Exercise 1: Estimating the Bayes Error

The Bayes decision rule for the two classes classification problem results in the Bayes error

$$P(error) = \int P(error|\boldsymbol{x})p(\boldsymbol{x})dx \tag{1}$$

where

$$P(error|\boldsymbol{x}) = \min\left[P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x})\right] \tag{2}$$

is the probability of error for a particular input $\boldsymbol{x}$. Interestingly, while class posteriors $P(\omega_1|\boldsymbol{x})$ and $P(\omega_2|\boldsymbol{x})$ can often be expressed analytically and are integrable, the error function has discontinuities that prevent its analytical integration, and therefore, direct computation of the Bayes error. Section 2.8 of Duda&Hart explains how considering an upper-bound to the Bayes error instead of the true error overcomes this difficulty.

a)

Show that the full error can be upper-bounded as follows:

$$P(error) \leq \int \frac{2}{\frac{1}{P(\omega_1|\boldsymbol{x})} + \frac{1}{P(\omega_2|\boldsymbol{x})}} p(\boldsymbol{x})d\boldsymbol{x} \tag{3}$$

Note that the integrand is now continuous and corresponds to the harmonic mean of class posteriors weighted by $p(\boldsymbol{x})$.

First we show, the upper-bound holds:

$$\begin{aligned}\frac{2}{\frac{1}{P(\omega_1|\boldsymbol{x})} + \frac{1}{P(\omega_2|\boldsymbol{x})}} &= 2\frac{P(\omega_1|\boldsymbol{x})P(\omega_2|\boldsymbol{x})}{P(\omega_1|\boldsymbol{x}) + P(\omega_2|\boldsymbol{x})} = 2\frac{P(\omega_1|\boldsymbol{x})P(\omega_2|\boldsymbol{x})}{1} \\ &= 2P(\omega_1|\boldsymbol{x})(1 - P(\omega_1|\boldsymbol{x})) = 2P(\omega_2|\boldsymbol{x})(1 - P(\omega_2|\boldsymbol{x})) \\ &\geq \min\left[P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x})\right],\end{aligned}$$

which leads to

$$P(error|\boldsymbol{x}) = \min\left[P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x})\right] \leq \frac{2}{\frac{1}{P(\omega_1|\boldsymbol{x})} + \frac{1}{P(\omega_2|\boldsymbol{x})}}.$$

We can multiply with the denominator to get

$$\left(\frac{1}{P(\omega_1|\boldsymbol{x})} + \frac{1}{P(\omega_2|\boldsymbol{x})}\right) \cdot \min\left[P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x})\right] = \min\left[1 + \frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})}, \frac{P(\omega_2|\boldsymbol{x})}{P(\omega_1|\boldsymbol{x})} + 1\right] \leq 2$$

we know that $P(\omega_1|\boldsymbol{x}) = 1 - P(\omega_2|\boldsymbol{x})$, since we decide for class 2, if we not decide for class 1. We can have a look to some examples in table 1.

For the minimum part holds $\min[...] \in [1, 2]$ which means that $\min[...] \leq 2$. The minimum is exactly 1 if

Tabelle 1: Some examples, remind that $P(\omega_1|\boldsymbol{x}) = 1 - P(\omega_2|\boldsymbol{x})$

$P(\omega_1\|\boldsymbol{x})$	$P(\omega_2\|\boldsymbol{x})$	$\frac{P(\omega_1\|\boldsymbol{x})}{P(\omega_2\|\boldsymbol{x})}$	$\frac{P(\omega_2\|\boldsymbol{x})}{P(\omega_1\|\boldsymbol{x})}$	$\min[1+\bullet, 1+\bullet]$
0.1	0.9	0.11	9.0	1.11
0.3	0.7	0.43	2.33	1.43
0.5	0.5	1.0	1.0	2.0
0.7	0.3	2.33	0.43	1.43
0.9	0.1	9.0	0.11	1.11

$P(\omega_1|\boldsymbol{x}) = 0$ or $P(\omega_2|\boldsymbol{x}) = 0$ and it is exactly 2 if $P(\omega_1|\boldsymbol{x}) = P(\omega_2|\boldsymbol{x})$.

Since $p(\boldsymbol{x}) \geq 0$, $P(error|\boldsymbol{x}) \geq 0$ and $\frac{2}{\frac{1}{P(\omega_1|\boldsymbol{x})}+\frac{1}{P(\omega_2|\boldsymbol{x})}} \geq 0$, the statements also holds for the integral:

$$P(error) \leq \int \frac{2}{\frac{1}{P(\omega_1|\boldsymbol{x})} + \frac{1}{P(\omega_2|\boldsymbol{x})}} p(\boldsymbol{x}) d\boldsymbol{x}$$

b)

Show using this result that for the univariate probability distributions

$$p(x|\omega_1) = \frac{\pi^{-1}}{1+(x-\mu)^2} \qquad \text{and} \qquad p(x|\omega_2) = \frac{\pi^{-1}}{1+(x+\mu)^2}$$

the Bayes error can be upper-bounded by:

$$P(error|\boldsymbol{x}) \leq \frac{2P(\omega_1)P(\omega_2)}{\sqrt{P(\omega_1)^2 + (4\mu^2+2)P(\omega_1)P(\omega_2) + P(\omega_2)^2}}$$

First we define some equations for later use:

Bayes formula:

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)} \tag{4}$$

Probabilities sum to 1:

$$P(\omega_1) + P(\omega_2) = 1 \tag{5}$$

Solution for integral (hint):

$$\int \frac{1}{ax^2+bx+c}\, dx = \frac{2\pi}{\sqrt{4ac-b^2}} \quad \text{if} \quad b^2 < 4ac \tag{6}$$

Now start with the upper bound from (a):

$$
\begin{aligned}
P(error) &= \int P(error|x)p(x)\,dx \\
&\overset{(a)}{\leq} \int \frac{2}{\frac{1}{P(\omega_1|x)} + \frac{1}{P(\omega_2|x)}} p(x)\,dx \overset{(4)}{=} \int \frac{2}{\frac{p(x)}{p(x|\omega_1)P(\omega_1)} + \frac{p(x)}{p(x|\omega_2)P(\omega_2)}} p(x)\,dx \\
&= \int \frac{2}{\frac{1}{p(x|\omega_1)P(\omega_1)} + \frac{1}{p(x|\omega_2)P(\omega_2)}}\,dx \overset{\text{Def.}}{=} \int \frac{2}{\frac{1+(x-\mu)^2}{\pi^{-1}P(\omega_1)} + \frac{1+(x+\mu)^2}{\pi^{-1}P(\omega_2)}}\,dx \\
&= \int \frac{2}{\frac{P(\omega_2)\pi(1+(x-\mu)^2)+P(\omega_1)\pi(1+(x+\mu)^2)}{P(\omega_1)P(\omega_2)}}\,dx \\
&= \frac{2P(\omega_1)P(\omega_2)}{\pi} \int \frac{1}{P(\omega_2)(1+x^2-2x\mu+\mu^2)+P(\omega_1)(1+x^2+2x\mu+\mu^2)}\,dx \\
&= \frac{2P(\omega_1)P(\omega_2)}{\pi} \int \frac{1}{P(\omega_1)+P(\omega_2)+x^2(P(\omega_1)+P(\omega_2))+2x\mu(P(\omega_1)-P(\omega_2))+\mu^2(P(\omega_1)+P(\omega_2))}\,dx \\
&\overset{(5)}{=} \frac{2P(\omega_1)P(\omega_2)}{\pi} \int \frac{1}{x^2+2x\mu(P(\omega_1)-P(\omega_2))+1+\mu^2}\,dx
\end{aligned}
$$

We can now use (6) and we can identify $a = 1$, $b = 2\mu(P(\omega_1) - P(\omega_2))$, $c = 1 + \mu^2$. Furthermore we can show that the condition $b^2 < 4ac$ holds. With $P(\omega_1) - P(\omega_2) \leq 1$ and $\mu \geq 0$, we can show:

$$
\begin{aligned}
4\mu^2(P(\omega_1) - P(\omega_2))^2 &< 4(1+\mu^2) \\
\mu^2 &< 1 + \mu^2
\end{aligned}
$$

Using (6) we can go on with:

$$
\begin{aligned}
P(error) &\leq \frac{2P(\omega_1)P(\omega_2)}{\pi} \frac{2\pi}{\sqrt{4(1+\mu^2) - 4\mu^2[P(\omega_1) - P(\omega_2)]^2}} \\
&= \frac{2P(\omega_1)P(\omega_2)}{\sqrt{(1+\mu^2)\cdot 1 - \mu^2[P(\omega_1) - P(\omega_2)]^2}} \\
&\overset{(5)}{=} \frac{2P(\omega_1)P(\omega_2)}{\sqrt{(1+\mu^2)[P(\omega_1) + P(\omega_2)]^2 - \mu^2[P(\omega_1) - P(\omega_2)]^2}} \\
&= \frac{2P(\omega_1)P(\omega_2)}{\sqrt{(1+\mu^2)[P(\omega_1)^2 + 2P(\omega_1)P(\omega_2) + P(\omega_2)^2] - \mu^2[P(\omega_1)^2 - 2P(\omega_1)P(\omega_2) + P(\omega_2)^2]}} \\
&= \frac{2P(\omega_1)P(\omega_2)}{\sqrt{P(\omega_1)^2 + (4\mu^2+2)P(\omega_1)P(\omega_2) + P(\omega_2)^2}}
\end{aligned}
$$

We can interpret the result as follows: μ gives us the distance between the data distributions $p(x|\omega_1)$ and $p(x|\omega_2)$. If μ is increased (more separated distributions), the denominator is increasing and the error decreases.

c)

Explain how you would estimate the error if there was no upper-bounds that are both tight and analytically integrable.

Discuss following two cases: (1) the data is low-dimensional and (2) the data is high-dimensional.

In a two-category classification problem it is possible to use an analytical approximation of an upper bound.

If we consider equation (2) to start with, first we can use the following upper bound for the minimum:

$$\min[a, b] \leq a^{\beta} b^{1-\beta} \quad \text{for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1$$

The prove for this inequality is simply shown: Without loss of generality we can assume that $a \geq b$ and we have $b \leq a^{\beta} b^{1-\beta} = \left(\frac{a}{b}\right)^2 b$. Since $\left(\frac{a}{b}\right)^2 \geq 1$ the inequality is proven.

Starting from (1) using (2), we can estimate:

$$\begin{aligned}
P(error) &= \int \min\left[P(\omega_1|x), P(\omega_2|x)\right] p(x)\, dx \\
&\overset{(4)}{=} \int \min\left[\frac{p(x|\omega_1)P(\omega_1)}{p(x)}, \frac{p(x|\omega_2)P(\omega_2)}{p(x)}\right] p(x)\, dx \\
&\leq \int \left(\frac{p(x|\omega_1)P(\omega_1)}{p(x)}\right)^{\beta} \left(\frac{p(x|\omega_2)P(\omega_2)}{p(x)}\right)^{1-\beta} p(x)\, dx \\
&= P(\omega_1)^{\beta} P(\omega_2)^{1-\beta} \int p(x|\omega_1)^{\beta} p(x|\omega_2)^{1-\beta} p(x)^{-\beta-1+\beta+1}\, dx \\
&= P(\omega_1)^{\beta} P(\omega_2)^{1-\beta} \int p(x|\omega_1)^{\beta} p(x|\omega_2)^{1-\beta}\, dx
\end{aligned}$$

Under the assumption of a *normal distribution*, we can identify two methods:

- **Chernoff Bound**: Minimizing the estimation with respect to β, where $0 \leq \beta \leq 1$. The advantage is a tighter bound, while on the other side the computational effort is higher.
- **Bhattacharyya Bound**: Minimizing the estimation with respect to a fixed $\beta = 0.5$. The bound become looser for $\beta \to 0$ and $\beta \to 1$, so fixing at 0.5 is not perfectly tight (in best case as tight as Chernoff Bound), as the best β depends on the parameters of the distribution and the prior probabilities, but is accelerating the computation.

In conclusion Chernoff Bound would be a good solution for low-dimensional data, since computation effort should be affordable. Bhattacharyya Bound should be a good solution for high-dimensional data, since we can save computation effort and have a solution which is probably tight enough for a usable solution.

Exercise 2: Bayes Decision Boundaries

One might speculate that, in some cases, the generated data $p(x|\omega_1)$ and $p(x|\omega_1)$ is of no use to improve the accuracy of a classifier, in which case one should only rely on prior class probabilities $P(\omega_1)$ and $P(\omega_2)$.

For the first part of this exercise, we assume that the data for each class is generated by the univariate Laplacian probability distributions:

$$p(x|\omega_1) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right) \quad \text{and} \quad p(x|\omega_2) = \frac{1}{2\sigma} \exp\left(-\frac{|x+\mu|}{\sigma}\right)$$

where $\sigma > 0$.

a)

Compute the Bayes optimal decision boundary (i.e. points $x \in \mathbb{R}$ where $P(\omega_1|x) = P(\omega_2|x)$).

$$
\begin{aligned}
& P(\omega_1|x) = P(\omega_2|x) \\
\overset{(4)}{\Rightarrow} \quad & \frac{p(x|\omega_1)P(\omega_1)}{p(x)} = \frac{p(x|\omega_2)P(\omega_2)}{p(x)} \\
\Rightarrow \quad & p(x|\omega_1)P(\omega_1) = p(x|\omega_2)P(\omega_2) \\
\overset{\text{Def.}}{\Rightarrow} \quad & \frac{1}{2\sigma}\exp\left(-\frac{|x-\mu|}{\sigma}\right)P(\omega_1) = \frac{1}{2\sigma}\exp\left(-\frac{|x+\mu|}{\sigma}\right)P(\omega_2) \\
\Rightarrow \quad & \frac{P(\omega_1)}{P(\omega_2)} = \exp\left(-\frac{|x+\mu|}{\sigma} + \frac{|x-\mu|}{\sigma}\right) \quad \text{for } P(\omega_1), P(\omega_2) > 0 \\
\Rightarrow \quad & \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = -\frac{|x+\mu|}{\sigma} + \frac{|x-\mu|}{\sigma}
\end{aligned}
$$

For the Bayes optimal decision boundary $D \subset \mathbb{R}$, we get:

$$D = \left\{ x \in \mathbb{R} \,\middle|\, -|x+\mu| + |x-\mu| = \sigma \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) \right\}$$

We now consider that $P(w_1) = 0$ or $P(w2) = 0$. In such case, we know that $p\,(x|w_1) \geq 0$, $p\,(x|w_2) \geq 0$ and $P(w_1) + P(w_2) = 1$. Hence, $P(\omega_1|x) = P(\omega_2|x)$ cannot be satisfied for all $x \in \mathbb{R}$ and it renders the optimal decision boundary D empty for either of the cases.

b)

Determine for which values of $P(\omega_1)$, $P(\omega_2)$, μ, σ the optimal decision is to always predict the first class (i.e. under which conditions $P(error|x) = P(\omega_2|x)\ \forall\ x \in \mathbb{R}$).

Firs we assume that $\mu > 0$ and $\sigma > 0$. Now starting from (2), we have that:

$$P(error|x) = \min\left[P(\omega_1|x), P(\omega_2|x)\right] = P(\omega_2|x) \quad \Leftrightarrow \quad P(\omega_1|x) > P(\omega_2|x) \quad \forall\ x \in \mathbb{R}$$

Doing the calculation from (a), the inequality will result in

$$P(\omega_1|x) > P(\omega_2|x) \quad \Rightarrow \quad \sigma \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > -|x+\mu| + |x-\mu|,$$

which mean, that in this case $P(\omega_1|x) > P(\omega_2|x)$ for all x. We now consider 4 cases and check in which case this condition holds.

Case 1: $x < -\mu$

$$\sigma \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > -|-\mu+\mu| + |-\mu-\mu| = 2\mu$$

$\Rightarrow$ Always predicts second class

Case 2: $-\mu < x < 0$

$$\sigma \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > -|x+(-x)| + |x-(-x)| = 2x$$

$\Rightarrow$ Both classes are possibly predicted

Case 3: $0 < x < \mu$

$$\sigma \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > -|x+x| + |x-x| = -2x$$

$\Rightarrow$ Both classes are possibly predicted

Case 4: $x > \mu$

$$\sigma \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) < -|\mu+\mu| + |\mu-\mu| = -2\mu$$

$\Rightarrow$ Always predicts first class

c)

Repeat the exercise for the case where the data for each class is generated by the univariate Gaussian probability distributions:

$$p(x|\omega_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{and} \quad p(x|\omega_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$.

