# Exercise 2 – Salami Quality

Mathematical Modelling
02526

The goal with this exercise is to use mathematical modelling to solve a practical problem: Quality assessment of salami. The exercise is structured to cover central elements of mathematical modelling:

- Identify significant issues
- Formulate a mathematical model, which solves these issues
- Implement the model in a computer program
- Use the implemented model for analysis of data
- Report the results from the analysis and discuss the validity of the model

The report for this exercise must be 5 pages at most, including graphs, tables, and images (excluding frontpage and appendices). In this exercise a number of questions are asked. The report, however, should not be a list of answers, but instead be a coherent documentation and discussion of the analysis performed. The details for this are described in the section *Reporting*.

The exercise consists of two parts. The first is an introduction to linear discriminant analysis, first in one dimension and then extended to higher dimensionality. This part is necessary to solve the second part, which considers quality assessment of salami based on multi-spectral images. It is only the second part that should be included in the report.

To solve the exercise data and code in MATLAB and PYTHON are placed on DTU Learn. It up to you which programming language you use. You are given the following:

- Six colour images `{color_day01.png, ..., color_day28.png}`
- Six multi-spectral images `{multispectral_day01.mat, ..., multispectral_day28.mat}`
- Images with annotations of fat and meat `{annotation_day01.png, ..., annotation_day28.png}`
- MATLAB functions for reading the data `loadMulti.m`, `getPix.m` for extractions of pixels, `setImagePix.m` for making RGD-images, and `showHistograms.m` to generate histograms. Place the files in your working directory and type `help "file name"` to see how the functions are used – e.g. `help loadMulti`.
- PYTHON functions ...

## 1 Linear Discriminant Analysis (Part 1)

**1.** Based on height measurements (in cm) of 489, 17 years old american men and 467, 17 years old american women we have calculated the following model

$$x \in \mathcal{N}(\mu, 6.7^2) \tag{1}$$

where

$$\mu = \begin{cases} 175.5 & \text{for men} \\ 162.9 & \text{for women} \end{cases} \tag{2}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal (Gaussian) distribution with mean value $\mu$ And variance $\sigma^2$.

**2.** Plot in a coordinate system the probability density function (PDF)

$$f(x|\mu) = \frac{1}{6.7\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{1}{6.7^2}(x-\mu)^2\right) \tag{3}$$

for the two values of $\mu$ given above.

**3.** We are now interested in finding the value of $x$ where

$$f(x|175.5) \geq f(x|162.9) \tag{4}$$

or equivalent

$$\frac{f(x|175.5)}{f(x|162.9)} \geq 1\ . \tag{5}$$

Find the solution graphical and by evaluation of the inequality.

**4.** Assume that we know the height of a person without knowing the sex of the person. For which values of height would you guess that the person is a man?

**5.** If we have the independent stochastic variables $x_1$ og $x_2$, which are normally distributed

$$\begin{aligned} x_1 &\in \mathcal{N}(0, 2^2) && (6)\\ x_2 &\in \mathcal{N}(1, 3^2) && (7)\\ & && (8) \end{aligned}$$

the joint PDF is given by the product of the marginal PDFs

$$f(x_1, x_2) = \frac{1}{2\pi}\frac{1}{2\cdot 3}\exp\left(-\frac{1}{2}\left[\frac{1}{4}x_1^2 + \frac{1}{9}(x_2-1)^2\right]\right)\ . \tag{9}$$

Plot this function.

**6.** If $x_1$ and $x_2$ is as given above, but correlated with the correlation coeffient

$$\rho = \frac{\text{Cov}(x_1, x_2)}{\sqrt{\text{Var}(x_1)\text{Var}(x_2)}} \tag{10}$$

the joint PDF becomes

$$g(x_1, x_2) = \frac{1}{2\pi}\frac{1}{6}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{1}{1-\rho^2}\left[\frac{1}{4}x_1^2 - 2\rho\frac{1}{6}x_1(x_2-1) + \frac{1}{9}(x_2-1)^2\right]\right)\ . \tag{11}$$

Plot this function for $\rho = \frac{2}{3}$.

