity is then

$$\text{Split B:} \qquad 0.5 - \frac{600}{800}H(R) = 0.5 - \frac{3}{4} \cdot \frac{1}{3} = 0.25,$$

i.e. the misclassification rate criterion does not care which split we pick.

Gini impurity First note that for a two-class classification problem, we can simplify the formula using the shorthand $p = p(y = 1)$

$$1 - \sum_{c=1}^{C} p(y = c)^2 = 1 - p^2 - (1 - p)^2 = 2p(1 - p).$$

In this case we have that the impurity of our node is again given as $H = 0.5$. However considering split A we now get that $p = 1/4$ and

$$H(L) = 2 \cdot \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{8} = H(R),$$

by symmetry. The overall reduction in impurity would then be given by

$$\text{Split A:} \qquad \frac{1}{2} - \frac{1}{2} \cdot \frac{3}{8} - \frac{1}{2} \cdot \frac{3}{8} = \frac{1}{8} = 0.125.$$

For split B we have $H(L) = 0$ and $H(R) = \frac{4}{9}$, giving us overall

$$\text{Split B:} \qquad \frac{1}{2} - \frac{3}{4} \cdot \frac{4}{9} = \frac{1}{6} \approx 0.167.$$

That is gini suggests to use split B as it results in a greater reduction in impurity.

Entropy Entropy gives us $H = -\log(0.5) \approx 0.693$.[1] We have for split A that $H(L) = H(R) \approx 0.562$. And overall

$$\text{Split A:} \qquad -\log(0.5) - \frac{1}{2}H(L) - \frac{1}{2}H(R) \approx 0.131.$$

Split B gives us $H(L) = 0$ as it is pure and together with $H(R) \approx 0.637$

$$\text{Split B:} \qquad -\log(0.5) - \frac{3}{4}H(R) \approx 0.216.$$

Again the split giving the pure node is favored.

(b) Applying a Random Forest(6pt). In practice you will often rely on already existing and optimized implementations for many algorithms. As discussed in the lecture the random forest is one of the best "off-the-shelf" classifiers we have. To get used to using existing models you will use the sklearn random forest implementation.[2] The goal is to learn how to classify digits, for which we rely on an existing data set provided by sklearn.[3] Perform the following steps:

i) Load the data set as follows

```
from sklearn.datasets import load_digits
digits, labels = load_digits(return_X_y=True)
```

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.

- Nr of trees in $\{5, 10, 20, 100\}$
- Split criterion either Gini or Entropy.
- Depth of the individual trees in $\{2, 5, 10, \text{pure}\}$[4]

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

Solution: See the jupyer notebook.

2 Bayes: Is it raining? (5 pt)

Let's say you assume a priori that it rains 20% of the days in Heidelberg, i.e.

$$p(\text{rainy}) = 0.2 \qquad p(\text{sunny}) = 0.8.$$

You have been inside all day working diligently on your exercise sheets without looking outside. Looking up you observe that a lot of your fellow students are wearing raincoats. You assume that

$$p(\text{raincoat}|\text{rainy}) = 0.95 \qquad p(\text{raincoat}|\text{sunny}) = 0.1.$$

Compute the posterior probability that it is rainy given this observation, i.e. compute $p(\text{rainy}|\text{raincoat})$.

[1] Note that we are using the logarithm to the basis e. Another popular choice is to use $\log_2$, which gives different numbers, but the same decision. The latter measures *bits*, the former *nats*.

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

[3] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html

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

Solution: Let's use the following abbreviations $r = \text{rainy}, s = \text{sunny}, c = \text{raincoat}$. This gives us

$$p(r|c) = \frac{p(c|r)p(r)}{p(c)} = \frac{p(c|r)p(r)}{p(c|r)p(r) + p(c|s)p(s)} \approx 0.7.$$

Importantly, note that $p(c|r) \neq p(r|c)$!

3 QDA & LDA (10 pt)

(a) QDA: Implementation and visualization of the posterior (5pt). Assume you are applying a QDA and have learned the mean and standard deviation in a one dimensional two-class problem. For each of the following two pairs of Normal distributions, plot the likelihoods on the range $[-7, 7]$ as well as the posterior $p(y = 2|x)$ assuming equal prior probabilities, i.e. $p(y = 1) = p(y = 2)$.

i) $p(x|y = 1) = \mathcal{N}(x| - 1, 1^2)$ and $p(x|y = 2) = \mathcal{N}(x|1, 1^2)$,

ii) $p(x|y = 1) = \mathcal{N}(x| - 1, 1.5^2)$ and $p(x|y = 2) = \mathcal{N}(x|1, 1^2)$.

What do you observe?

Solution: See jupyter notebook.

(b) Generalization to LDA (5pt). In the lecture we saw that assuming we can approximate the likelihood for each class with a multivariate Gaussian with separate $\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$ for each class, we get a decision boundary that is quadratic in $\mathbf{x}$. Assume that we are still in a two-class classification setting, but have even less data available. A further simplification is to then assume that the covariance matrix between the two classes is shared, i.e. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Derive the posterior decision boundary where

$$p(y = 1|\mathbf{x}) = p(y = 2|\mathbf{x})$$

analogously to the lecture and show that we end up with a linear decision boundary.