First we calculate the Bayes optimal decision boundary:

$$
\begin{aligned}
& P(\omega_1|x) = P(\omega_2|x) \\
\overset{(4)}{\Rightarrow} \quad & \frac{p(x|\omega_1)P(\omega_1)}{p(x)} = \frac{p(x|\omega_2)P(\omega_2)}{p(x)} \\
\Rightarrow \quad & p(x|\omega_1)P(\omega_1) = p(x|\omega_2)P(\omega_2) \\
\overset{\text{Def.}}{\Rightarrow} \quad & \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) P(\omega_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right) P(\omega_2) \\
\Rightarrow \quad & \frac{P(\omega_1)}{P(\omega_2)} = \exp\left(-\frac{(x+\mu)^2}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for } P(\omega_1), P(\omega_2) > 0 \\
\Rightarrow \quad & \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = -\frac{(x+\mu)^2}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^2} \\
\Rightarrow \quad & 2\sigma^2 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = -(x^2 + 2x\mu + \mu^2) + (x^2 - 2x\mu + \mu^2) = -4x\mu
\end{aligned}
$$

For the Bayes optimal decision boundary $D \subset \mathbb{R}$, we get:

$$
D = \left\{ x \in \mathbb{R} \,\middle|\, -2x\mu = \sigma^2 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) \right\}
$$

We can regard a special case: If $\mu = 0$, the decision boundary is $D = \mathbb{R}$, if $P(\omega_1) = P(\omega_2)$.

In the next step, we want to find out when the optimal decision is to always predict the first class ω_1. Analog to the calculation above, using Bayes formular, we get:

$$
P(\omega_1|x) > P(\omega_2|x) \quad \Rightarrow \quad \sigma^2 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > -2x\mu
$$

We now consider 2 cases (case 1 and 4 from a) and check in which case this condition holds.

Case 1: $x < -\mu$

$$
\sigma^2 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > -2(-\mu)\mu = 2\mu^2
$$

$\Rightarrow$ Always predicts second class

Case 2: $x > \mu$

$$
\sigma^2 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) < -2\mu\mu = -2\mu^2
$$

$\Rightarrow$ Always predicts first class

Exercise 1: Maximum-Likelihood Estimation

We consider the problem of estimating using the maximum-likelihood approach the parameters $\lambda, \eta > 0$ of the probability distribution:

$$p(x, y) = \lambda\eta e^{-\lambda x - \eta y}$$

supported on $\mathbb{R}_+^2$. We consider a dataset $\mathcal{D} = ((x_1, y_1), ..., (x_N, y_N))$ composed of N independent draws from this distribution.

a)

Show that x and y are independent.

A family of random variables $(X_i)_{i \in I}$ are independent if and only if they have a joint density $p_I(x_1, ..., x_n)$ for that holds:

$$p_I(x_1, ..., x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

where $p_{X_i}(x_i)$ are the marginal densities of X_i.

First we calculate the marginal densities:

$$p(x) = \int_{\mathbb{R}_+} p(x, y)\, \lambda(dy) = \int_0^\infty \lambda\eta e^{-\lambda x - \eta y}\, dy = \lambda\eta \left[-\frac{1}{\eta} e^{-\lambda x - \eta y} \right]_0^\infty = -\lambda(0 - e^{-\lambda x}) = \lambda e^{-\lambda x}$$

$$p(y) = \int_{\mathbb{R}_+} p(x, y)\, \lambda(dx) = \int_0^\infty \lambda\eta e^{-\lambda x - \eta y}\, dx = \lambda\eta \left[-\frac{1}{\lambda} e^{-\lambda x - \eta y} \right]_0^\infty = -\eta(0 - e^{-\eta y}) = \eta e^{-\eta y}$$

Using the marginal densities, we can show that the product results in the joint density

$$p(x) \cdot p(y) = \lambda e^{-\lambda x} \cdot \eta e^{-\eta y} = \lambda\eta e^{-\lambda x - \eta y} = p(x, y)$$

which proofs the independence of the random variables X and Y.

b)

Derive a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$.

First we derive the log-likelihood with parameter

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \lambda \\ \eta \end{pmatrix}.$$

The log-likelihood can be written as:

$$l(\theta) = \sum_{i=1}^{N} \ln p(x_i, y_i|\lambda, \eta) = \sum_{i=1}^{N} (\ln(\lambda) + \ln(\eta) - \lambda x_i - \eta y_i)$$

To estimate λ, we calculate the maximum of the log-likelihood, building the derivation regarding λ:

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{N} \left(\frac{1}{\lambda} - x_i\right) \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \frac{N}{\lambda} = \sum_{i=1}^{N} x_i \quad \Leftrightarrow \quad \hat{\lambda} = \frac{1}{\bar{x}} \quad \text{, where} \quad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Finally we can check the second derivation

$$\frac{\partial^2 l}{\partial \lambda^2} = \sum_{i=1}^{N} \left(-\frac{1}{\lambda^2}\right) < 0 \quad \Rightarrow \quad \max,$$

which indicates a maximum, since $\lambda > 0$ and therefore the second derivation is always smaller than 0.

c)

Derive a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$ under the constraint $\eta = 1/\lambda$.

We start again with the log-likelihood where $\eta = 1/\lambda$:

$$l(\theta) = \sum_{i=1}^{N} \left(\ln(\lambda) + \ln\left(\frac{1}{\lambda}\right) - \lambda x_i - \frac{y_i}{\lambda}\right) = \sum_{i=1}^{N} \left(\ln(\lambda) - \ln(\lambda) - \lambda x_i - \frac{y_i}{\lambda}\right) = \sum_{i=1}^{N} \left(-\lambda x_i - \frac{y_i}{\lambda}\right)$$

Calculating the derivation results in

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{N} \left(-x_i + \frac{y_i}{\lambda^2}\right) \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \sum_{i=1}^{N} x_i = \frac{1}{\lambda^2}\sum_{i=1}^{N} y_i \quad \Leftrightarrow \quad \hat{\lambda} = \sqrt{\frac{\bar{y}}{\bar{x}}}$$

for $\bar{x} \neq 0$, where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

Checking the second derivation results in

$$\frac{\partial^2 l}{\partial \lambda^2} = \sum_{i=1}^{N} \left(-2\frac{y_i}{\lambda^3}\right) < 0 \quad \Rightarrow \quad \max,$$

since $\lambda > 0$ and $y_i > 0$.

d)

Derive a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$ under the constraint $\eta = 1 - \lambda$.

Again we start with the log-likelihood where $\eta = 1 - \lambda$:

$$l(\theta) = \sum_{i=1}^{N} (\ln(\lambda) + \ln(1-\lambda) - \lambda x_i - (1-\lambda) y_i)$$

Calculating the derivation:

$$\begin{aligned}
\frac{\partial l}{\partial \lambda} &= \sum_{i=1}^{N} \left(\frac{1}{\lambda} - \frac{1}{1-\lambda} - x_i + y_i \right) \stackrel{!}{=} 0 \\
\Leftrightarrow \quad & \frac{1-2\lambda}{\lambda(1-\lambda)} = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i) \\
\Leftrightarrow \quad & 1 - 2\lambda = \lambda(1-\lambda)(\bar{x} - \bar{y}) \\
\Leftrightarrow \quad & 1 - 2\lambda - \lambda(\bar{x} - \bar{y}) + \lambda^2(\bar{x} - \bar{y}) = 0 \\
\Leftrightarrow \quad & \lambda^2 - \lambda \left(\frac{2 + \bar{x} - \bar{y}}{\bar{x} - \bar{y}} \right) + \frac{1}{\bar{x} - \bar{y}} = 0
\end{aligned}$$

We can now use the p-q-formula, to calculate the estimator:

$$\begin{aligned}
\hat{\lambda}_{1/2} &= \frac{(2 + \bar{x} - \bar{y})}{2(\bar{x} - \bar{y})} \pm \sqrt{\frac{(2 + \bar{x} - \bar{y})^2}{4(\bar{x} - \bar{y})^2} - \frac{1}{\bar{x} - \bar{y}}} \\
&= \frac{(2 + \bar{x} - \bar{y})}{2(\bar{x} - \bar{y})} \pm \sqrt{\frac{(2 + \bar{x} - \bar{y})^2 - 4(\bar{x} - \bar{y})}{4(\bar{x} - \bar{y})^2}} \\
&= \frac{(2 + \bar{x} - \bar{y}) \pm \sqrt{4 + 4(\bar{x} - \bar{y}) + (\bar{x} - \bar{y})^2 - 4(\bar{x} - \bar{y})}}{2(\bar{x} - \bar{y})} \\
&= \frac{(2 + \bar{x} - \bar{y}) \pm \sqrt{4 + (\bar{x} - \bar{y})^2}}{2(\bar{x} - \bar{y})}
\end{aligned}$$

for $\bar{x} \neq \bar{y}$. Finally we check the second deviation:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \lambda^2}(\hat{\lambda}_1) &= \sum_{i=1}^{N} \left(-\frac{1}{\hat{\lambda}_1^2} + \frac{1}{(1-\hat{\lambda}_1)^2} \right) \stackrel{(*)}{=} \sum_{i=1}^{N} (\bar{y} - \bar{x}) \left(\sqrt{(\bar{x} - \bar{y})^2 + 4} + 2 \right) < 0 \quad \text{if} \quad \bar{x} > \bar{y} \\
\frac{\partial^2 l}{\partial \lambda^2}(\hat{\lambda}_2) &= \sum_{i=1}^{N} \left(-\frac{1}{\hat{\lambda}_2^2} + \frac{1}{(1-\hat{\lambda}_2)^2} \right) \stackrel{(*)}{=} \sum_{i=1}^{N} (\bar{x} - \bar{y}) \left(\sqrt{(\bar{x} - \bar{y})^2 + 4} - 2 \right) < 0 \quad \text{if} \quad \bar{y} > \bar{x}
\end{aligned}$$

Since the second terms $\sqrt{(\bar{x} - \bar{y})^2 + 4} + 2$ and $\sqrt{(\bar{x} - \bar{y})^2 + 4} - 2$ are strictly positive (if $\bar{x} \neq \bar{y}$ holds), the result only depends on the signs of $\bar{x}$ and $\bar{y}$ in the first terms $(\bar{y} - \bar{x})$ and $(\bar{x} - \bar{y})$ respectively.

$(*)$ Wolframalpha was used here

Exercise 2: Multiple Linear Regression

Consider the multiple linear regression problem $y = \boldsymbol{x}^T \boldsymbol{\beta} + \epsilon$, where $\boldsymbol{x} \in \mathbb{R}^d$ are the predictor variables, $y \in \mathbb{R}$ is the response variable, and $\boldsymbol{\beta} \in \mathbb{R}^d$ are the linear regression coefficients. We have again a dataset $\mathcal{D} = ((\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N))$ of N independent draws of pairs $(\boldsymbol{x}_i, y_i)$. We summarize data and noise into the vectors $\boldsymbol{y} = (y_1, ..., y_N) \in \mathbb{R}^N$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_N) \in \mathbb{R}^N$, and the matrix $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N) \in \mathbb{R}^{N \times d}$.

In the lecture, we have derived the maximum-likelihood solution for $\boldsymbol{\beta}$ under the assumption of zero-mean Gaussian distributed noise (denoted by $\epsilon \sim N(0, \sigma^2)$):

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

a)

Show that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$; i.e., $\hat{\boldsymbol{\beta}}$ is Gaussian distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2 (X^T X)^{-1}$.

It holds that linear transformation of normal random variables result again in normal distributed random variables. Since $\boldsymbol{y} = X\beta + \boldsymbol{\epsilon}$ is a linear transformation of $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_N)$, $\boldsymbol{y}$ is normal distributed again. Same holds for $\hat{\boldsymbol{\beta}}$, since $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ is a linear transformation of $\boldsymbol{y}$. Then, as we know that $\hat{\boldsymbol{\beta}}$ is normal distributed, it remains to show the expectation value and the variance.

To calculate the expectation value, we use that $\boldsymbol{y} = X\beta + \boldsymbol{\epsilon}$ and $\mathbb{E}[\boldsymbol{\epsilon}] = 0$, since $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_N)$.

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(X^T X)^{-1} X^T \boldsymbol{y}] \\
&= \mathbb{E}[(X^T X)^{-1} X^T (X\beta + \boldsymbol{\epsilon})] \\
&= \mathbb{E}[(X^T X)^{-1} (X^T X)\boldsymbol{\beta}] + \mathbb{E}[(X^T X)^{-1} X^T \boldsymbol{\epsilon}] \\
&= \mathbb{E}[\boldsymbol{\beta}] + (X^T X)^{-1} X^T \mathbb{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta}
\end{aligned}$$

Note that the expectation value of a deterministic value is its value $\mathbb{E}[C] = C$ (where C is a deterministic value).

To calculate the variance, we use that $\boldsymbol{y} = X\beta + \boldsymbol{\epsilon}$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 I_N$, since $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_N)$.

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= Var((X^T X)^{-1} X^T \boldsymbol{y}) \\
&= Var((X^T X)^{-1} X^T (X\beta + \boldsymbol{\epsilon})) \\
&= Var((X^T X)^{-1} (X^T X)\boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{\epsilon}) \\
&= Var((X^T X)^{-1} X^T \boldsymbol{\epsilon}) \\
&= (X^T X)^{-1} X^T Var(\boldsymbol{\epsilon}) ((X^T X)^{-1} X^T)^T \\
&= (X^T X)^{-1} X^T (\sigma^2 I_N) X (X^T X)^{-1^T} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

Note that the variance of a deterministic value is zero $Var(Z + C) = Var(Z)$ (where Z is any random variable and C a deterministic value). Also note that $X^T X$ is symmetric and therefore it holds that $(X^T X)^T = X^T X$ and furthermore $(X^T X)^{-1^T} = (X^T X)^{-1}$.

We can now conclude that:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

b)

Discuss the benefit of knowing the full distribution of $\hat{\boldsymbol{\beta}}$ rather than only the estimate itself. What additional statements about $\boldsymbol{\beta}$ can be made (hint: variable selection)? Assume that σ^2 is known and does not need to be estimated.

If we know the distribution of $\hat{\boldsymbol{\beta}}$, we can derive probabilities. For example we calculate a confidence interval for $\boldsymbol{\beta}$. It allows us to calculate in which range the true value will be with a given probability, e.g. $1-\alpha = 0.95$. More precise: the confidence interval will include the true value with a probability of $1-\alpha$. Another advantage is to have the ability to do hypothesis testing. For example one can check if $\hat{\boldsymbol{\beta}}$ is different from zero.

We want to show the benefit, deriving a confidence interval for $\boldsymbol{\beta}$ with a given σ^2.

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}) \quad \Leftrightarrow \quad \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\sigma^2(X^TX)^{-1}}} \sim N(0,1)$$

For the probability we say that:

$$\mathbb{P}\left(q_{\alpha/2} \leq \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\sigma^2(X^TX)^{-1}}} \leq q_{1-\alpha/2}\right) \geq 1-\alpha$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}} - q_{1-\alpha/2}\sqrt{\sigma^2(X^TX)^{-1}} \leq \boldsymbol{\beta} \leq \hat{\boldsymbol{\beta}} + q_{1-\alpha/2}\sqrt{\sigma^2(X^TX)^{-1}}\right) \geq 1-\alpha$$

Where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$-quantiles of the standard normal distribution, and $q_{\alpha/2} = -q_{1-\alpha/2}$, because of symmetry. We can now directly state the confidence interval:

$$C = \left[\hat{\boldsymbol{\beta}} - q_{1-\alpha/2}\sqrt{\sigma^2(X^TX)^{-1}}, \quad \hat{\boldsymbol{\beta}} + q_{1-\alpha/2}\sqrt{\sigma^2(X^TX)^{-1}}\right]$$

c)

Assume we have measured a new data point, $\boldsymbol{x}_*$. We use our regression model to predict the response for: $\boldsymbol{x}_* : \hat{y}_* = \boldsymbol{x}_*^T\hat{\boldsymbol{\beta}}$. Derive the distribution of $\hat{y}_*$.

Arguing in the same way as in (a), $\hat{y}_*$ is normal distributed, since every linear transformation of a normal distributed random variable is normal distributed again.

For the expectation value we get:

$$\begin{aligned}
\mathbb{E}[\hat{y_*}] &= \mathbb{E}[\boldsymbol{x}_*^T\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{x}_*^T(X^TX)^{-1}X^T\boldsymbol{y}] \\
&= \mathbb{E}[\boldsymbol{x}_*^T(X^TX)^{-1}X^T(X\beta + \boldsymbol{\epsilon})] \\
&= \mathbb{E}[\boldsymbol{x}_*^T\boldsymbol{\beta}] + \boldsymbol{x}_*(X^TX)^{-1}X^T\mathbb{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{x}_*^T\boldsymbol{\beta} = y_*
\end{aligned}$$

And for the variance we can state:

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= Var(\boldsymbol{x}_*^T\hat{\boldsymbol{\beta}}) = Var(\boldsymbol{x}_*^T(X^TX)^{-1}X^T\boldsymbol{y}) \\ &= Var(\boldsymbol{x}_*^T(X^TX)^{-1}X^T(X\beta+\boldsymbol{\epsilon})) \\ &= Var(\boldsymbol{x}_*^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}) \\ &= \boldsymbol{x}_*^T(X^TX)^{-1}X^TVar(\boldsymbol{\epsilon})(\boldsymbol{x}_*^T(X^TX)^{-1}X^T)^T \\ &= \boldsymbol{x}_*^T(X^TX)^{-1}X^T(\sigma^2I_N)X(X^TX)^{-1}\boldsymbol{x}_* \\ &= \sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_* \end{aligned}$$

So it follows:

$$\hat{y_*} \sim N(y_*, \sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*)$$

d)

Discuss the benefit of also knowing that distribution in an application of your choice.