**7.** We now introduce

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{12}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{13}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 4 & 4 \\ 4 & 9 \end{pmatrix}. \tag{14}$$

Write the expression for $g(x_1, x_2)$ as a function of $\mathbf{x}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$!

## 2 Introduction to sausages and salami (Part 2)

Sausages of good quality are important for the nourishment of the Danish population (not to mention the German) and many other nations with a love of sausages. If the sausage quality was too poor many people would risk mycotoxins from fungi, and bacterial infections. Mycotoxins are toxins produced by moulds and shows in the form of moulden sausages. Bacteria could for example be *Listeria monocytogenes* which causes listeriosis, killing 20-30 % of patients, or *Clostridium botulinum* which causes botulism (aptly named 'sausage poisoning' in Danish *pølseforgiftning* and German *Wurstvergiftung*). The toxin produced is named Botulinum toxin, and is one of the most lethal poisons in the world with a $\mathrm{LD}_{50}$ of 1.3-2.1 ng/kg. The toxin blocks the signal transmission from nerves to muscles and leads to paralysis and death. This demonstrates the importance of ensuring sausage quality!

Salami is produced by mixing minced meat with spices and salt and stuffing it into the casing which is either intestine or artificial. The salami is then dried for many days hereby changing character – image examples are given in figure 1. The last step is cold smoking for conservation and flavouring. During the whole process the meat is not heated, so micro organisms grows and ferment the meat which lowers the pH. The conservation is partly through this and partly through the smoking. Due to the lack of heat treatment it is important that the product is not contaminated by dangerous bacteria or fungi. Even under clean production circumstances the production can lead to salami of varying quality. In this exercise we will focus on some of these quality parameters. Important parameters are the meat/fat ratio and the dessication over time. These parameters can be measured multi spectrally which we will use in this exercise.

The purpose of this exercise is to develop a technique for automatic determination of the amount of meat and fat in the sausage. The sensor for this task is a camera and controlled lightning. The task is to

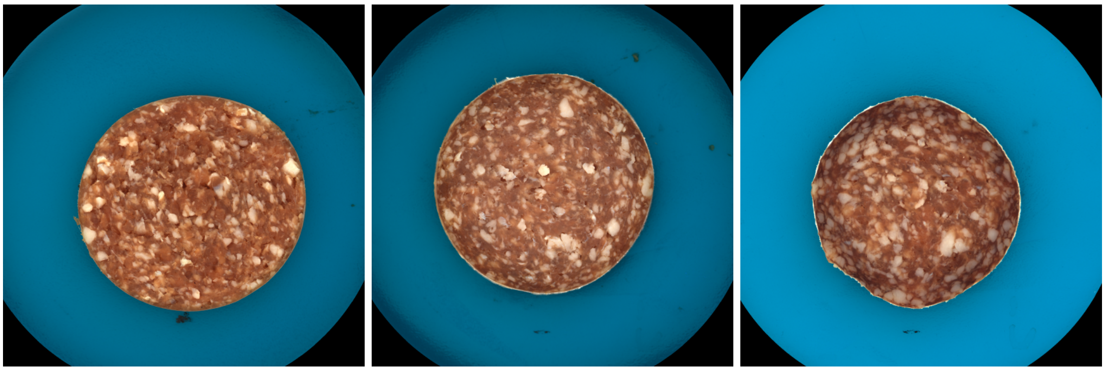

Figure 1: Salami imaged day 1, 6 and 28 respectively. The image colours are reconstructed from multi spectral images.