Solution: We follow the QDA approach given in the lecture, where we now have

$$\begin{aligned} p(y = 1|\mathbf{x}) &= p(y = 2|\mathbf{x}) \\ \Leftrightarrow \frac{p(\mathbf{x}|y = 1)p(y = 1)}{p(\mathbf{x})} &= \frac{p(\mathbf{x}|y = 1)p(y = 2)}{p(\mathbf{x})}. \end{aligned}$$

Taking logarithms and dropping terms independent of $\mathbf{x}$ writing $\stackrel{c}{=}$[5] for equality up to a constant, we get

$$\begin{aligned} 0 &\stackrel{c}{=} \log p(\mathbf{x}|y = 1) - \log p(\mathbf{x}|y = 2) \\ &\stackrel{c}{=} -\frac{1}{2}\left((\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right) \\ &= -\frac{1}{2}\left(\mathbf{x}^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \mathbf{x}^T \boldsymbol{\Sigma}^{-1}\mathbf{x} + 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) \\ &\stackrel{c}{=} -\frac{1}{2}\left(-2\mathbf{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \end{aligned}$$

[5] My notation, and not standardized one. Usually you only have $a \propto b$ to mean equality up to a multiplicative constant.

4 The Multivariate Normal (technical +10pt)

In the lecture, we stated that the marginal and the conditional distributions of a multivariate Normal distribution are again Normal. In this exercise, you will show this.

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

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

Note that while $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ it is not the case that $\Lambda_{11} = \Sigma_{11}^{-1}$.

i) Conditional distribution. 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_a 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 $x_2, \mu_1, \mu_2, \Sigma_{11}, \Sigma_{12}, \Sigma_{21}$.

ii) Marginal distribution. Derive $p(x_1) = \int p(x_1, x_2)dx_2 = \mathcal{N}(x_1|\tilde{\mu}_1, \tilde{\Sigma}_1)$ 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:

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 form which, having the form of Gaussian exponential, can then be integrated 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.

Solution: Short answer: Have a look at Bishop, *Pattern Recognition and Machine Learning* (p. 85-89) for a very nice, detailed derivation and discussion. Here we will only look at a very rough sketch of the essential ideas.

i) Conditional distribution. Expanding the exponential we have[6]

$$\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) = &-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\boldsymbol{\Lambda}_{11}(\mathbf{x}_1-\boldsymbol{\mu}_1)\\
&-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\boldsymbol{\Lambda}_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2)\\
&-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\boldsymbol{\Lambda}_{21}(\mathbf{x}_1-\boldsymbol{\mu}_1)\\
&-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\boldsymbol{\Lambda}_{22}(\mathbf{x}_2-\boldsymbol{\mu}_2),
\end{aligned} \tag{1}$$

i.e. an exponential that is quadratic in $\mathbf{x}_1$, hence a Normal distribution.[7] In general we have

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const.} \tag{2}$$

This pattern appears again and again. Now we only need to expand the terms in (1) and compare them with the corresponding terms in (2) and get the forms for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. E.g. we have one term quadratic in $\mathbf{x}_1$, giving us $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Lambda}_{11}^{-1}$. Analogously we get

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\Sigma}_{1|2}\big(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2)\big).$$

Using the expressions for the precision subsets given on the exercise sheet one can finally simplify to

$$\begin{aligned}
\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\\
\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.
\end{aligned}$$

ii) Marginal distribution. Starting again from (1), in order to marginalize over $\mathbf{x}_2$, we this time collect all the terms relevant to $\mathbf{x}_2$ and as a second step add the necessary terms to *complete the square*, i.e.

$$-\frac{1}{2}\mathbf{x}_2^T\boldsymbol{\Lambda}_{22}\mathbf{x}_2 + \mathbf{x}_2^T\mathbf{m} = -\frac{1}{2}\left(\mathbf{x}_2 - \boldsymbol{\Lambda}_{22}^{-1}\mathbf{m}\right)^T\boldsymbol{\Lambda}_{22}\left(\mathbf{x}_2 - \boldsymbol{\Lambda}_{22}^{-1}\mathbf{m}\right) + \underbrace{\frac{1}{2}\mathbf{m}^T\boldsymbol{\Lambda}_{22}^{-1}\mathbf{m}}_{\text{added}},$$

for suitable $\mathbf{m}$ similar to above. This allows us to analytically integrate over the first term. Combining the second term with the remaining terms from (1), rearranging with respect to $\mathbf{x}_1$, and again comparing with (2), gives us the expressions for $\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1$. These can then be further simplified and we end up with the satisfying

$$\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 \qquad \text{and} \qquad \tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11}.$$

[6] I give the general multivariate approach here. In the exercise it was fine if you stayed in the 2d case.
[7] Note that I will sometimes refer to it as “Normal” and sometimes as “Gaussian”. These terms are interchangeable.