Similar to (b) we can derive probabilities. We can for example calculate a confidence interval. A confidence interval of $1-\alpha = 0.95$ would say, that the true label is in that confidence interval with a probability of 95%. We can derive it as follows:

$$\hat{y_*} \sim N(y_*, \sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*) \quad \Leftrightarrow \quad \frac{\hat{y_*} - y_*}{\sqrt{\sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*}} \sim N(0,1)$$

For the probability we say that:

$$\mathbb{P}\left(q_{\alpha/2} \leq \frac{\hat{y_*} - y_*}{\sqrt{\sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*}} \leq q_{1-\alpha/2}\right) \geq 1-\alpha$$

$$\mathbb{P}\left(\hat{y_*} - q_{1-\alpha/2}\sqrt{\sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*} \leq y_* \leq \hat{y_*} + q_{1-\alpha/2}\sqrt{\sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*}\right) \geq 1-\alpha$$

Where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$-quantiles of the standard normal distribution again, and $q_{\alpha/2} = -q_{1-\alpha/2}$, because of symmetry. We can now directly state the confidence interval:

$$C = \left[\hat{y_*} - q_{1-\alpha/2}\sqrt{\sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*}, \quad \hat{y_*} + q_{1-\alpha/2}\sqrt{\sigma^2\boldsymbol{x}_*^T(X^TX)^{-1}\boldsymbol{x}_*}\right]$$

Exercise 1: Maximum Likelihood vs. Bayes

An unfair coin is tossed seven times and the event (head or tail) is recorded at each iteration. The observed sequence of events is

$$\mathcal{D} = (x_1, x_2, \ldots, x_7) = (\text{head, head, tail, tail, head, head, head}) .$$

We assume that all tosses $x_1, x_2, \ldots$ have been generated independently following the Bernoulli probability distribution

$$P(x \,|\, \theta) = \begin{cases} \theta & \text{if } x = \text{head} \\ 1 - \theta & \text{if } x = \text{tail,} \end{cases}$$

where $\theta \in [0, 1]$ is an unknown parameter.

a)

State the likelihood function $P(\mathcal{D} \,|\, \theta)$, that depends on the parameter θ.

For the Bernoulli distribution we have the following likelihood function:

$$P(\mathcal{D} \,|\, \theta) = \prod_{k=1}^{n} \theta^{X_k} \, (1-\theta)^{1-X_k} = \theta^{\sum_{k=1}^{n} X_k} \, (1-\theta)^{n - \sum_{k=1}^{n} X_k} ,$$

where $X_k := \begin{cases} 1 & \text{if } X_k = \text{head} \\ 0 & \text{if } X_K = \text{tail} \end{cases}$, $k = 1, \ldots, n$

b)

Compute the maximum likelihood solution $\hat{\theta}$, and evaluate for this parameter the probability that the next two tosses are "head", that is, evaluate

$$P\left(x_8 = \text{head}, x_9 = \text{head} \,\middle|\, \hat{\theta}\right) . \tag{1}$$

We use the loglikelihood function to compute the maximum likelihood solution $\hat{\theta}$, which is

$$\begin{aligned} l(\theta) &= ln(P(\mathcal{D} \,|\, \theta)) \\ &= ln(\theta) \sum_{k=1}^{n} x_k + ln(1-\theta) \left(n - \sum_{k=1}^{n} x_k \right) \end{aligned}$$

First, we need to set the first derivation of $l(\theta)$ to zero:

$$\begin{aligned}
& \frac{\partial l}{\partial \theta} = \frac{1}{\theta}\sum_{k=1}^{n} x_k + \frac{(-1)}{(1-\theta)}\left(n - \sum_{k=1}^{n} x_k\right) \stackrel{!}{=} 0 \\
\Leftrightarrow \quad & \frac{1-\theta}{\theta}\sum_{k=1}^{n} x_k = n - \sum_{k=1}^{n} x_k \\
\Leftrightarrow \quad & \left(\frac{1}{\theta} - 1\right)\sum_{k=1}^{n} x_k = n - \sum_{k=1}^{n} x_k \\
\Leftrightarrow \quad & \frac{1}{\theta}\sum_{k=1}^{n} x_k = n \\
\Leftrightarrow \quad & \hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{x},
\end{aligned}$$

where $\bar{X}$ denotes the sample mean.

Now, we use the second derivation to verify that $\hat{\theta}$ is a maximum:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \theta^2} &= -\frac{1}{\theta^2}\sum_{k=1}^{n} x_k + \frac{(-1)}{(1-\theta)^2}\left(n - \sum_{k=1}^{n} x_k\right) \\
&= \sum_{k=1}^{n} -\frac{x_K}{\theta^2} - \frac{1}{(1-\theta)^2}\sum_{k=1}^{n} 1 - x_k \\
&= \sum_{k=1}^{n} \underbrace{-\frac{x_k}{\theta^2}}_{\leq 0} \underbrace{-\frac{1-x_k}{(1-\hat{\theta})^2}}_{\leq 0} < 0,
\end{aligned}$$

since $x_k \in \{0, 1\}$ for each $k = 1, \ldots, n$ and $\hat{\theta} \in (0, 1)$ (for the non-trivial cases).

Hence, depending on our sample data $\mathcal{D}$ we get:

$$\hat{\theta} = \bar{x} = \frac{1}{7} \cdot (1 + 1 + 0 + 0 + 1 + 1 + 1) = \frac{5}{7}$$

Using this parameter (and considering the independence of the single experiments X_k) we get

$$P\left(x_8 = \text{head}, x_9 = \text{head} \,\middle|\, \hat{\theta}\right) = P\left(x_8 = \text{head} \,\middle|\, \hat{\theta}\right) \cdot P\left(x_9 = \text{head} \,\middle|\, \hat{\theta}\right) = \left(\frac{5}{7}\right)^2 \approx 0.5102$$

c)

We now adopt a Bayesian view on this problem, where we assume a prior distribution for the parameter θ defined as:

$$p(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{else} \end{cases}$$

Compute the posterior distribution $p(\theta|\mathcal{D})$, and evaluate the probability that the next two tosses are head, that is,

$$\int P(x_8 = \text{head}, x_9 = \text{head} \,|\, \theta) p(\theta|\mathcal{D}) \mathrm{d}\theta.$$

The Bayes theorem states that

$$P(\theta|D) = \frac{P(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)d\theta} = \frac{P(D|\theta)p(\theta)}{\int_0^1 p(D|\theta)p(\theta)d\theta},$$

where the numerator is given by

$$P(D|\theta) \cdot p(\theta) = \theta^5(1-\theta)^2 \cdot 1 = \theta^7 - 2\theta^6 + \theta^5,$$

where $p(\theta) = 1$ in the interval $\theta \in [0, 1]$. For the denominator holds

$$\int_0^1 p(D|\theta)p(\theta)d\theta = \int_0^1 (\theta^7 - 2\theta^6 + \theta^5)d\theta = \left[\frac{1}{8}\theta^8 - \frac{2}{7}\theta^7 + \frac{1}{6}\theta^6\right]_0^1 = \frac{1}{8} - \frac{2}{7} + \frac{1}{6} = \frac{1}{168},$$

where the integration limits are from 0 to 1, regarding to the definition of $p(\theta)$, which limits the parameter space to $\Theta = [0, 1]$. The probability $P(\theta|D)$ is finally given by

$$P(\theta|D) = \frac{P(D|\theta)p(\theta)}{\int_0^1 p(D|\theta)p(\theta)d\theta} = \frac{\theta^5(1-\theta)^2}{\frac{1}{168}} = 168\ \theta^5(1-\theta)^2.$$

Now we can calculate the probability that the next two tosses are "head":

$$\begin{aligned}
\int p(x_8 = \text{head}, x_9 = \text{head}|\theta)p(\theta|D)d\theta &= \int_0^1 p(x_8 = \text{head}|\theta) \cdot p(x_9 = \text{head}|\theta) \cdot 168\ (\theta^7 - 2\theta^6 + \theta^5)d\theta \\
&= 168 \int_0^1 \theta^2 \cdot (\theta^7 - 2\theta^6 + \theta^5)d\theta \\
&= 168 \int_0^1 \theta^9 - 2\theta^6 + \theta^5 d\theta \\
&= 168 \left[\frac{1}{10}\theta^{10} - \frac{2}{9}\theta^9 + \frac{1}{8}\theta^8\right]_0^1 \\
&= 168 \left(\frac{1}{10} - \frac{2}{9} + \frac{1}{8}\right) \\
&= \frac{7}{15} \approx 0.4667
\end{aligned}$$

Exercise 2: Convergence of Bayes Parameter Estimation

We consider Section 3.4.1 of Duda et al., where the data is generated according to the univariate probability density

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

where σ^2 is known and where μ is unknown with prior distribution

$$p(\mu_0) \sim N(\mu_0, \sigma_0^2)$$

Having sampled a dataset $\mathcal{D}$ from the data-generating distribution, the posterior probability distribution over the unknown parameter μ becomes

$$p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$$

where

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}, \qquad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}, \qquad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k. \tag{2}$$

a)

Show that the variance of the posterior can be upper-bounded as follows:

$$\sigma_n^2 \leq \min\left(\frac{\sigma^2}{n}, \sigma_0^2\right) \tag{3}$$

that is, the variance of the posterior is contained both by the uncertainty of the data mean and of the prior.

First we rearrange equation (2):

$$\begin{aligned} & \frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \\ \Leftrightarrow \quad & \frac{1}{\sigma_n^2} = \frac{n\sigma_0^2 + \sigma^2}{\sigma^2\sigma_0^2} \\ \Leftrightarrow \quad & \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} \end{aligned} \tag{4}$$

Now we consider two cases. In the **first case** we assume that $\frac{\sigma^2}{n} \leq \sigma_0^2$ and with (3) we get:

$$\begin{aligned} & \sigma_n^2 \leq \frac{\sigma^2}{n} \\ \overset{(4)}{\Leftrightarrow} \quad & \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} \leq \frac{\sigma^2}{n} \\ \Leftrightarrow \quad & n\sigma_0^2 \leq n\sigma_0^2 + \sigma^2 \\ \Leftrightarrow \quad & 0 \leq \sigma^2 \end{aligned}$$

Since $\sigma^2 > 0$, the equation $0 \leq \sigma^2$ holds. In the **second case** we assume that $\frac{\sigma^2}{n} \geq \sigma_0^2$ and with (3) we get:

$$\begin{aligned} & \sigma_n^2 \leq \sigma_0^2 \\ \overset{(4)}{\Leftrightarrow} \quad & \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} \leq \sigma_0^2 \\ \Leftrightarrow \quad & \sigma^2 \leq n\sigma_0^2 + \sigma^2 \\ \Leftrightarrow \quad & 0 \leq n\sigma_0^2 \end{aligned}$$

Since $\sigma_0^2 > 0$, the equation $0 \leq n\sigma_0^2$ holds. Hence we have shown that (3) holds.

b)

Show that the mean of the posterior can be lower- and upper-bounded as follows:

$$\min\left(\hat{\mu}_n, \mu_0\right) \leq \mu_n \leq \max\left(\hat{\mu}_n, \mu_0\right) \tag{5}$$

that is, the mean of the posterior distribution lies somewhere on the segment between the mean of the prior distribution and the sample mean.

First we rearrange equation (2):

$$
\begin{aligned}
& \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \\
\Leftrightarrow \quad & \mu_n = \left(\frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}\right) \cdot \sigma_n^2 \\
\overset{(4)}{\Leftrightarrow} \quad & \mu_n = \left(\frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}\right) \cdot \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} \\
\Leftrightarrow \quad & \mu_n = \frac{n}{\sigma^2}\hat{\mu}_n \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \frac{\mu_0}{\sigma_0^2}\frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} \\
\Leftrightarrow \quad & \mu_n = \hat{\mu}_n \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \mu_0 \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \\
\Leftrightarrow \quad & \mu_n = \hat{\mu}_n \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \mu_0 \frac{\sigma^2 + n\sigma_0^2 - n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \\
\Leftrightarrow \quad & \mu_n = \hat{\mu}_n \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \mu_0 \left(\frac{\sigma^2 + n\sigma_0^2}{n\sigma_0^2 + \sigma^2} - \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right) \\
\Leftrightarrow \quad & \mu_n = \hat{\mu}_n \eta + \mu_0 (1 - \eta) \qquad \text{where } \eta \in (0, 1)
\end{aligned}
$$

We can interpret the result as μ_n is a weighted sum of the sample mean $\hat{\mu}_n$ and the mean of the prior distribution μ_0. We can consider two cases.

In the **first case** we assume that $\hat{\mu}_n \leq \mu_0$ and we get:

$$
\begin{aligned}
\mu_n &= \hat{\mu}_n\eta + \mu_0(1-\eta) \leq \mu_0\eta + \mu_0(1-\eta) = \mu_0 = \max(\hat{\mu}_n, \mu_0) \\
\mu_n &= \hat{\mu}_n\eta + \mu_0(1-\eta) \geq \hat{\mu}_n\eta + \hat{\mu}_n(1-\eta) = \hat{\mu}_n = \min(\hat{\mu}_n, \mu_0)
\end{aligned}
$$

In the **second case** we assume that $\hat{\mu}_n \geq \mu_0$ and we get:

$$
\begin{aligned}
\mu_n &= \hat{\mu}_n\eta + \mu_0(1-\eta) \geq \mu_0\eta + \mu_0(1-\eta) = \mu_0 = \min(\hat{\mu}_n, \mu_0) \\
\mu_n &= \hat{\mu}_n\eta + \mu_0(1-\eta) \leq \hat{\mu}_n\eta + \hat{\mu}_n(1-\eta) = \hat{\mu}_n = \max(\hat{\mu}_n, \mu_0)
\end{aligned}
$$

Overall we could show that equation (5) holds.

Exercise 1: Lagrange Multipliers

Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^d$ be a dataset of n samples. We consider the objective function

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{n} \|\boldsymbol{\theta} - \boldsymbol{x}_k\|^2$$

to be minimized with respect to the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. It can be shown that in absence of constraints for $\boldsymbol{\theta}$, parameter $\boldsymbol{\theta}^*$ that minimizes this objective is given by the empirical mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k$. However, this is not necessarily the case when the parameter $\boldsymbol{\theta}$ is constrained.

a)

Using the method of Lagrange multipliers, find the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{\theta}^T\boldsymbol{b} = 0$, where $b \in \mathbb{R}^d$. Give a geometrical interpretation to your solution.

First we formulate the Lagrangian function:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \lambda) &= J(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{b} = \sum_{k=1}^{n} \|\boldsymbol{\theta} - \boldsymbol{x}_k\|^2 + \lambda\boldsymbol{\theta}^T\boldsymbol{b} \\
&= \sum_{k=1}^{n} (\boldsymbol{\theta} - \boldsymbol{x}_k)^T(\boldsymbol{\theta} - \boldsymbol{x}_k) + \lambda\boldsymbol{\theta}^T\boldsymbol{b} \\
&= \sum_{k=1}^{n} \left(\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\theta}^T\boldsymbol{x}_k - \boldsymbol{x}_k^T\boldsymbol{\theta} + x_k^T x_k\right) + \lambda\boldsymbol{\theta}^T\boldsymbol{b} \\
&= n\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\sum_{k=1}^{n} \boldsymbol{\theta}^T\boldsymbol{x}_k + \sum_{k=1}^{n} x_k^T x_k + \lambda\boldsymbol{\theta}^T\boldsymbol{b}
\end{aligned}$$

Note that $\boldsymbol{\theta}^T\boldsymbol{x}_k = \boldsymbol{x}_k^T\boldsymbol{\theta}$ since they are scalar.

Next we check the necessary condition $\nabla\mathcal{L} = 0$ to find the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ under the constraint. We start with the partial derivation by $\boldsymbol{\theta}$

$$\begin{aligned}
\frac{\partial\mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial\boldsymbol{\theta}} &= 2n\boldsymbol{\theta} - 2\sum_{k=1}^{n} \boldsymbol{x}_k + \lambda\boldsymbol{b} \stackrel{!}{=} 0 \\
\Leftrightarrow \quad \boldsymbol{\theta}^* &= \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k - \frac{\lambda\boldsymbol{b}}{2n} = \bar{\boldsymbol{x}} - \frac{\lambda\boldsymbol{b}}{2n}
\end{aligned}$$

which is a shift of the unconstrained parameter. Now we calculate the partial derivation by λ

$$\frac{\partial\mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial\lambda} = \boldsymbol{\theta}^T\boldsymbol{b} \stackrel{!}{=} 0$$

where we just get the initial constraint. We can now plug in $\boldsymbol{\theta}^*$ into the second condition and we obtain:

$$\left(\bar{\boldsymbol{x}} - \frac{\lambda\boldsymbol{b}}{2n}\right)^T \boldsymbol{b} = \bar{x}^T\boldsymbol{b} - \frac{\lambda}{2n}\boldsymbol{b}^T\boldsymbol{b} = 0 \;\Rightarrow\; \lambda^* = 2n\frac{\bar{x}^T\boldsymbol{b}}{\boldsymbol{b}^T\boldsymbol{b}}$$

Plugging this into the first condition gives

$$\boldsymbol{\theta}^* = 2n\frac{\bar{x}^T\boldsymbol{b}}{\boldsymbol{b}^T\boldsymbol{b}} \cdot \boldsymbol{b} = \bar{x} - p_b(\bar{\boldsymbol{x}}),$$

where $p_b(\bar{\boldsymbol{x}})$ is a projection of $\bar{\boldsymbol{x}}$ to $\boldsymbol{b}$.