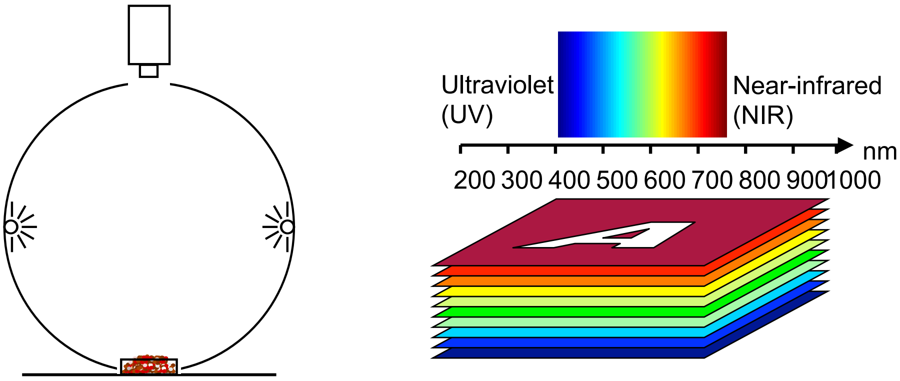

Figure 2: Sketch showing the principals of VideometerLab (left) and spectral images (right). VideometerLab is a sphere lit by coloured (specific wavelength) LEDs placed around the equator of the sphere. The resulting multi spectral image is a stack of images, where each layer is of a specific wavelength. The wavelengths range from blue to infra-red light.

develop a mathematical model which takes these measurements as input and gives a measure of the fat and meat as output.

We use a multi spectral computer vision system named VideometerLab [1] – illustrated in figure 2. With VideometerLab a multi spectral image is recorded, which means an image where each layer is an image recorded in a specific wavelength. The sphere surrounding the sample ensures diffuse even lighting of the sample, thus avoiding reflections from the surface. Each pixel in the image contains spectral information. The images of the salami is recorded at 19 different wavelengths, giving 19 layers in the image and 19 measurements for each pixel.

**Describtion of the problem:**

1. Look at the images and consider how to distinguish between fat and meat.
2. Which visual changes are apparent in the salami over the 28 days?
3. How can these changes be measured from the images?
4. Load a multi spectral image in MATLAB or PYTHON and investigate the data. As mentioned code is available on DTU Inside.

**Mathematical model** Each pixel contains spectral information regarding the sample. Depending on the sample's chemical and physical characteristics different amounts of light are absorbed and scattered in the sample. Meat appears red because the blue and yellow wavelengths are absorbed, while the white fat only absorbs a little light from all wavelengths. If we consider each pixel as an independent multivariate stochastic variable, we can formulate a statistical model for classification of meat and fat.

First we will introduce some notation for images. A grey scale image is given by $I_g \subset \Omega_g \in \mathbb{R}^2$, where the image $I_g$ is a subset of the plane $\Omega_g$. Since an image is discretely sampled – that means only measured on specific discrete points on the plane – we get the pixel value $x = I_g(\mathbf{c})$, where $\mathbf{c} = (r, c)^T$ is a coordiante in the plane. The multi spectral image $I \subset \Omega \in \mathbb{R}^3$ consists of layers of grey scale images in the volume $\Omega$. A pixel value is here given by $x = I(\mathbf{c}, l)$, where $l = 1, \ldots, n$ are the layers of grey scale images. We denote the vector of pixels over all layers at the same spatial coordiante by $\mathbf{x} = I(\mathbf{c})$. It is the intensity of the pixels, that we use to solve the discrimination problem.

A simple model for classification of a continuous variable into two classes is by using a threshold. The classification model

[1] www.videometer.com

$$\tau(x) = \begin{cases} C_1 & \text{if} \quad x \geq t \\ C_2 & \text{if} \quad x < t \end{cases} \tag{15}$$

classifies the pixel $x$ to one of two classes $C_1$ og $C_2$ based on the threshold value $t$. This can done for a single spectral band and would be very fast to calculate. The threshold value can be determined from the intensity distribution in the two classes. Typically this is unknown, but a reasonable assumption is that they are normally distributed $\mathcal{N}_i(\mu_i, \sigma^2)$, $i \in \{1, 2\}$, where $\mu_i$ is the mean value for class $i$ and $\sigma^2$ is the common variance. The threshold can then be determined as that intensity where the probability for belonging to the two classes is equal, and which is placed between the two mean values.

