Machine Learning
Winter Semester 2019/2020

# Exercise Sheet #3

Due: 07/11/2019 – 16:00

Fred Hamprecht
Manuel Haußmann

---

**General Regulations.**

- You should hand in your solutions in groups of at least two people (recommended are three).
- The theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX.
- Practical exercises should be implemented in python and can be submitted either as scripts (`.py`) or jupyter notebooks (`.ipynb`). In both cases always provide both the (commented) code as well as the output, and don't forget to explain/interpret the latter.
- Submit all your files in a single `.zip` archive to `mlhd1920@gmail.com` using the following standardized format: The subject line should consist of the full names of all team members as well as the exercise, and the title of the zip archive the last names. I.e. assuming your group consists of Ada Lovelace, Geoffrey Hinton and Michael Jordan, this means

  Subject: [EX03] Michael Jordan, Geoffrey Hinton, Ada Lovelace
  Zip Archive: `ex03-jordan-hinton-lovelace.zip`
- *****NEWS*****: Remember that there won't be a lecture next Monday, 4.11. We will continue on Wednesday, 6.11.

## 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

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

**(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.[1] The goal is to learn how to classify digits, for which we rely on an existing data set provided by sklearn.[2] 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)
```

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

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 $\{2, 5, 10, \text{pure}\}$[3]

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

## 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 $p(\text{rainy}|\text{raincoat})$.

## 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?

**(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.

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

# 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 $\mathbf{x} = (x_1, x_2)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$,

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

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 $\mu_1, \Sigma_1$. 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.