To verify the minimum, we have to calculate the second derivation and have a look to the hessian matrix. First we calculate the partial derivations:

$$\frac{\partial^2\mathcal{L}(\boldsymbol{\theta},\lambda)}{\partial\boldsymbol{\theta}^2} = 2n\boldsymbol{I}_d, \quad \frac{\partial^2\mathcal{L}(\boldsymbol{\theta},\lambda)}{\partial\boldsymbol{\theta}\partial\lambda} = \boldsymbol{b}, \quad \frac{\partial^2\mathcal{L}(\boldsymbol{\theta},\lambda)}{\partial\lambda^2} = 0$$

The hessian matrix $H_{\mathcal{L}}(\boldsymbol{\theta}^*,\lambda^*)$ has the following form:

$$H_{\mathcal{L}}(\boldsymbol{\theta}^*,\lambda^*) = \begin{pmatrix} \frac{\partial^2\mathcal{L}(\boldsymbol{\theta}^*,\lambda^*)}{\partial\boldsymbol{\theta}^2} & \frac{\partial^2\mathcal{L}(\boldsymbol{\theta}^*,\lambda^*)}{\partial\boldsymbol{\theta}\partial\lambda} \\ \frac{\partial^2\mathcal{L}(\boldsymbol{\theta}^*,\lambda^*)}{\partial\lambda\partial\boldsymbol{\theta}} & \frac{\partial^2\mathcal{L}(\boldsymbol{\theta}^*,\lambda^*)}{\partial\lambda^2} \end{pmatrix} = \begin{pmatrix} 2n & & 0 & b_1 \\ & \ddots & & \vdots \\ 0 & & 2n & b_d \\ b_1 & \dots & b_d & 0 \end{pmatrix}$$

If $\boldsymbol{b} > 0$, $H_{\mathcal{L}}(\boldsymbol{\theta}^*,\lambda^*)$ is positive definite and therefore $J(\boldsymbol{\theta})$ under the constraint of $\boldsymbol{\theta}^T\boldsymbol{b} = 0$ is minimized by $\boldsymbol{\theta}^* = \bar{\boldsymbol{x}} - \frac{\lambda\boldsymbol{b}}{2n}$.

More simply we can argue that a problem including a norm $\|\cdot\|^2$ is always a convex problem and therefore calculating the second derivative is not necessary here.

Geometrical interpretation: The new minimizing parameter $\boldsymbol{\theta}^*$ is a shift of the unconstrained parameter $\boldsymbol{\theta}$ (the empirical mean $\bar{\boldsymbol{x}}$) by $\frac{\lambda\boldsymbol{b}}{2n}$. More precise it's a projection of $\bar{\boldsymbol{x}}$ to $\boldsymbol{b}$ (as obtained above).

b)

Using the same method, find the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to $\|\boldsymbol{\theta} - \boldsymbol{c}\|^2 = 1$, where $c \in \mathbb{R}^d$. Give a geometrical interpretation to your solution.

We formulate the Lagrangian function again, analog to (a):

$$\mathcal{L}(\boldsymbol{\theta},\lambda) = n\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\sum_{k=1}^{n}\boldsymbol{\theta}^T\boldsymbol{x}_k + \sum_{k=1}^{n}x_k^Tx_k + \lambda(\|\boldsymbol{\theta} - \boldsymbol{c}\|^2 - 1)$$

Next we check the necessary condition $\nabla\mathcal{L} = 0$ to find the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ under the constraint.

We start with the partial derivation by $\boldsymbol{\theta}$

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} &= 2n\boldsymbol{\theta} - 2\sum_{k=1}^{n} \boldsymbol{x}_k + 2\lambda(\boldsymbol{\theta} - \boldsymbol{c}) \\ &= 2\boldsymbol{\theta}(n+\lambda) - 2\sum_{k=1}^{n} \boldsymbol{x}_k - 2\lambda\boldsymbol{c} \stackrel{!}{=} 0 \\ \Leftrightarrow \quad \boldsymbol{\theta}^* &= \frac{1}{n+\lambda}\sum_{k=1}^{n} \boldsymbol{x}_k + \frac{n}{n+\lambda}\boldsymbol{c} \\ &= \frac{n\bar{\boldsymbol{x}} + \lambda\boldsymbol{c}}{n+\lambda} \end{aligned} \tag{1}$$

Now we calculate the partial derivation by λ

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \lambda} = \|\boldsymbol{\theta} - \boldsymbol{c}\|^2 - 1 \stackrel{!}{=} 0 \tag{2}$$

where we just get the initial constraint. If we put (1) into (2), we obtain

$$\begin{aligned}&\left\| \frac{n\bar{\boldsymbol{x}} + \lambda\boldsymbol{c}}{n+\lambda} - \boldsymbol{c} \right\|^2 = \left\| \frac{n\bar{\boldsymbol{x}} + \lambda\boldsymbol{c}}{n+\lambda} - \frac{\boldsymbol{c}(n+\lambda)}{n+\lambda} \right\|^2 = n^2 \left\| \frac{\bar{\boldsymbol{x}} - \boldsymbol{c}}{n+\lambda} \right\|^2 = 1 \\ &\Leftrightarrow \ (n+\lambda)^2 = n^2 \|\bar{\boldsymbol{x}} - \boldsymbol{c}\|^2 \\ &\Leftrightarrow \ n + \lambda = \pm n \|\bar{\boldsymbol{x}} - \boldsymbol{c}\|, \quad \text{where } `+` \text{ if } n+\lambda > 0 \text{ and } `-` \text{ if } n+\lambda < 0 \\ &\Leftrightarrow \ \lambda^* = \pm n \|\bar{\boldsymbol{x}} - \boldsymbol{c}\| - n.\end{aligned}$$

If we plug λ^* into 1, we get

$$\begin{aligned}\boldsymbol{\theta}^* &= \frac{n\bar{\boldsymbol{x}} + (\pm n \|\bar{\boldsymbol{x}} - \boldsymbol{c}\| - n)\boldsymbol{c}}{n + (\pm n \|\bar{\boldsymbol{x}} - \boldsymbol{c}\| - n)} = \frac{n\bar{\boldsymbol{x}} \pm n \|\bar{\boldsymbol{x}} - \boldsymbol{c}\| \boldsymbol{c} - n\boldsymbol{c}}{n \pm n \|\bar{\boldsymbol{x}} - \boldsymbol{c}\| - n} \\ &= \frac{\bar{\boldsymbol{x}} - \boldsymbol{c} \pm \|\bar{\boldsymbol{x}} - \boldsymbol{c}\| \boldsymbol{c}}{\pm \|\bar{\boldsymbol{x}} - \boldsymbol{c}\|} = \frac{\bar{\boldsymbol{x}} - \boldsymbol{c}}{\pm \|\bar{\boldsymbol{x}} - \boldsymbol{c}\|} + \frac{\pm \|\bar{\boldsymbol{x}} - \boldsymbol{c}\| \boldsymbol{c}}{\pm \|\bar{\boldsymbol{x}} - \boldsymbol{c}\|} \\ &= \pm \frac{\bar{\boldsymbol{x}} - \boldsymbol{c}}{\|\bar{\boldsymbol{x}} - \boldsymbol{c}\|} + \boldsymbol{c}, \quad \text{where } `+` \text{ if } n+\lambda > 0 \text{ and } `-` \text{ if } n+\lambda < 0.\end{aligned}$$

To verify the minimum, we have to calculate the second derivation and have a look to the hessian matrix. First we calculate the partial derivations:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}^2} = 2(n+\lambda)\boldsymbol{I}_d, \quad \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta} \partial \lambda} = 2(\boldsymbol{\theta} - \boldsymbol{c}), \quad \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \lambda^2} = 0$$

Plugging in the parameter $\boldsymbol{\theta}^*$ to the mixes derivation, we get:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^*, \lambda^*)}{\partial \boldsymbol{\theta} \partial \lambda} = 2(\boldsymbol{\theta}^* - \boldsymbol{c}) = 2\left(\frac{n\bar{\boldsymbol{x}} + \lambda\boldsymbol{c}}{n+\lambda} - \boldsymbol{c} \right) = 2\left(\frac{n\bar{\boldsymbol{x}} + \lambda\boldsymbol{c} - (n+\lambda)\boldsymbol{c}}{n+\lambda} \right) = \frac{2n}{n+\lambda}(\bar{\boldsymbol{x}} - \boldsymbol{c})$$

Defining $\boldsymbol{h} = \frac{2n}{n+\lambda}(\bar{\boldsymbol{x}} - \boldsymbol{c})$, the hessian matrix $H_{\mathcal{L}}(\boldsymbol{\theta}^*, \lambda^*)$ can be written as:

$$H_{\mathcal{L}}(\boldsymbol{\theta}^*, \lambda^*) = \begin{pmatrix} \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^*, \lambda^*)}{\partial \boldsymbol{\theta}^2} & \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^*, \lambda^*)}{\partial \boldsymbol{\theta} \partial \lambda} \\ \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^*, \lambda^*)}{\partial \lambda \partial \boldsymbol{\theta}} & \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^*, \lambda^*)}{\partial \lambda^2} \end{pmatrix} = \begin{pmatrix} 2(n+\lambda) & & 0 & h_1 \\ & \ddots & & \vdots \\ 0 & & 2(n+\lambda) & h_d \\ h_1 & \dots & h_d & 0 \end{pmatrix}$$

If $\boldsymbol{c} < \bar{\boldsymbol{x}}$, $H_{\mathcal{L}}(\boldsymbol{\theta}^*, \lambda^*)$ is positive definite and therefore $J(\boldsymbol{\theta})$ under the constraint of $\|\boldsymbol{\theta} - \boldsymbol{c}\|^2 = 1$ is minimized by $\boldsymbol{\theta}^* = \frac{n\bar{\boldsymbol{x}} + \lambda \boldsymbol{c}}{n+\lambda}$.

More simply we can argue that a problem including a norm $\|\cdot\|^2$ is always a convex problem and therefore calculating the second derivative is not necessary here.

Geometrical interpretation: The new parameter $\boldsymbol{\theta}^* = \pm \frac{\bar{\boldsymbol{x}} - \boldsymbol{c}}{\|\bar{\boldsymbol{x}} - \boldsymbol{c}\|} + \boldsymbol{c}$ is shifted by $\boldsymbol{c}$ and normalized. It is restricted to a circle with center $\boldsymbol{c}$ and radius 1.

Exercise 2: Bounds on Eigenvalues

We consider a dataset $x_1, ..., x_n \in \mathbb{R}^d$. The empirical mean m, and the scatter matrix S are given by

$$\boldsymbol{m} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k \qquad \text{and} \qquad \boldsymbol{S} = \sum_{k=1}^{n} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^T$$

Let λ_1 be the largest eigenvalue of the matrix $\boldsymbol{S}$. The eigenvalue λ_1 quantifies the amount of variation in the data on the first principal component. Because computation of the full scatter matrix and respective eigenvalues can be slow, it can be useful to relate them to the diagonal elements of the scatter matrix $\{\boldsymbol{S}_{ii}\}$ that can be computed in linear time.

a)

Show that $\sum_{i=1}^{d} \boldsymbol{S}_{ii}$ is an upper bound to the eigenvalue λ_1.

We use that the trace of a diagonalizable matrix is the sum of its eigenvalues. Since $\boldsymbol{S}$ is symmetric, it is also diagonalizable (spectral theorem), hence we get:

$$Tr(\boldsymbol{S}) = \sum_{i=1}^{d} \boldsymbol{S}_{ii} = \sum_{i=1}^{d} \lambda_i$$

It holds that $\lambda_i \geq 0 \; \forall \; i = 1, ..., d$, since $\boldsymbol{S}$ is positive semi-definite, which means that $\boldsymbol{z}^T \boldsymbol{S} \boldsymbol{z} \geq 0 \; \forall \; z \in \mathbb{R}^d, \; z \neq 0$. $\lambda_i \geq 0$ implies that:

$$\lambda_1 \leq \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{d} \boldsymbol{S}_{ii}$$

which shows that $\sum_{i=1}^{d} \boldsymbol{S}_{ii}$ is an upper bound to the eigenvalue λ_1.

b)

State the conditions on the data for which the upper bound is tight.

It holds that $\lambda_1 = \sum_{i=1}^{d} \boldsymbol{S}_{ii}$ if and only if $\lambda_i = 0$ for all $i = 2, ..., d$.

In this case the matrix $\boldsymbol{S}$ is of rank 1, which means that all features are linearly dependent.

c)

Show that $\max_{i=1}^{d} \boldsymbol{S}_{ii}$ is a lower bound to the eigenvalue λ_1.

Let's define $\lambda_1 = \lambda_{\max}$ for clearity in notation. Since we know from (a) that $\boldsymbol{S}$ is diagonalizable, the matrix can be represented as

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

where the matrix $\boldsymbol{U} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_d)$ contains the eigenvectors $\boldsymbol{u}_1, ..., \boldsymbol{u}_d$ and has the property that:

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}_d$$

The matrix $\boldsymbol{\Lambda} = diag(\lambda_1, ..., \lambda_d)$ contains the eigenvalues $\lambda_1, ..., \lambda_d$.

Solution 1

For any standardized vector $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\| = 1$ we get:

$$\boldsymbol{v}^T\boldsymbol{S}\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{v}$$

And we define $\boldsymbol{w} = \boldsymbol{U}\boldsymbol{v} \in \mathbb{R}^d$. Now we get:

$$\begin{aligned}
\boldsymbol{v}^T\boldsymbol{S}\boldsymbol{v} &= \boldsymbol{v}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{v} = \boldsymbol{w}^T\boldsymbol{\Lambda}\boldsymbol{w} \\
&= \lambda_1 w_1^2 + ... + \lambda_d w_d^2 \\
&\leq \lambda_{\max}(w_1^2 + ... + w_d^2) \\
&= \lambda_{\max}\|\boldsymbol{w}\|^2 = \lambda_{\max}\boldsymbol{w}^T\boldsymbol{w} \\
&= \lambda_{\max}\boldsymbol{v}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{v} = \lambda_{\max}\boldsymbol{v}^T\boldsymbol{I}_d\boldsymbol{v} \\
&= \lambda_{\max}\boldsymbol{v}^T\boldsymbol{v} = \lambda_{\max}\|v\|^2 = \lambda_{\max}
\end{aligned}$$

If we choose $\boldsymbol{v} = \boldsymbol{u}_i$, we get

$$\boldsymbol{S}_{ii} = \boldsymbol{u}_i^T\boldsymbol{S}\boldsymbol{u}_i \leq \lambda_{\max}$$

for any $i = 1, ..., d$, which implies $\max_{i=1}^{d} \boldsymbol{S}_{ii} \leq \lambda_{\max}$.

Solution 2

Without loss of generalization we can state that the eigenvectors are normalized, which means that

$$\boldsymbol{U} = \{\boldsymbol{u} : \|\boldsymbol{u}\| = 1\}$$

and we define $\boldsymbol{C} \subset \boldsymbol{U}$ as

$$\boldsymbol{C} = \{(1, ..., 0), ..., (0, ..., 1)\} = \{\delta_1, ..., \delta_d\}.$$

We can now simply follow

$$\lambda_{\max} = \max_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u}^T \boldsymbol{S} \boldsymbol{u} \geq \max_{\boldsymbol{u} \in \boldsymbol{C}} \boldsymbol{u}^T \boldsymbol{S} \boldsymbol{u} = \max_{i=1,...,d} \boldsymbol{\delta}_i^T \boldsymbol{S} \boldsymbol{\delta}_i = \max_{i=1,...,d} \boldsymbol{S}_{ii}$$

d)

State the conditions on the data for which the lower bound is tight.

$\max_{i=1}^{d} \boldsymbol{S}_{ii} = \lambda_{\max}$ holds in following cases:

- The condition holds if $\boldsymbol{S}$ is diagonal. If we consider $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, the values of $\boldsymbol{\Lambda}$ and $\boldsymbol{S}$ are related if $\boldsymbol{S}$ is diagonal.
- There is just one value λ, such that $\lambda_1, ..., \lambda_d \in \{\lambda, 0\}$.
- In means that the lower bound is tight if the algebraic multiplicity is $m_\lambda = rank(\boldsymbol{S})$.

Exercise 3: Iterative PCA

When performing principal component analysis, computing the full eigendecomposition of the scatter matrix $\boldsymbol{S}$ is typically slow, and we are often only interested in the few first principal components. An efficient procedure to find the first eigenvector is the power iteration method, which starts with a random vector $\boldsymbol{w} \in \mathbb{R}^d$, and iteratively applies the parameter update

$$\boldsymbol{w} \leftarrow \frac{\boldsymbol{S}\boldsymbol{w}}{\|\boldsymbol{S}\boldsymbol{w}\|}$$

until some convergence criterion is met.

a)

Show that application of the power iteration method is equivalent to defining the unconstrained objective

$$J(\boldsymbol{w}) = \|\boldsymbol{S}\boldsymbol{w}\| - \frac{1}{2}\boldsymbol{w}^T \boldsymbol{S} \boldsymbol{w} \tag{3}$$

and performing the gradient ascent $\boldsymbol{v} \leftarrow \boldsymbol{v} + \gamma \frac{\partial J}{\partial \boldsymbol{v}}$, where $\boldsymbol{v} = \boldsymbol{S}^{0.5}\boldsymbol{w}$ is a reparameterization of $\boldsymbol{v}$, for some learning rate γ. We assume that the matrix $\boldsymbol{S}$ is invertible.