As $\mu$ and $\sigma$ are unknown,we have to estimate them from a training set chosen from the data. Estimated values are denoted $\hat{\mu}$ and $\hat{\sigma}$ and are calculated by

$$\hat{\mu} = \frac{1}{m}\sum_{j=1}^{m} x_j \ , \tag{16}$$

$$\hat{\sigma} = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(x_j - \hat{\mu})^2} \ , \tag{17}$$

where $m$ is the number of observations of $x$.

If a single spectral band is used, the discriminative value will be limited, and we must expect larger discriminative power by using all bands. We still want to classify to the most probable class under the assumption that our observations belong to one of two normally distributed populations. This time the normal distribution is multivariate with $\mathcal{N}_i(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, $i \in \{1, 2\}$. Notice, that the observations are now vectors $\mathbf{x} \in \mathbb{R}^n$. From the normal disribution model the values of the PDF $f_i$ of observation $\mathbf{x}$ is estimated as

$$f_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det \hat{\mathbf{\Sigma}}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^T \hat{\mathbf{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)\right) \ , \tag{18}$$

where $n$ is the dimensionality of the observation. Notice, that it is assumed that the covariance $\hat{\mathbf{\Sigma}}$ is equal for the two classes. For one class the covariance is calculated as

$$\hat{\mathbf{\Sigma}}(a, b) = \frac{1}{m-1}\sum_{j=1}^{m}(x_{aj} - \hat{\mu}_a)(x_{bj} - \hat{\mu}_b) \ , \tag{19}$$

where $\hat{\mathbf{\Sigma}}(a, b)$ is the covariance for dimension $a$ and $b$. $x_a$ and $x_b$ is the pixel value in dimension $a$ and $b$ respectively and $m$ is the number of observations. The calculation for the pooled covariance matrix becomes

$$\hat{\mathbf{\Sigma}} = \frac{1}{\sum_{i=1}^{k}(m_i - 1)}\sum_{i=1}^{k}(m_i - 1)\hat{\mathbf{\Sigma}}_i \ , \tag{20}$$

where $k$ is the number of classes and $\hat{\mathbf{\Sigma}}_i$ is the covariance matrix for each class. $m_i$ is the number of observations in $i$.

The observation can be classified as the most probable class, and a decision criterion is then

$$\tau(\mathbf{x}) = \begin{cases} C_1 & \text{if} \quad f_1 \geq f_2 \\ C_2 & \text{if} \quad f_2 > f_1 \end{cases} . \tag{21}$$

If we already know before hand the probability of the classes, we can use this to enhance the classification model. The PDF is here the probability that a randomly chosen pixel belongs to a certain class, e.g. fat. This is called the *prior* distribution. By use of Bayes theorem we get

$$g_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det \hat{\mathbf{\Sigma}}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^T \hat{\mathbf{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)\right) p_i \ , \tag{22}$$

where $p_i$ is the prior probability. If we take the logarithm of $g_i$ and remove common elements from the two classes we get the discriminant function

$$S_i(\mathbf{x}) = \mathbf{x}^T \hat{\mathbf{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_i - \frac{1}{2} \hat{\boldsymbol{\mu}}_i{}^T \hat{\mathbf{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_i + \ln p_i \ . \tag{23}$$

Again we classify to the most probable class with the decision criterion

$$\tau(\mathbf{x}) = \begin{cases} C_1 & \text{if} \quad S_1 \geq S_2 \\ C_2 & \text{if} \quad S_2 > S_1 \end{cases} . \tag{24}$$

This model can be used to discriminate between fat and meat in the salami, and to classify the change that occurs due to dessication of the meat.