Since $\boldsymbol{S}$ is symmetric, we can state that:

$$\begin{aligned} \boldsymbol{v} = \boldsymbol{S}^{1/2}\boldsymbol{w} \quad \text{and} \quad \boldsymbol{v}^T = \boldsymbol{w}^T(\boldsymbol{S}^{1/2})^T = \boldsymbol{w}^T\boldsymbol{S}^{1/2} \\ \boldsymbol{w} = \boldsymbol{S}^{-1/2}\boldsymbol{v} \quad \text{and} \quad \boldsymbol{w}^T = \boldsymbol{v}^T(\boldsymbol{S}^{-1/2})^T = \boldsymbol{v}^T\boldsymbol{S}^{-1/2} \end{aligned}$$

We can now replace $\boldsymbol{w}$ in (3) by $\boldsymbol{S}^{-1/2}\boldsymbol{v}$, so we get:

$$\begin{aligned} J(\boldsymbol{v}) &= \|\boldsymbol{S}\boldsymbol{S}^{-1/2}\boldsymbol{v}\| - \frac{1}{2}\boldsymbol{v}^T\boldsymbol{v} \\ &= \|\boldsymbol{S}^{1/2}\boldsymbol{v}\| - \frac{1}{2}\|\boldsymbol{v}\|^2 \end{aligned}$$

Calculating the derivation of $J(\boldsymbol{v})$ results in

$$\frac{\partial J(\boldsymbol{v})}{\partial \boldsymbol{v}} = \frac{\boldsymbol{S}^{1/2}\boldsymbol{v}}{\|\boldsymbol{S}^{1/2}\boldsymbol{v}\|} \cdot \boldsymbol{S}^{1/2} - \boldsymbol{v} = \frac{\boldsymbol{S}\boldsymbol{v}}{\|\boldsymbol{S}^{1/2}\boldsymbol{v}\|} - \boldsymbol{v},$$

since it holds that $\frac{\partial}{\partial \boldsymbol{x}}\|\boldsymbol{a}\boldsymbol{x}^n - \boldsymbol{b}\| = \frac{\boldsymbol{a}\boldsymbol{x}^n - \boldsymbol{b}}{\|\boldsymbol{a}\boldsymbol{x}^n - \boldsymbol{b}\|} n\boldsymbol{a}\boldsymbol{x}^{n-1}$.

Now we use the result to plug it into $\boldsymbol{v} + \gamma \frac{\partial J}{\partial \boldsymbol{v}}$

$$\boldsymbol{v} + \gamma \left(\frac{\boldsymbol{S}\boldsymbol{v}}{\|\boldsymbol{S}^{1/2}\boldsymbol{v}\|} - \boldsymbol{v} \right) = \boldsymbol{v} + \frac{\boldsymbol{S}\boldsymbol{v}}{\|\boldsymbol{S}^{1/2}\boldsymbol{v}\|} - \boldsymbol{v} = \frac{\boldsymbol{S}\boldsymbol{v}}{\|\boldsymbol{S}^{1/2}\boldsymbol{v}\|}$$

where we chose $\gamma = \boldsymbol{I}$. If we replace $\boldsymbol{v}$ by $\boldsymbol{S}^{1/2}\boldsymbol{w}$

$$\boldsymbol{v} = \boldsymbol{S}^{1/2}\boldsymbol{w} \leftarrow \frac{\boldsymbol{S}\boldsymbol{v}}{\|\boldsymbol{S}^{1/2}\boldsymbol{v}\|} = \frac{\boldsymbol{S}\boldsymbol{S}^{1/2}\boldsymbol{w}}{\|\boldsymbol{S}^{1/2}\boldsymbol{S}^{1/2}\boldsymbol{w}\|} = \frac{\boldsymbol{S}^{3/2}\boldsymbol{w}}{\|\boldsymbol{S}\boldsymbol{w}\|},$$

we can show by dividing by $\boldsymbol{S}^{1/2}$

$$\boldsymbol{w} \leftarrow \frac{\boldsymbol{S}\boldsymbol{w}}{\|\boldsymbol{S}\boldsymbol{w}\|},$$

that the power iteration method is equivalent to defining the unconstrained objective $J(\boldsymbol{w})$ and performing the gradient ascent.

b)

Show that a necessary condition for $\boldsymbol{w}$ to maximize the objective $J(\boldsymbol{w})$ is to be a unit vector (i.e. $\|\boldsymbol{w}\| = 1$).

We can simply show that:

$$\|\boldsymbol{w}\| \overset{(a)}{=} \left\| \frac{\boldsymbol{S}\boldsymbol{w}}{\|\boldsymbol{S}\boldsymbol{w}\|} \right\| = \frac{1}{\|\boldsymbol{S}\boldsymbol{w}\|}\|\boldsymbol{S}\boldsymbol{w}\| = 1$$

Exercise 1: Finding the direction of maximal correlation between datasets

In some applications, one might want to perform dimensionality reduction for two simultaneously acquired multivariate datasets. For example, in a neuroimaging experiment, one might simultaneously acquire functional magnetic resonance images (fMRI) and electroencephalographic (EEG) signals of the same participant. These two techniques measure quite different aspects of brain activity (fMRI measures blood flow while EEG measures electrical activity), and the measurements are generally differently scaled. Submitting the concatenated data to a joint PCA dimensionality reduction therefore bears the risk that the decomposition is dominated by one measurement modality at the expense of neglecting the other. On the other hand, performing separate PCAs bears the risk that the found principal subspaces miss interesting relationships between the two datasets that may be present in the original data.

The correlation coefficient

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{1}$$

is a measure of the linear relationship between two variables that is independent of their scale and mean value. Assume the presence of two data sets $\mathcal{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ and $\mathcal{Y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_N\}$, where the samples $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are measured at the same time, and where all $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ and $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$. Assume that all data have already been centered ($\sum_{i=1}^N \boldsymbol{x}_i = 0$ and $\sum_{i=1}^N \boldsymbol{y}_i = 0$). We are interested in finding linear projections $\boldsymbol{w}_x \in \mathbb{R}^{d_x}$ and $\boldsymbol{w}_y \in \mathbb{R}^{d_y}$ that, when applied to $\mathcal{X}$ and $\mathcal{Y}$, maximize the correlation coefficient of the projected data.

a)

Establish the objective of this optimization problem as a function of $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$.

First we determine that $\mathrm{Cov}(X,Y) = \Sigma_{X,Y}$, $\mathrm{Var}(X) = \Sigma_{X,X}$ and $\mathrm{Var}(Y) = \Sigma_{Y,Y}$. Furthermore we have that

$$\boldsymbol{\mu}_X = \sum_{i=1}^N \boldsymbol{x}_i = 0 \quad \text{and} \quad \boldsymbol{\mu}_Y = \sum_{i=1}^N \boldsymbol{y}_i = 0. \tag{2}$$

For simplification of notation we say that $X = \mathcal{X}$ and $Y = \mathcal{Y}$, where X and Y is now the data set and not the random variable from equation (1).

We also define $\tilde{X} = \boldsymbol{w}_x^T X$ and $\tilde{Y} = \boldsymbol{w}_y^T Y$.

Starting with the sample correlation, we get:

$$\begin{aligned}
\rho(\boldsymbol{w}_x, \boldsymbol{x}_y) &= \frac{\Sigma_{\tilde{X},\tilde{Y}}}{\sqrt{\Sigma_{\tilde{X},\tilde{X}}\Sigma_{\tilde{Y},\tilde{Y}}}} = \frac{(\tilde{X}-\boldsymbol{\mu}_{\tilde{X}})(\tilde{Y}-\boldsymbol{\mu}_{\tilde{Y}})^T}{\sqrt{\|\tilde{X}-\boldsymbol{\mu}_{\tilde{X}}\|^2\|\tilde{Y}-\boldsymbol{\mu}_{\tilde{Y}}\|^2}} \\
&= \frac{(\boldsymbol{w}_x^T X - \boldsymbol{w}_x^T\boldsymbol{\mu}_X)(\boldsymbol{w}_y^T Y - \boldsymbol{w}_y^T\boldsymbol{\mu}_Y)^T}{\sqrt{\|\boldsymbol{w}_x^T X - \boldsymbol{w}_x^T\boldsymbol{\mu}_X\|^2\|\boldsymbol{w}_y^T Y - \boldsymbol{w}_y^T\boldsymbol{\mu}_Y\|^2}} \\
&\overset{(2)}{=} \frac{\boldsymbol{w}_x^T X(\boldsymbol{w}_y^T Y)^T}{\sqrt{\|\boldsymbol{w}_x^T X\|^2\|\boldsymbol{w}_y^T Y\|^2}} \\
&= \frac{\boldsymbol{w}_x^T XY^T\boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T\|X\|^2\boldsymbol{w}_x}\sqrt{\boldsymbol{w}_y^T\|Y\|^2\boldsymbol{w}_y}} \\
&= \frac{\boldsymbol{w}_x^T\Sigma_{X,Y}\boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T\Sigma_{X,X}\boldsymbol{w}_x}\sqrt{\boldsymbol{w}_y^T\Sigma_{Y,Y}\boldsymbol{w}_y}}
\end{aligned}$$

b)

Derive analytic expressions for $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ at the optimum.

The correlation needs to stay standardized. Since the denominator does not influence the maximization, we choose variances which makes the derivation easy. So the problem we have to solve is:

$$\max \rho(\boldsymbol{w}_x, \boldsymbol{x}_y) \quad \text{s.t.} \quad \boldsymbol{w}_x^T\Sigma_{X,X}\boldsymbol{w}_x = 1 \quad \text{and} \quad \boldsymbol{w}_y^T\Sigma_{Y,Y}\boldsymbol{w}_y = 1$$

Regarding (a) the problem can be written as (if we consider the constraints):

$$\max \boldsymbol{w}_x^T\Sigma_{X,Y}\boldsymbol{w}_y \quad \text{s.t.} \quad \boldsymbol{w}_x^T\Sigma_{X,X}\boldsymbol{w}_x = 1 \quad \text{and} \quad \boldsymbol{w}_y^T\Sigma_{Y,Y}\boldsymbol{w}_y = 1$$

We can now formulate the Lagrange function:

$$\mathcal{L}(\boldsymbol{w}_x, \boldsymbol{x}_y; \lambda, \mu) = \boldsymbol{w}_x^T\Sigma_{X,Y}\boldsymbol{w}_y - \frac{1}{2}\lambda(\boldsymbol{w}_x^T\Sigma_{X,X}\boldsymbol{w}_x - 1) - \frac{1}{2}\mu(\boldsymbol{w}_y^T\Sigma_{Y,Y}\boldsymbol{w}_y - 1)$$

Calculating the derivations leads to:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}_x} = \Sigma_{X,Y}\boldsymbol{w}_y - \lambda\Sigma_{X,X}\boldsymbol{w}_x \overset{!}{=} 0 \tag{3}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}_y} = \boldsymbol{w}_x^T\Sigma_{X,Y} - \mu\Sigma_{Y,Y}\boldsymbol{w}_y = \Sigma_{X,Y}\boldsymbol{w}_x - \mu\Sigma_{Y,Y}\boldsymbol{w}_y \overset{!}{=} 0 \tag{4}$$

The reshape in (4) is possible, because $\frac{\partial}{\partial\boldsymbol{x}}\boldsymbol{x}^T\boldsymbol{y} = \frac{\partial}{\partial\boldsymbol{x}}\boldsymbol{y}^T\boldsymbol{x}$ and $\Sigma_{X,Y} = \Sigma_{X,Y}^T$.

Now we multiply $\boldsymbol{w}_x^T$ and $\boldsymbol{w}_y^T$ from the left:

$$\begin{aligned}
\boldsymbol{w}_x^T\Sigma_{X,Y}\boldsymbol{w}_y - \lambda\boldsymbol{w}_x^T\Sigma_{X,X}\boldsymbol{w}_x = \boldsymbol{w}_x^T\Sigma_{X,Y}\boldsymbol{w}_y - \lambda = 0 \quad &\Leftrightarrow \quad \lambda = \boldsymbol{w}_x^T\Sigma_{X,Y}\boldsymbol{w}_y \\
\boldsymbol{w}_y^T\Sigma_{X,Y}\boldsymbol{w}_x - \mu\boldsymbol{w}_y^T\Sigma_{Y,Y}\boldsymbol{w}_y = \boldsymbol{w}_y^T\Sigma_{X,Y}\boldsymbol{w}_x - \mu = 0 \quad &\Leftrightarrow \quad \mu = \boldsymbol{w}_y^T\Sigma_{X,Y}\boldsymbol{w}_x
\end{aligned}$$

Note that $\Sigma_{\tilde{X},\tilde{Y}} = \boldsymbol{w}_x^T \Sigma_{X,Y} \boldsymbol{w}_y = \boldsymbol{w}_y^T \Sigma_{X,Y} \boldsymbol{w}_x = \Sigma_{\tilde{Y},\tilde{X}}$ and regarding the constraint it is $\boldsymbol{w}_x^T \Sigma_{X,X} \boldsymbol{w}_x = \boldsymbol{w}_y^T \Sigma_{Y,Y} \boldsymbol{w}_y = 1$.

It directly follows that

$$\lambda = \mu = \boldsymbol{w}_x^T \Sigma_{X,Y} \boldsymbol{w}_y = \rho(\boldsymbol{w}_x, \boldsymbol{x}_y) \qquad \text{(under the constraint)} \tag{5}$$

Now (3) and (4) have the following form, since $\lambda = \mu$:

$$\Sigma_{X,Y} \boldsymbol{w}_y - \lambda \Sigma_{X,X} \boldsymbol{w}_x = 0 \tag{6}$$
$$\Sigma_{X,Y} \boldsymbol{w}_x - \lambda \Sigma_{Y,Y} \boldsymbol{w}_y = 0 \tag{7}$$

Going on with (7), we get:

$$\Sigma_{X,Y} \boldsymbol{w}_x = \lambda \Sigma_{Y,Y} \boldsymbol{w}_y \quad \Leftrightarrow \quad \boldsymbol{w}_y = \frac{1}{\lambda} \Sigma_{Y,Y}^{-1} \Sigma_{X,Y} \boldsymbol{w}_x$$

Plugging $\boldsymbol{w}_y$ into (6) results in:

$$\begin{aligned} & \frac{1}{\lambda} \Sigma_{X,Y} \Sigma_{Y,Y}^{-1} \Sigma_{X,Y} \boldsymbol{w}_x - \lambda \Sigma_{X,X} \boldsymbol{w}_x = 0 \\ \Leftrightarrow \quad & \lambda^2 \Sigma_{X,X} \boldsymbol{w}_x = \Sigma_{X,Y} \Sigma_{Y,Y}^{-1} \Sigma_{X,Y} \boldsymbol{w}_x \\ \Leftrightarrow \quad & \lambda^2 \boldsymbol{w}_x = \Sigma_{X,X}^{-1} \Sigma_{X,Y} \Sigma_{Y,Y}^{-1} \Sigma_{X,Y} \boldsymbol{w}_x \end{aligned}$$

We now have an eigenvalue problem

$$A \boldsymbol{w}_x = \tilde{\lambda} \boldsymbol{w}_x$$

where $\tilde{\lambda} = \lambda^2$ and $A = \Sigma_{X,X}^{-1} \Sigma_{X,Y} \Sigma_{Y,Y}^{-1} \Sigma_{X,Y}$.

The solution for $\boldsymbol{w}_y$ is analog. We have to start from equation (6) and plug the result into (7). We get

$$B \boldsymbol{w}_y = \tilde{\lambda} \boldsymbol{w}_y$$

where $B = \Sigma_{Y,Y}^{-1} \Sigma_{X,Y} \Sigma_{X,X}^{-1} \Sigma_{X,Y}$.

c)

Derive an analytic expression for the correlation coefficient at the optimum.

With (b) we know that $\lambda^2 = \tilde{\lambda}$ solves the eigenvalue problem.

Since we know from (5), that (under the constraint) $\lambda = \boldsymbol{w}_x^T \Sigma_{X,Y} \boldsymbol{w}_y = \rho(\boldsymbol{w}_x, \boldsymbol{w}_y)$, we already have the solution for correlation coefficient at the optimum:

$$\rho(\boldsymbol{w}_x, \boldsymbol{w}_y) = \lambda = \sqrt{\tilde{\lambda}}$$

Exercise 2: Fisher and Bayes

In the asymptotic case where the sampled data for two classes tends to the probability density functions $p(\boldsymbol{x}|\omega_1)$ and $p(\boldsymbol{x}|\omega_2)$ of means and covariances $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, and with class priors $P(\omega_1)$, $P(\omega_2)$, Fisher linear discriminant finds a projection vector $\boldsymbol{w}$ that maximizes the objective.

$$J_{\text{Fisher}}(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \Sigma_B \boldsymbol{w}}{\boldsymbol{w}^T \Sigma_W \boldsymbol{w}}$$

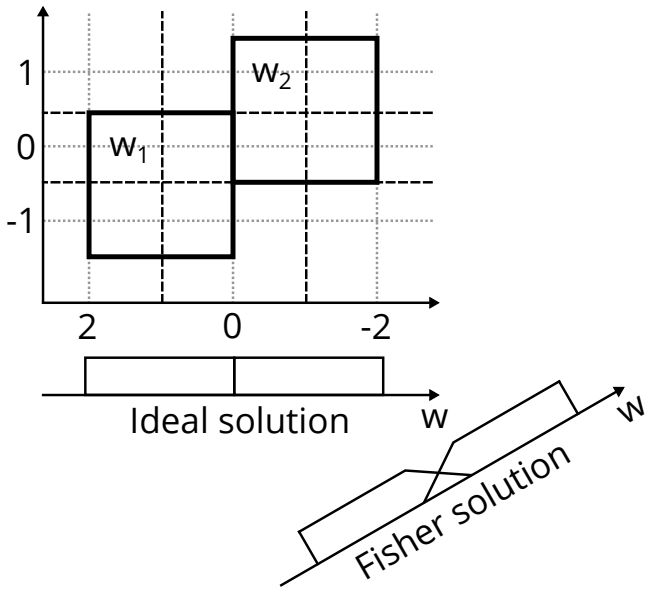
where $\Sigma_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ is the between-class covariance matrix and $\Sigma_W = P(\omega_1)\boldsymbol{\Sigma}_1 + P(\omega_2)\boldsymbol{\Sigma}_2$ is the within-class covariance matrix. A solution to this problem is given in closed form as:

$$\boldsymbol{w}* = \Sigma_W^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

The function $\phi(\boldsymbol{x}) = \langle \boldsymbol{w}*, \boldsymbol{x} \rangle$ is called the Fisher linear discriminant. Let ϕ be one element of the set of possible mappings Φ (e.g. given by the set of all possible one-dimensional linear projections $\mathbb{R}^d \to \mathbb{R}$ of the the data). We say that ϕ is optimal in the Bayes sense, if when building the decision boundary in the image of ϕ, no other mapping in the set Φ supports a decision boundary with lower expected error. For example, in the special case of two Gaussian distributions with covariances $\Sigma_1 = \Sigma_2$, the Fisher linear discriminant is optimal in the Bayes sense when considering the set of all possible one-dimensional linear projections, and even when considering all possible functions.

a)

Find two non-Gaussian two-dimensional probability distributions $p(\boldsymbol{x}|\omega_1)$ and $p(\boldsymbol{x}|\omega_2)$ with same covariances matrices $\Sigma_1 = \Sigma_2$ such that the best one-dimensional linear projection in the Bayes sense is different from the best one-dimensional linear projection in the Fisher sense. Sketch these two distributions along with the best linear projection in the Fisher and Bayes sense.



Consider two distributions $p(\boldsymbol{x}|\omega_1) \sim U(a, b)^2 = U(-1, 1)^2$ and $p(\boldsymbol{x}|\omega_2) \sim U(-1, 1)^2$, which are 2 dimensional.

The distributions can be seen in the figure. Define the expectation as:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$$

Let's define X as a random variable for the first dimension, and Y as a random variable for the second dimension. While we assume, that the 2 dimensions are independent and the covariance is symmetric, we can follow that $\text{Cov}(X,Y) = \text{Cov}(Y,X) = 0$ and since both dimensions are identically distributed, we also have that $\text{Var}(X) = \text{Var}(Y) = \frac{1}{12}(b-a)^2 = \frac{1}{12}(1-(-1))^2) = \frac{1}{3}$. We obtain

$$\boldsymbol{\Sigma} = \text{Cov}\left(\begin{pmatrix} X \\ Y \end{pmatrix}, \begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(Y,X) & \text{Var}(Y) \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{3}\boldsymbol{I}_2$$

We can now solve the fisher distribution, using

$$\boldsymbol{w} = \boldsymbol{\Sigma}_W^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = 4\boldsymbol{I}\left(\begin{pmatrix} 1 \\ 0.5 \end{pmatrix} - \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}\right) = 4\begin{pmatrix} 2 \\ 1 \end{pmatrix} \neq \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

where $\binom{1}{0}$ would be the optimal Bayes solution. Both solutions are shown in the figure. In that case Fisher linear discriminant differs from best linear prediction in Bayes sense. The linear prediction in Bayes sense can easily be obtained by

$$\begin{aligned} p(\omega_1|\boldsymbol{x}) &> p(\omega_2|\boldsymbol{x}) \\ p(\boldsymbol{x}|\omega_1)P(\omega_1) &> p(\boldsymbol{x}|\omega_2)P(\omega_2) \\ p(\boldsymbol{x}|\omega_1) &> p(\boldsymbol{x}|\omega_2), \end{aligned}$$

where $P(\omega_1) = P(\omega_2) = \mathbf{1}_{[-1,1]^2}(\boldsymbol{x})$ (?).

a)

Consider now that the two classes are generated by two d-dimensional Gaussian distributions $p(\boldsymbol{x}|\omega_1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $p(\boldsymbol{x}|\omega_2) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. Find a mapping $\phi \in \mathbb{R}^d \to \mathbb{R}$ which for fixed mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ is optimal in the Bayes sense. (*Hint: your function ϕ should depend on $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, but not on the prior probabilities $P(\omega_1)$ and $P(\omega_2)$.*)

The classes are generated by Gaussian distributions, their densites are given by:

$$\begin{aligned} p(\boldsymbol{x}|\omega_1) &= \frac{1}{\sqrt{2\pi \det(\boldsymbol{\Sigma}_1)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right) \\ p(\boldsymbol{x}|\omega_2) &= \frac{1}{\sqrt{2\pi \det(\boldsymbol{\Sigma}_2)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right) \end{aligned}$$

We start with the posterior distribution, which decides for ω_1, if probability for ω_1 is higher than for ω_2 (given $\boldsymbol{x}$).

$$p(\omega_1|\boldsymbol{x}) > p(\omega_2|\boldsymbol{x})$$

Using Bayes theorem, we get

$$\begin{aligned}\frac{p(\boldsymbol{x}|\omega_1)P(\omega_1)}{p(\boldsymbol{x})} &> \frac{p(\boldsymbol{x}|\omega_2)P(\omega_2)}{p(\boldsymbol{x})} \\ p(\boldsymbol{x}|\omega_1)P(\omega_1) &> p(\boldsymbol{x}|\omega_2)P(\omega_2),\end{aligned}$$

where we can now take the logarithm to simplify the calculation:

$$\begin{aligned}\ln p(\boldsymbol{x}|\omega_1) + \ln P(\omega_1) &> \ln p(\boldsymbol{x}|\omega_2) + \ln P(\omega_2) \\ \ln p(\boldsymbol{x}|\omega_1) - \ln p(\boldsymbol{x}|\omega_2) &> \ln P(\omega_2) - \ln P(\omega_1)\end{aligned}$$

Plugging in the distributions from above, we obtain:

$$\begin{aligned}&-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(\boldsymbol{\Sigma}_1)) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\det(\boldsymbol{\Sigma}_2)) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \\ &\quad > \ln P(\omega_2) - \ln P(\omega_1) \\ &(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) > \ln P(\omega_2) - \ln P(\omega_1) + \ln(\det(\boldsymbol{\Sigma}_1)) - \ln(\det(\boldsymbol{\Sigma}_2))\end{aligned}$$

We can identify the first part of the inequation as the mapping $\phi : \mathbb{R}^d \to \mathbb{R}$

$$\phi(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1),$$

which leads to an optimal solution in Bayes sense.

Exercise 1: Discrete EM: Coin Tosses from Multiple Distributions

Consider the following coin tossing experiment. The experimenter has three (potentially unfair) coins with different probabilities for heads and tails, one secret/hidden coin, call it coin H, and two coins that are undistinguishable in the observation, call them coin A and coin B. First, the secret coin H is tossed. If it shows heads, the experimenter tosses coin A m times, if it shows tails he tosses coin B m times. What is observed is the results from either coin A or coin B, but not the fact from which coin the results originated.

Mathematically, this is modeled by two random variables: Z for the secret/hidden coin H, X for the second coin which is either coin A or coin B. Z is a Bernoulli variable with

$$P(Z = \text{heads}) \,|\, \theta) = \lambda \qquad\qquad P(Z = \text{tails}) \,|\, \theta) = 1 - \lambda$$

and X is an m-vector with each component an independent conditional Bernoulli variable, i.e.,

$$\begin{aligned} P(X_i = \text{heads}) \,|\, Z = \text{heads}, \theta) &= p_1 & P(X_i = \text{tails}) \,|\, Z = \text{heads}, \theta) &= 1 - p_1 \\ P(X_i = \text{heads}) \,|\, Z = \text{tails}, \theta) &= p_2 & P(X_i = \text{tails}) \,|\, Z = \text{tails}, \theta) &= 1 - p_2 \end{aligned}$$

for each $1 \leq i \leq m$, where the components X_i are assumed independent, and where p_1 is the heads-probability of coin A, and p_2 the heads-probability of coin B. X is the generative observable (= variable that is observed) for the experiment.

Now, the experiment above is repeated N times. That is, coin H is tossed N times, and each time coin A or B is tossed m times. In total, $m \cdot N$ tosses are observed in the sample. We model this by a N-tuple of independent copies $\mathcal{X} := (X^{(1)}, ..., X^{(N)})$ of X, which form the statistical sample.

The goal is now to estimate the parameter vector $\theta = (\lambda, p_1, p_2)$ which contains all information on the experiment. This cannot be done directly, since λ is a hidden parameter, corresponding to the latent (=lat. hidden) variable Z, but it can be done by Expectation Maximization. Due to independence, the joint distribution of the observable X and latent variables Z for a fixed sample $x = (x^{(1)}, ..., x^{(N)})$ (an N-tuple of m-vectors with heads or tails), and for the results of the hidden coin $z = (z^{(1)}, ..., z^{(N)})$ (an N-tuple of heads or tails), is

$$P(\mathcal{X} = x, Z = z \,|\, \theta) = \prod_{i=1}^{N} P(Z = z^{(i)}) \,|\, \theta) \prod_{j=1}^{m} P(X_j = x_j^{(i)} \,|\, Z = z^{(i)}, \theta),$$

The M-step of EM determines the new parameters θ^{new} given the values of the old parameters θ^{old} by solving the optimization problem

$$\theta^{\text{new}} = \text{argmax}_\theta \, Q(\theta, \theta^{\text{old}}) \tag{1}$$

with

$$Q(\theta, \theta^{\text{old}}) = \sum_{z \in \{\text{heads,tails}\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \log P(\mathcal{X} = x, Z = z \,|\, \theta)$$

Derive the closed-form solutions for the M-step (Equation 1). That is, derive an expression for each element of the new parameter vector $\theta^{\text{new}} = (\hat{\lambda}, \hat{p}_1, \hat{p}_2)$.

To simplify notation, we define tail = 0 and head = 1. We start to formulate the function $Q(\theta, \theta^{\text{old}})$:

$$
\begin{aligned}
Q(\theta, \theta^{\text{old}}) &= \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \ln P(\mathcal{X} = x, Z = z \,|\, \theta) \\
&\overset{\text{iid}}{=} \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \ln \left[\prod_{i=1}^{N} \left(P(Z = z^{(i)}) \,|\, \theta) \prod_{j=1}^{m} P(X_j = x_j^{(i)} \,|\, Z = z^{(i)}, \theta) \right) \right] \\
&= \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \left[\sum_{i=1}^{N} \left(\ln P(Z = z^{(i)}) \,|\, \theta) + \sum_{j=1}^{m} \ln P(X_j = x_j^{(i)} \,|\, Z = z^{(i)}, \theta) \right) \right] \\
&= \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \left[\sum_{i=1}^{N} \ln P(Z = z^{(i)}) \,|\, \theta) + \sum_{i=1}^{N} \sum_{j=1}^{m} \ln P(X_j = x_j^{(i)} \,|\, Z = z^{(i)}, \theta) \right]
\end{aligned}
$$

Now we plug in Bernoulli distribution and we continue as follows:

$$
\begin{aligned}
&= \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \left[\sum_{i=1}^{N} \ln \left(\lambda^{z^{(i)}} (1 - \lambda^{1 - z^{(i)}}) \right) + \right. \\
&\qquad \left. \sum_{i=1}^{N} \sum_{j=1}^{m} \ln \left(\left[p_1^{x_j^{(i)}} (1 - p_1^{1 - x_j^{(i)}}) \right]^{z^{(i)}} \left[p_2^{x_j^{(i)}} (1 - p_2^{1 - x_j^{(i)}}) \right]^{1 - z^{(i)}} \right) \right] \\
&= \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \left[\sum_{i=1}^{N} \left(z^{(i)} \ln(\lambda) + (1 - z^{(i)}) \ln(1 - \lambda) \right) + \right. \\
&\qquad \left. \sum_{i=1}^{N} \sum_{j=1}^{m} \left(z^{(i)} \ln \left[p_1^{x_j^{(i)}} (1 - p_1^{1 - x_j^{(i)}}) \right] + (1 - z^{(i)}) \ln \left[p_2^{x_j^{(i)}} (1 - p_2^{1 - x_j^{(i)}}) \right] \right) \right] \\
&= \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \left[\sum_{i=1}^{N} \left(z^{(i)} \ln(\lambda) + (1 - z^{(i)}) \ln(1 - \lambda) \right) + \right. \\
&\qquad \left. \sum_{i=1}^{N} \sum_{j=1}^{m} \left(z^{(i)} \left[x_j^{(i)} \ln(p_1) + (1 - x_j^{(i)}) \ln(1 - p_1) \right] + (1 - z^{(i)}) \ln \left[x_j^{(i)} \ln(p_2) + (1 - x_j^{(i)}) \ln(1 - p_2) \right] \right) \right]
\end{aligned}
$$

For minimization we calculate the partial derivation of $Q(\theta, \theta^{\text{old}})$ with respect to $\theta = (\lambda, p_1, p_2)$.

$$\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial \lambda} = \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \left(\frac{z^{(i)}}{\lambda} - \frac{1 - z^{(i)}}{1 - \lambda} \right) \stackrel{!}{=} 0$$

Multiply the whole line with $\lambda(1 - \lambda)$:

$$\Leftrightarrow \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \left(z^{(i)} - z^{(i)}\lambda - \lambda + z^{(i)}\lambda \right) = 0$$

$$\Leftrightarrow \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \left(\sum_{i=1}^{N} z^{(i)} - N\lambda \right) = 0$$

$$\Leftrightarrow \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \cdot \sum_{i=1}^{N} z^{(i)} = N\lambda \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}})$$

$$\Leftrightarrow \hat{\lambda} = \frac{\sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \cdot \sum_{i=1}^{N} z^{(i)}}{N \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}})}$$

$$\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial p_1} = \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \sum_{j=1}^{m} \left(z^{(i)} \left[\frac{x_j^{(i)}}{p_1} - \frac{1 - x_j^{(i)}}{1 - p_1} \right] \right) \stackrel{!}{=} 0$$

Multiply the whole line with $p_1(1 - p_1)$:

$$\Leftrightarrow \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \sum_{j=1}^{m} \left(z^{(i)} \left[x_j^{(i)} - x_j^{(i)} p_1 - p_1 + x_j^{(i)} p_1 \right] \right) = 0$$

$$\Leftrightarrow \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \left(\sum_{i=1}^{N} \sum_{j=1}^{m} z^{(i)} x_j^{(i)} - m\, p_1 \sum_{i=1}^{N} z^{(i)} \right) = 0$$

$$\Leftrightarrow \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \sum_{j=1}^{m} z^{(i)} x_j^{(i)} = m\, p_1 \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} z^{(i)}$$

$$\Leftrightarrow \hat{p}_1 = \frac{\sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \sum_{j=1}^{m} z^{(i)} x_j^{(i)}}{m \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} z^{(i)}}$$

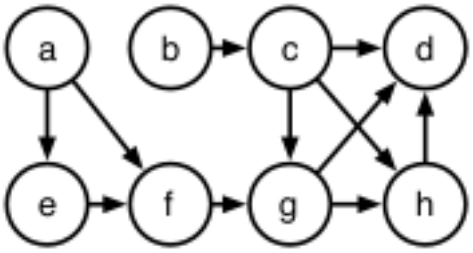
The calculation of $\hat{p}_2$ follows analog, where we have $1 - z^{(i)}$ instead of $z^{(i)}$, we get:

$$\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial p_2} = \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \sum_{j=1}^{m} \left((1 - z^{(i)}) \left[\frac{x_j^{(i)}}{p_2} - \frac{1 - x_j^{(i)}}{1 - p_2} \right] \right) \stackrel{!}{=} 0$$

$$\Leftrightarrow \hat{p}_2 = \frac{\sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} \sum_{j=1}^{m} \left[(1 - z^{(i)}) \, x_j^{(i)} \right]}{m \sum_{z \in \{0,1\}^N} P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}}) \sum_{i=1}^{N} (1 - z^{(i)})}$$

We have a closed-form solution if the derivations are solvable. Since $P(Z = z \,|\, \mathcal{X} = x, \theta^{\text{old}})$ depends on θ^{old}, which is known, and thus calculable, we can solve the derivations and have a closed-form solution.