**Formulation of the mathematical model:**

1. Describe the mathematical model
2. Why do we expect it to work?
3. What are the limitations of the model?

**Data** The original data set consists of 64 multi spectral images of salami. Each multi spectral image is composed of 19 individual images recorded at a specific wavelength. The individual images in the composed image are spectral bands. The salamis have been imaged over 28 days of drying. At day 1, 2, 3, 6, 13, 20, and 28 images have been taken. The dataset is made from a batch of salamis in production. On each trial day, nine salamis is taken from the batch and a piece cut from each and imaged. No salami has thus been recorded more than once, but we have 9 repetitions from each trial day. The data is illustrated in figure 3.

Only five of the days and only a single image from each have been selected for this exercise: day 1, 6, 13, 20, and 28. The number of images has been reduced to shorten the computation time. Further, the resolution of the images have been downsampled to $1/4$ of the original side length and has a spatial resolution of $514 \times 514 \times 19$. The grey scale resolution has been reduced from 16 bit to 8 bit. In this way the size of the image files have been reduced from roughly 300 Mb to around 1.5 Mb.

**Analysis of data**

1. Investigate the data. A manual segmentation has been made of fat and meat in a single image for each day – have a look at day 1. (On DTU Learn code is available for loading data, making histograms, etc.).
2. What is the spectral distribution of meat and fat respectively?
3. Is it a reasonable assumption that each pixel is either meat or fat?