Exercise 1: Graphical Models and Conditional Independence



For the directed graphical model above, show by using the basic rules of probability theory and by using the rules of d-separation.

a)

Whether a and b are conditionally independent given the empty set?

Rule of probability theory

We have to check if $p(a,b) \stackrel{?}{=} p(a)p(b)$. We factorize $p(a,b,c,d,e,f,g,h)$ and marginalize by c,d,e,f,g,h:

$$
\begin{aligned}
p(a,b) &= \sum_{c,d,e,f,g,h} p(a,b,c,d,e,f,g,h) \\
&= \sum_{c,d,e,f,g,h} p(a)p(e|a)p(f|a,e)p(b)p(c|b)p(d|c,g,h)p(g|c,f)p(h|c,d,g) \\
&= p(a)p(b) \sum_{c,d,e,f,g,h} p(c|b)p(d|c,g,h)p(e|a)p(f|a,e)p(g|c,f)p(h|c,d,g) \\
&= p(a)p(b) \sum_c p(c|b) \left(\sum_d p(d|c,g,h) \left(\sum_e p(e|a) \left(\sum_f p(f|a,e) \left(\sum_g p(g|c,f) \left(\sum_h p(h|c,d,g) \right) \right) \right) \right) \right) \\
&= p(a)p(b) \sum_c p(c|b) \left(\sum_d p(d|c,g,h) \left(\sum_e p(e|a) \left(\sum_f p(f|a,e) \left(\sum_g p(g|c,f) \cdot 1 \right) \right) \right) \right) \\
&= p(a)p(b) \sum_c p(c|b) \left(\sum_d p(d|c,g,h) \left(\sum_e p(e|a) \left(\sum_f p(f|a,e) \cdot 1 \right) \right) \right) \\
&= p(a)p(b) \sum_c p(c|b) \left(\sum_d p(d|c,g,h) \left(\sum_e p(e|a) \cdot 1 \right) \right) \\
&= p(a)p(b) \sum_c p(c|b) \left(\sum_d p(d|c,g,h) \cdot 1 \right) = p(a)p(b) \sum_c p(c|b) \cdot 1 \\
&= p(a)p(b) \cdot 1 = p(a)p(b)
\end{aligned}
$$

$\Rightarrow a \perp\!\!\!\perp b \mid \emptyset$

Rule of d-separation

g, h and d are head-to-head nodes, and neither the nodes, nor any of its descendants is in the conditioning set (empty set). Therefore all paths are blocked and a is d-separated from b by the empty set.

$\Rightarrow a \perp\!\!\!\perp b \mid \emptyset$

b)

Whether a and b are conditionally independent given the variable h?

Rule of probability theory

We have to check if $p(a, b|h) \stackrel{?}{=} p(a|h)p(b|h)$. Using Bayes theorem we get:

$$p(a, b|h) = \frac{p(h|a, b)p(a, b)}{p(h)} \stackrel{(a)}{=} \frac{p(h|a, b)p(a)p(b)}{p(h)} = p(a)p(b)\frac{p(h|a, b)}{p(h)}$$

and

$$p(a|h)p(b|h) = \frac{p(h|a)p(a)}{p(h)}\frac{p(h|b)p(b)}{p(h)} = p(a)p(b)\frac{p(h|a)p(h|b)}{p(h)^2}$$

And we can see that $p(a, b|h) \neq p(a|h)p(b|h)$, which means that $a \not\perp\!\!\!\perp b \mid h$.

Rule of d-separation

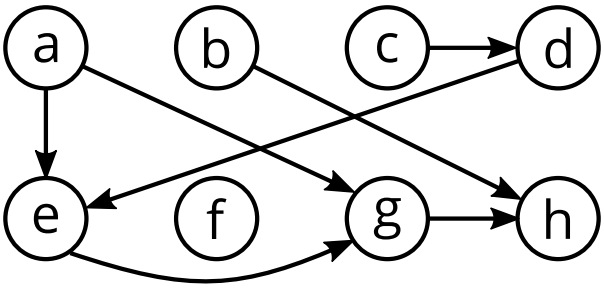
Although g, h and d are head-to-head nodes, h is part of the conditioning set. Therefore this path is not blocked by h. We can follow that a is not d-separated from b by h.

$\Rightarrow a \not\perp\!\!\!\perp b \mid h$

c)

Draw the directed graphical model for the following joint probability distribution:

$$p(a)p(b)p(c)p(d|c)p(e|a, d)p(g|e, a)p(h|b, g)$$



Exercise 2: Graphical Models and Discriminants

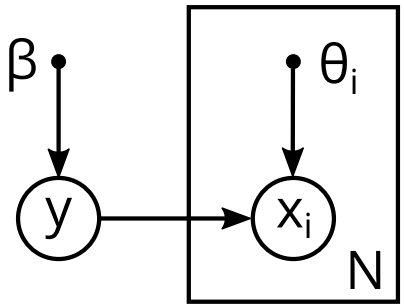
Assume that we have an image x consisting d binary pixels x_i. Each image depicts a handwritten digit with class $y \in \{0, ..., 9\}$. Such an image can be modeled as follows. We make the assumption, that observed variables $x_1, ..., x_d$ are conditionally independent given y. That is, the joint probability distribution factorizes as

$$P(\boldsymbol{x}, y) = P(y) \cdot \prod_{i=1}^{d} P(x_i, y),$$

where each factor has its own parameters. In particular, we set $\beta_k = P(y = k)$, and $\theta_{ik} = P(x_i = 1|y = k)$. Please note that each pixel x_i is modeled independently.

a)

Draw the directed graphical model (i.e. the set of nodes, the node labels, the directed connections between nodes, and the parameters of the network to be defined).



b)

Show that the most likely class $k = \text{argmax}_k P(y = k|\boldsymbol{x})$ for a given image $\boldsymbol{x}$ can be found by a set of linear discriminants $g_k(\boldsymbol{x}) = \boldsymbol{w}_k\boldsymbol{x} + b_k$, where $k \in \{0, ..., 9\}$ is the class, and where the parameters of the discriminants $\{\boldsymbol{w}_k, b_k\}$ can be written as a function of $\{\beta_k, \theta_{ik}\}$.

The pixels follow a Bernoulli distribution $x_i|y = k \sim Bern(\theta_{i,k})$ with density $p(x_i|y) = \theta_{i,k}^{x_i}(1 - \theta_{i,k})^{1-x_i}$, since they are binary (either 0 or 1).

$$
\begin{aligned}
k &= \text{argmax}_k P(y=k|\boldsymbol{x}) = \text{argmax}_k \frac{P(\boldsymbol{x}, y=k)}{P(x)} \\
&= \text{argmax}_k P(\boldsymbol{x}, y=k), \quad \text{since } P(\boldsymbol{x}) = \text{const.} \\
&= \text{argmax}_k \ln P(\boldsymbol{x}, y=k) \\
&= \text{argmax}_k \ln \left(P(y=k) \cdot \prod_{i=1}^{d} P(x_i, y=k) \right) \\
&= \text{argmax}_k \left[\ln P(y=k) + \sum_{i=1}^{d} \ln P(x_i, y=k) \right] \\
&= \text{argmax}_k \left[\ln \beta_k + \left(\sum_{i \in d:\, x_i=1} \ln P(x_i=1, y=k) + \sum_{j \in d:\, x_j=0} \ln P(x_i=0, y=k) \right) \right] \\
&= \text{argmax}_k \left[\ln \beta_k + \sum_{i \in d:\, x_i=1} \ln \theta_{i,k} + \sum_{j \in d:\, x_j=0} \ln(1-\theta_{j,k}) \right] \\
&= \text{argmax}_k \left[\ln \beta_k + (\ln \boldsymbol{\theta}_k)^T \boldsymbol{x} + (\ln(1-\boldsymbol{\theta}_k))^T (1-\boldsymbol{x}) \right] \\
&= \text{argmax}_k \left[\ln \beta_k + (\ln \boldsymbol{\theta}_k)^T \boldsymbol{x} + (\ln(1-\boldsymbol{\theta}_k))^T - (\ln(1-\boldsymbol{\theta}_k))^T \boldsymbol{x} \right] \\
&= \text{argmax}_k \left[\left(\ln \beta_k + (\ln(1-\boldsymbol{\theta}_k))^T \right) + \left((\ln \boldsymbol{\theta}_k)^T - (\ln(1-\boldsymbol{\theta}_k))^T \right) \boldsymbol{x} \right] \\
&= \text{argmax}_k (b_k + \boldsymbol{w}_k \boldsymbol{x}) = \text{argmax}_k \; g_k(\boldsymbol{x})
\end{aligned}
$$

Another alternative (more straight forward) solution is given by

$$
\begin{aligned}
k &= \text{argmax}_k P(y=k|\boldsymbol{x}) = \text{argmax}_k \frac{P(\boldsymbol{x}|y=k)P(y=k)}{P(\boldsymbol{x})} \\
&= \text{argmax}_k P(y=k) P(\boldsymbol{x}|y=k) \\
&= \text{argmax}_k \ln P(y=k) + \ln P(\boldsymbol{x}|y=k) \\
&= \text{argmax}_k \ln \beta_k + \sum_{i=1}^{d} \ln P(x_i|y=k) \\
&= \text{argmax}_k \ln \beta_k + \sum_{i=1}^{d} \ln \left(\theta_{i,k}^{x_i} (1-\theta_{i,k})^{1-x_i} \right) \\
&= \text{argmax}_k \ln \beta_k + \sum_{i=1}^{d} \left[x_i \ln(\theta_{i,k}) + (1-x_i) \ln(1-\theta_{i,k}) \right] \\
&= \text{argmax}_k \ln \beta_k + \ln(\boldsymbol{\theta}_k)^T \boldsymbol{x} + \ln(\mathbf{1}-\boldsymbol{\theta}_k)^T (\mathbf{1}-\boldsymbol{x}) \\
&= \text{argmax}_k \left(\ln \beta_k + \ln(\mathbf{1}-\boldsymbol{\theta}_k)^T \right) + \left(\ln(\boldsymbol{\theta}_k)^T - \ln(\mathbf{1}-\boldsymbol{\theta}_k)^T \right) \boldsymbol{x} \\
&= \text{argmax}_k (b_k + \boldsymbol{w}_k \boldsymbol{x}) = \text{argmax}_k \; g_k(\boldsymbol{x}),
\end{aligned}
$$

where $b_k = \ln \beta_k + \ln(\mathbf{1}-\boldsymbol{\theta}_k)^T$ and $\boldsymbol{w}_k = \ln(\boldsymbol{\theta}_k)^T - \ln(\mathbf{1}-\boldsymbol{\theta}_k)^T$.

Exercise 3: Finding the Most Likely Solution

Last year for Christmas we received a robotic vacuum-cleaner that was programmed badly. Each day, we switch it on in the living room. Today, 7 minutes, after turning it on, we found the vacuum cleaner in the office. Unlike

most vacuum cleaners, ours is rather fast and spends only exactly one minute in each room before moving onto and cleaning the next one (rather poorly unfortunately). What was the most likely sequence of rooms that our robot visited?

From almost a year of intense observation of robot-cleaning activities we have learned that when the robot is in the living room, it will be one minute later in the kitchen with probability 0.5, in the office with probability 0.3 and in the bedroom with probability 0.2. If the robot is in the office then it will get stuck under the desk for a minute with probability 0.1, with probability 0.662 it will end up in the kitchen. With a probability of just 0.008, it is rather unlikely that the robot goes from the office to the bedroom, which should come as no surprise since it has to climb a series of stairs to get to the bedroom. Finally, with probability 0.23 it will decide to clean the living room next. If the robot is in the bedroom, a minute later it will be in the living room with probability 0.3, with probability 0.6, it will have fallen down the stairs and into the office. Only in 10% of our observations, we found the robot in the kitchen the minute after cleaning the bedroom. Surprisingly, after spending a minute in the kitchen, the robot will always move to a different room, and each rooms is equally likely.

First we will structure the observed probabilities $P(T_{i+1}|T_i)$ from room T_i to room T_{i+1}. Or in other words, it is the probability that the cleaner moves to room T_{i+1}, given the information that it is in room T_i.

		T_i			
		L	K	O	B
T_{i+1}	L	0	0.33	0.23	0.3
	K	0.5	0	0.662	0.1
	O	0.3	0.33	0.1	0.6
	B	0.2	0.33	0.008	0

In the table L means *living room*, K is the *kitchen*, O the *office* and B the *bedroom*. We know that the cleaner starts in L (living room) and we observe it in O (office) 7 minutes later. It will change the room every minute, so we have 7 steps (room changes).

We have to find those 7 steps (under constraint of starting end ending position), such that the likelihood is maximizes, which means that:

$$\max L = \max \prod_{i=1}^{7} P(T_{i+1}|T_i)$$

Equivalently we search for those $T_2, ..., T_7$, given T_1 and T_8, such that

$$\underset{T_2,...,T_6}{\operatorname{argmax}} \; P(T_2, T_3, T_4, T_5, T_6, T_7|T_1, T_8) = \underset{T_2,...,T_6}{\operatorname{argmax}} \; \frac{P(T_1, ..., T_8)}{P(T_1, T_8)}.$$

With just trying all possible paths and calculating how likely each path is, we obtain that the most likely path is

$$L \xrightarrow{0.3} O \xrightarrow{0.662} K \xrightarrow{0.33} B \xrightarrow{0.6} O \xrightarrow{0.662} K \xrightarrow{0.33} B \xrightarrow{0.6} O,$$

which has a likelihood of:

$$L = \prod_{i=1}^{7} P(T_{i+1}|T_i) \approx 0.005$$

Exercise 1: Bias and Variance of Mean Estimators

Assume we have an estimator $\hat{\theta}$ for a parameter θ. The bias of the estimator $\hat{\theta}$ is the difference between the true value for the estimator, and its expected value:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \theta].$$

If $\text{Bias}(\hat{\theta}) = 0$ then $\hat{\theta}$ is called unbiased. The variance of the estimator $\hat{\theta}$ is the expected square deviation from its expected value:

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

The mean squared error of the estimator $\hat{\theta}$ is

$$\text{Error}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

Let $X_1, ..., X_N$ be a sample of i.i.d random variables. Assume that X_i has mean μ and variance σ^2. Calculate bias, variance and mean squared error of the following mean estimators:

a)

$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ (i.e. the sample mean)

First we will calculate the expectation value of $\hat{\mu}$:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] \overset{\text{lin.}}{=} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_i] = \mathbb{E}[X_i] = \mu$$

Using the expectation value, we can easily calculate the bias

$$\text{Bias}(\hat{\mu}) = \mathbb{E}[\hat{\mu} - \mu] = \mathbb{E}[\hat{\mu}] - \mu = \mu - \mu = 0$$

For the variance we get:

$$\text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \overset{\text{iid}}{=} \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}(X_i) = \frac{1}{N}\text{Var}(X_i) = \frac{\sigma^2}{N}$$

So the error will be:

$$\text{Error}(\hat{\mu}) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}) = 0^2 + \frac{\sigma^2}{N} = \frac{\sigma^2}{N}$$

b)

$\hat{\mu} = 0$

We start again calculating the expectation value:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[0] = 0$$

For the bias we can follow:

$$\text{Bias}(\hat{\mu}) = \mathbb{E}[\hat{\mu} - \mu] = \mathbb{E}[\hat{\mu}] - \mu = 0 - \mu = -\mu$$

And for the variance we get:

$$\text{Var}(\hat{\mu}) = \text{Var}(0) = 0$$

So the error will be:

$$\text{Error}(\hat{\mu}) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}) = (-\mu)^2 + 0 = \mu^2$$

Exercise 2: Bias-Variance Decomposition for Regression

Let $y = f(x)$ be a function mapping input to output and evaluated at some out-of-sample data point x. Consider an estimator $\hat{f}(x)$ that is obtained by training a regression model on some random sample $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ of the function $y = f(x)$.

a)

Prove the bias-variance decomposition

$$\text{Error}(\hat{f}(x)) = \text{Bias}(\hat{f}(x))^2 + \text{Var}(\hat{f}(x))$$

where the mean squared error, bias and variance are given by

$$\text{Error}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - f(x))^2], \qquad \text{Bias}(\hat{f}(x)) = \mathbb{E}[\hat{f}(x) - f(x)], \qquad \text{Var}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2].$$

$$\begin{aligned}
\text{Error}(\hat{f}(x)) &= \mathbb{E}[(\hat{f}(x) - f(x))^2] \\
&= \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - f(x))^2] \\
&= \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2 + 2\,(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])(\mathbb{E}[\hat{f}(x)] - f(x)) + (\mathbb{E}[\hat{f}(x)] - f(x))^2] \\
&= \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + 2\,\mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])(\mathbb{E}[\hat{f}(x)] - f(x))] + \mathbb{E}[(\mathbb{E}[\hat{f}(x)] - f(x))^2] \\
&= \text{Var}(\hat{f}(x)) + 2\,(\mathbb{E}[\hat{f}(x)] - f(x))\,\mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])] + (\mathbb{E}[\hat{f}(x)] - f(x))^2 \\
&= \text{Var}(\hat{f}(x)) + 2\,(\mathbb{E}[\hat{f}(x)] - f(x))\,(\mathbb{E}[\hat{f}(x)] - \mathbb{E}[\hat{f}(x)]) + \text{Bias}(\hat{f}(x))^2 \\
&= \text{Var}(\hat{f}(x)) + 2\,(\mathbb{E}[\hat{f}(x)] - f(x)) \cdot 0 + \text{Bias}(\hat{f}(x))^2 \\
&= \text{Var}(\hat{f}(x)) + \text{Bias}(\hat{f}(x))^2
\end{aligned}$$