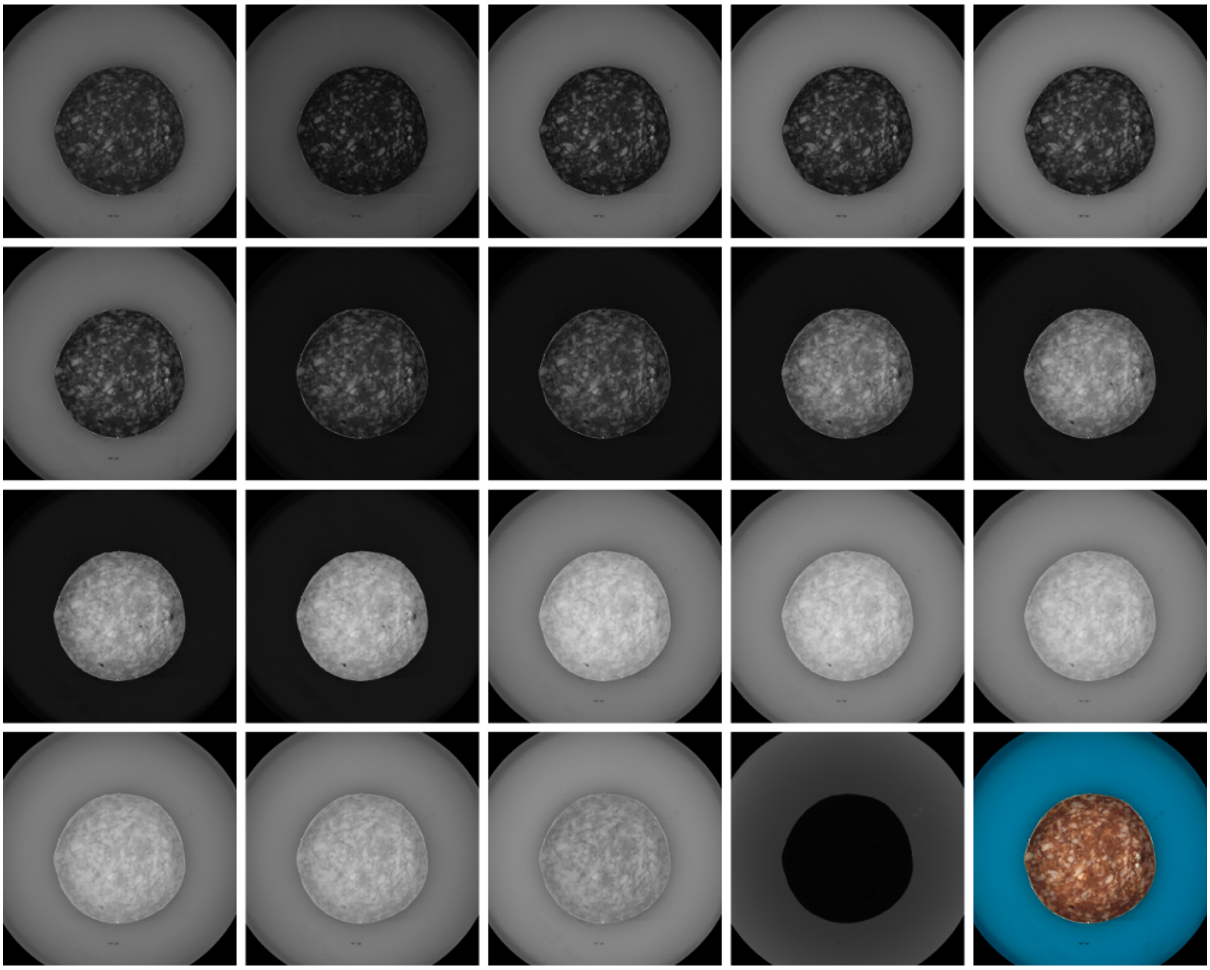

Figure 3: Multi spectral image of salami after 20 days of drying – each image is a layer in the multi spectral image. The wavelengths are row wise from left to right: 410 nm, 438 nm, 450 nm, 468 nm, 502 nm, 519 nm, 572 nm, 591 nm, 625 nm, 639 nm, 653 nm, 695 nm, 835 nm, 863 nm, 880 nm, 913 nm, 929 nm, 940 nm og 955 nm. Last image is a RGB image made by a weighted average of the images in the visible spectrum, which means the first 12 spectral bands

**Experiments** The salami data set makes it possible to investigate the fat and meat distribution over time. The determination of meat and fat is a classification problem and the outlined classification model is well suited for the problem.

We also see a visual change in the colour of the salami over time. This is partly due to drying and partly due to physical and chemical changes. This is a gradual change and the salami develops a radial gradient with a darker colour around the edges than in the middle. These changes makes the fat and meat classification more difficult and we want to quantify the error by classifying manually annotated images from the different days.

We need an automatic method for determination of the amount of fat and meat in the salami. The method must be based on the outlined classification model. The model must be implemented as computer program and trained on annotated data for day 1.

First the simplest model should be implemented. That is, to find a threshold value for a single spectral band. As there is 19 spectral bands, we will need to choose the best one, the band that gives the best classification. We can investigate which spectral band gives the best classification from the training data. The simplest model assumes that the measured pixel values for meat and fat are normally distributed and have the same variance. We seek a threshold value $t$, that optimally separates the salami in meat with the mean value $\mu_{\mathrm{m}}$ and fat with the mean value $\mu_{\mathrm{f}}$. What is the threshold value for this simple model and for what spectral band?

**Threshold value for a single spectral band**

1. Calculate the threshold value $t$ for all spectral bands for day 1.
2. Calculate the error rate for each spectral band.
3. Identify the spectral band, which has the best discriminative properties for meat and fat.
4. Classify the entire image of the salami for day 1, and visualise it.

A more advanced model takes all spectral bands into consideration at the same time. Here we need to perform the multi spectral discriminant analysis described in (23) and perform the classification as described in (24). Again, we assume the same variance for the two classes, but as each pixel contains 19 values from the 19 spectral bands we use the covariance matrix as described in (19) and (20).

**Classification by means of all spectral bands**

1. Calculate the multivariate linear discriminant function as described in (23) for day 1.
2. Calculate the error rate (disagreement between the model and the annotations) for the training set.
3. Classify the entire image of the salami for day 1 and visualise it.

We have now made two models for automatic determination of the fat and meat content in salami. To test these models me must apply them to new data. That is data the model has not been trained on (used for the estimation of the parameters in the model). Since the training was on day 1, we can use the data from the remaining days, i.e. 6, 13, 20, and 28, calculate the error rate without chaning the model parameters, and thus check the validity of the model.

**Calculation for all days:**

1. Classify fat and meat for the remaining days with the models trained on day 1.
2. Calculate the error rate for the annotated areas for the remaining days. How is the performance of the two models?
3. Classify again the entire images for the remaining days and, and visualise them. Judged on the visualisations, which model performs best?

It is not certain that day 1 is the best day to train on, as the salami dries over time. Perhaps it would be better to choose another day later in the drying process and train on that instead.

**Training on each day and error calculated on all the other days:**

1. Judging from the images, which day would you choose to train on and why??
2. Train the linear discriminant function on day 1, day 6 and so on, and for each day calculate the error rate on all the other days.
3. Show the error rate for all days (5 training days x 4 test days = 20 error rates) in an appropriate plot or table.
4. Why do we exclude the day we have trained on from the comparison?
5. Which day is the best to train the model on, and why?
6. What are the error sources in the model? Can we trust that the calculated fat and meat content in the salami is the same as the real content?

The salami producer has informed you that there are 30 % percent fat in this batch of salamis.

**Prior knowledge:**

1. Incorporate this information into your linear discriminant model
2. Does it change your estimates?

**Reporting** Reporting has the goal to convey the performed analysis such that the conclusion is clear and well founded. That is, the report must contain a short description of the problem and its relevance. The experimental work must be described. That includes a description of the data, the models used, how the model parameters have been found or estimated, and how they influence the model. The relevance of the model to the problem at hand must be discussed - in this section limitations of the model should be described. Finally, the conclusion of the analysis and how it contributes to either a solution or a better understanding of the problem.

This frame for reporting of a scientific analysis – here based on mathematical modelling – is very broad. It is however important to carefully consider which elements to include, and choose those salient for the conclusion. Both, so the report is sufficient to argue for all claims made, but not more than what is necessary to make the argument short and convincingly. This is certainly not easy, but it is essential to make the analysis useful: It is first when the analysis can be read, understood, and used by others that the analysis becomes valuable.

In this exercise the report has been limited to five pages. The length of the report is in itself not that important. It is the contents of the report that matters. It is unfortunately often seen, that a report contains a lot of results and partial results, where insufficient effort has been made in considering whether they actually matter to the final conclusions. Such reports are poor. The message is lost, and it difficult to judge the validity of any claims made. The other extreme is also bad. In very short reports important arguments or results will have to be omitted, leading to an insufficient and thus poor report.

The good report is achieved by iterating the report several times: It will never be perfect on the first try. It is important to critically consider all claims made. It is a good idea to ask one self, 'why' it is so, each time a result is reported. If the claims stands up to your own criticism, it is vital to check that the claim is supported in the text, either through experimentation, logical arguments or reference. By working the report over with these goals in mind, the result will typically be a shorter, clearer, and at least more precise report than the first iteration. Be careful about notation, and make sure that all symbols are defined and used consistently.

**Contents of the report**

1. Describe the problem and the background – what is modelled and why?
2. Describe data and experiments
3. Describe the mathematical model - how and why?
4. Describe results
5. Discuss results – how good are the results? To what degree do they reflect the true problem?
6. Conclude – what is the contribution of the analysis?

**Hand in** The report is uploaded to peergrade through DTU LEar. The code used (`*.m`-files `*.py`-files) must also be uploaded. The code must *also* be included in an appendix in the report.

Anders Nymark Christensen, Anders Bjorholm Dahl & Knut Conradsen
Revised 2020, ANYM