Exercise 3: Bias-Variance Decomposition for Classification

The bias-variance decomposition usually applies to regression data. In this exercise, we would like to obtain similar decomposition for classification, in particular, when the prediction is given as a probability distribution over C classes. Let $P = [P_1, ..., P_C]$ be the ground truth class distribution associated to a particular input pattern. Assume a random estimator of class probabilities $\hat{P} = [\hat{P}_1, ..., \hat{P}_C]$ for the same input pattern. The error function is given by the KL-divergence between the ground truth and the estimated probability distribution:

$$\text{Error} = \mathbb{E}[D_{KL}(P||\hat{P})].$$

First, we would like to determine the mean of of the class distribution estimator $\hat{P}$. We define the mean as the distribution that minimizes its expected KL divergence from the the class distribution estimator, that is, the distribution R that optimizes

$$\min \mathbb{E}[D_{KL}(R||\hat{P})].$$

a)

Show that the solution to the optimization problem above is given by

$$R = [R_1, ..., R_C], \quad \text{where} \quad R_i = \frac{\exp \mathbb{E}[\ln \hat{P}_i]}{\sum_j \exp \mathbb{E}[\ln \hat{P}_j]} \quad \forall\, 1 \leq i \leq C.$$

First we want to formulate the KL divergence:

$$D_{KL}(P||\hat{P}) = \sum_{i=1}^{C} P_i \ln\left(\frac{P_i}{\hat{P}_i}\right) = \sum_{i=1}^{C} P_i(\ln P_i - \ln \hat{P}_i)$$

It holds that $\sum_i R_i = 1$, since R is a probability distribution. Considering this constraint, we can formulate the Lagrange function:

$$\begin{aligned}
\mathcal{L}(R, \lambda) &= \mathbb{E}[D_{KL}(R||\hat{P})] + \lambda\left(\sum_{i=1}^{C} R_i - 1\right) \\
&= \mathbb{E}\left[\sum_{i=1}^{C} R_i(\ln R_i - \ln \hat{P}_i)\right] + \lambda\left(\sum_{i=1}^{C} R_i - 1\right) \\
&\overset{\text{lin.}}{=} \sum_{i=1}^{C} R_i(\ln R_i - \mathbb{E}[\ln \hat{P}_i]) + \lambda\left(\sum_{i=1}^{C} R_i - 1\right)
\end{aligned}$$

Calculating the derivations, we obtain:

$$\begin{aligned}
\frac{\partial \mathcal{L}(R_i, \lambda)}{\partial R_i} &= \ln R_i - \mathbb{E}[\ln \hat{P}_i] + R_i\left(\frac{1}{R_i}\right) + \lambda = \ln R_i - \mathbb{E}[\ln \hat{P}_i] + 1 + \lambda \overset{!}{=} 0 \\
&\Rightarrow R_i = \exp(\mathbb{E}[\ln \hat{P}_i] - 1 - \lambda) \quad \text{(I)}
\end{aligned}$$

$$\frac{\partial \mathcal{L}(R_i, \lambda)}{\partial \lambda} = \sum_{i=1}^{C} R_i - 1 \stackrel{!}{=} 0$$
$$\Rightarrow \sum_{i=1}^{C} R_i = 1 \quad \text{(II)}$$

Plugging (I) into (II), we get:

$$\begin{aligned}
&\sum_{j=1}^{C} \exp(\mathbb{E}[\ln \hat{P}_j] - 1 - \lambda) = 1 \\
\Rightarrow\ & e^{-\lambda} \sum_{j=1}^{C} \exp(\mathbb{E}[\ln \hat{P}_j] - 1) = 1 \\
\Rightarrow\ & -\lambda + \ln \sum_{j=1}^{C} \exp(\mathbb{E}[\ln \hat{P}_j] - 1) = 0 \\
\Rightarrow\ & \lambda = \ln \sum_{j=1}^{C} \exp(\mathbb{E}[\ln \hat{P}_j] - 1) = \ln \sum_{j=1}^{C} \exp \mathbb{E}[\ln \hat{P}_j] - 1
\end{aligned}$$

Plugging λ into (I) results in:

$$\begin{aligned}
R_i &= \exp\left(\mathbb{E}[\ln \hat{P}_i] - 1 - \ln \sum_{j=1}^{C} \exp \mathbb{E}[\ln \hat{P}_j] + 1\right) \\
&= \exp\left(\mathbb{E}[\ln \hat{P}_i] - \ln \sum_{j=1}^{C} \exp \mathbb{E}[\ln \hat{P}_j]\right) \\
&= \frac{\exp \mathbb{E}[\ln \hat{P}_i]}{\sum_{j=1}^{C} \exp \mathbb{E}[\ln \hat{P}_j]} \quad \forall\, 1 \leq i \leq C
\end{aligned}$$

We obtain an unique minimum, since the Lagrange function is convex on $R \in [0,1]^C$.

b)

Prove the bias-variance decomposition

$$Error(\hat{P}) = Bias(\hat{P}) + Var(\hat{P})$$

where the error, bias and variance are given by

$$\text{Error}(\hat{P}) = \mathbb{E}[D_{KL}(P||\hat{P})], \qquad \text{Bias}(\hat{P}) = D_{KL}(P||R), \qquad \text{Var}(\hat{P}) = \mathbb{E}[D_{KL}(R||\hat{P})].$$

$$
\begin{aligned}
\text{Error}(\hat{P}) &= \mathbb{E}[D_{KL}(P||\hat{P})] = \mathbb{E}\left[\sum_{i=1}^{C} P_i \ln\left(\frac{P_i}{\hat{P}_i}\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{C} P_i \ln\left(\frac{P_i}{\hat{P}_i}\cdot\frac{R_i}{R_i}\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{C} P_i \left(\ln\frac{P_i}{R_i} + \ln\frac{R_i}{\hat{P}_i}\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{C} P_i \ln\frac{P_i}{R_i} + \sum_{i=1}^{C} P_i \ln\frac{R_i}{\hat{P}_i}\right] \\
&= D_{KL}(P||R) + \sum_{i=1}^{C} P_i(\ln R_i - \mathbb{E}[\ln \hat{P}_i]) \\
&\overset{(a)}{=} D_{KL}(P||R) + \sum_{i=1}^{C} P_i \left[\ln\left(\frac{\exp\mathbb{E}[\ln\hat{P}_i]}{\sum_{j=1}^{C}\exp\mathbb{E}[\ln\hat{P}_j]}\right) - \mathbb{E}[\ln\hat{P}_i]\right] \\
&= D_{KL}(P||R) + \sum_{i=1}^{C} P_i \left(\mathbb{E}[\ln\hat{P}_i] - \ln\sum_{j=1}^{C}\exp\mathbb{E}[\ln\hat{P}_j] - \mathbb{E}[\ln\hat{P}_i]\right) \\
&= D_{KL}(P||R) - \left(\sum_{i=1}^{C} P_i\right)\left(\ln\sum_{j=1}^{C}\exp\mathbb{E}[\ln\hat{P}_j]\right) \\
&= D_{KL}(P||R) - 1\cdot\left(\ln\sum_{j=1}^{C}\exp\mathbb{E}[\ln\hat{P}_j]\right) \\
&= D_{KL}(P||R) - \left(\sum_{i=1}^{C} R_i\right)\left(\ln\sum_{j=1}^{C}\exp\mathbb{E}[\ln\hat{P}_j]\right) \\
&= D_{KL}(P||R) + \sum_{i=1}^{C} R_i \left(\mathbb{E}[\ln\hat{P}_i] - \ln\sum_{j=1}^{C}\exp\mathbb{E}[\ln\hat{P}_j] - \mathbb{E}[\ln\hat{P}_i]\right) \\
&= D_{KL}(P||R) + \sum_{i=1}^{C} R_i \left[\ln\left(\frac{\exp\mathbb{E}[\ln\hat{P}_i]}{\sum_{j=1}^{C}\exp\mathbb{E}[\ln\hat{P}_j]}\right) - \mathbb{E}[\ln\hat{P}_i]\right] \\
&= D_{KL}(P||R) + \sum_{i=1}^{C} R_i(\ln R_i - \mathbb{E}[\ln\hat{P}_i]) \\
&= D_{KL}(P||R) + \mathbb{E}\left[\sum_{i=1}^{C} R_i \ln\left(\frac{R_i}{\hat{P}_i}\right)\right] \\
&= D_{KL}(P||R) + \mathbb{E}[D_{KL}(R||\hat{P})] = Bias(\hat{P}) + Var(\hat{P})
\end{aligned}
$$

A kernel function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ generalizes the linear scalar product between two vectors and allows to treat non-linearities in data by linear means. In order for that to work, the kernel function must fulfill certain properties which are shared by the scalar product. In particular, it must satisfy the *Mercer's condition*, which verifies that for any sequence of data points $x_1, ..., x_n \in \mathbb{R}^d$ and coefficients $c_1, ..., c_n \in \mathbb{R}$ the inequality

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

is satisfied. If it is the case, the kernel is called a *Mercer kernel*.

Conversely, the *representer theorem* states that if k is a Mercer kernel on $\mathbb{R}^d$, then there exists a Hilbert space (i.e., a finite or infinite dimensional $\mathbb{R}$-vector space with norm and scalar product) $\mathcal{F}$, the so-called feature space, and a continuous map $\varphi : \mathbb{R}^d \to \mathcal{F}$, such that

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \quad \forall\, x, x' \in \mathbb{R}$$

I.e., the kernel k implicitly allows to calculate scalar products in the feature space $\mathcal{F}$ without need for an algorithm to calculate φ. This scalar product form can also be useful as an alternate representation (φ-representation) of Mercer kernels, in particular, when showing whether a particular composition of Mercer kernels is or isn't a Mercer kernel itself.

Exercise 1: Kernology

b)

Show that

i. $k(x, x') = a \qquad a \in \mathbb{R}^+$,

ii. $k(x, x') = \langle x, x' \rangle$,

iii. $k(x, x') = f(x) \cdot f(x')$ where $f : \mathbb{R}^d \to \mathbb{R}$ is an arbitrary continuous function,

are Mercer kernels.

---

i.)

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j a = a\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \\ = a\left(\sum_{i=1}^{n} c_i\right)\left(\sum_{j=1}^{n} c_j\right) = a\left(\sum_{i=1}^{n} c_i\right)^2 \geq 0\end{aligned}$$

ii.)

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle x_i, x_j \rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j x_i x_j$$
$$= \left(\sum_{i=1}^{n} c_i x_i\right)\left(\sum_{j=1}^{n} c_j x_j\right) = \left(\sum_{i=1}^{n} c_i x_i\right)^2 \geq 0$$

iii.)

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j f(x_i) f(x_j)$$
$$= \left(\sum_{i=1}^{n} c_i f(x_i)\right)\left(\sum_{i=1}^{n} c_j f(x_j)\right) = \left(\sum_{i=1}^{n} c_i f(x_i)\right)^2 \geq 0$$

b)

Let k_1, k_2 be two Mercer kernels. Show that

i. $k(x, x') = k_1(x, x') + k_2(x, x')$,
ii. $k(x, x') = k_1(x, x') \cdot k_2(x, x')$

are again Mercer kernels.

---

i.)

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \left[k_1(x_i, x_j) + k_2(x_i, x_j)\right]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_1(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_2(x_i, x_j) \geq 0$$

since we know that $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_1(x_i, x_j) \geq 0$ and $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_2(x_i, x_j) \geq 0$.

ii.)

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_1(x_i, x_j) \cdot k_2(x_i, x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle \varphi_1(x_i), \varphi_1(x_j)\rangle_{\mathcal{F}} \langle \varphi_2(x_i), \varphi_2(x_j)\rangle_{\mathcal{F}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \varphi_1(x_i)\varphi_1(x_j)\varphi_2(x_i)\varphi_2(x_j) \\
&= \left(\sum_{i=1}^{n} c_i \varphi_1(x_i)\varphi_2(x_i)\right)\left(\sum_{j=1}^{n} c_j \varphi_1(x_j)\varphi_2(x_j)\right) \\
&= \left(\sum_{i=1}^{n} c_i \varphi_1(x_i)\varphi_2(x_i)\right)^2 \geq 0
\end{aligned}$$

c)

Show using the results above that the polynomial kernel of degree d, where $k(x, x') = (\langle x, x'\rangle + \vartheta)^d$ and $\vartheta \in \mathbb{R}^+$, is a Mercer kernel.

There are two ways to show that the polynomial of degree d is a kernel.

1) The simple solution is a combination of what we already know. From 1.a)ii) We know that $\langle x, x'\rangle$ is a kernel and from 1.a)ii) we know that ϑ is also a kernel. Using the result from 1.b)i) we can conclude that also the sum of kernels $\langle x, x'\rangle + \vartheta$ is a kernel. Last we know from 1.b)ii) that a product of kernels is also a kernel which proofs that $(\langle x, x'\rangle + \vartheta)^d$ is a kernel.

2) The other solutions follows by calculating (for $d = 2$):

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (x_i x_j + \vartheta)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (x_i^2 x_j^2 + 2\vartheta x_i x_j + \vartheta^2) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j x_i^2 x_j^2 + 2\vartheta \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j x_i x_j + \vartheta^2 \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \\
&= \left(\sum_{i=1}^{n} c_i x_i^2\right)\left(\sum_{j=1}^{n} c_j x_j^2\right) + 2\vartheta \left(\sum_{i=1}^{n} c_i x_i\right)\left(\sum_{j=1}^{n} c_j x_j\right) + \vartheta^2 \left(\sum_{i=1}^{n} c_i\right)\left(\sum_{j=1}^{n} c_j\right) \\
&= \left(\sum_{i=1}^{n} c_i x_i^2\right)^2 + 2\vartheta \left(\sum_{i=1}^{n} c_i x_i\right)^2 + \vartheta^2 \left(\sum_{i=1}^{n} c_i\right)^2 \geq 0
\end{aligned}$$

If we just use *multinomial theorem*, it can be shown that it holds for any d.

d)

Show using the results above that the Gaussian kernel of width σ, where $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ is a Mercer kernel.

First we need to show that $\exp(k(x, x'))$ is a kernel ($\star$), using Taylor expansion:

$$\exp(k(x, x')) = \sum_{m=1}^{\infty} \frac{1}{m!} k(x, x')^m$$

Since we know from 1.b)ii) that products of kernels are kernels again, we can conclude that $k(x, x')^m$ is a kernel. Using 1.b)i) we can see that also the sum is a kernel again and therefore $\exp(k(x, x'))$ is a Mercer kernel.

In the next step we reshape the given function:

$$\begin{aligned}
\exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) &= \exp\left(-\frac{(x-x')^T(x-x')}{2\sigma^2}\right) \\
&= \exp\left(-\frac{x^Tx - 2x^Tx' + x'^Tx'}{2\sigma^2}\right) \\
&= \exp\left(-\frac{x^Tx}{2\sigma^2} + \frac{x^Tx'}{\sigma^2} - \frac{x'^Tx'}{2\sigma^2}\right) \\
&= \exp\left(-\frac{x^Tx}{2\sigma^2}\right) \exp\left(-\frac{x'^Tx'}{2\sigma^2}\right) \exp\left(\frac{x^Tx'}{\sigma^2}\right) \\
&= f(x)f(x') \cdot g(x, x') \\
&\overset{(\star)}{=} f(x)f(x') \cdot k_2(x, x') \\
&\overset{(1.a)iii))}{=} k_1(x, x') \cdot k_2(x, x') \\
&\overset{(1.a)ii))}{=} k(x, x')
\end{aligned}$$

Using the results from above, we could easily show that the Gaussian kernel is a Mercer kernel.

Exercise 2: The Feature Map

Consider the homogenous polynomial kernel k of degree 2 which is $k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, where

$$k(x, y) = \langle x, y \rangle^2 = \left(\sum_{i=1}^{d} x_i y_i\right)^2 .$$

a)

Show that $\mathcal{F} = \mathbb{R}^3$ and $\varphi : \mathbb{R}^3 \to \mathbb{R}^2$ where

$$\varphi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

are possible choices for feature space and feature map.

We have to check if $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{F}}$ holds.

$$\begin{aligned} \langle \varphi(x), \varphi(y) \rangle_{\mathcal{F}} = \langle \varphi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \varphi \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \rangle_{\mathcal{F}} &= \left\langle \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}, \begin{pmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{pmatrix} \right\rangle_{\mathcal{F}} \\ &= x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 \\ &= (x_1y_1 + x_2y_2)^2 = \left(\sum_{i=1}^{2} x_iy_i \right)^2 \\ &= \langle x, y \rangle^2 = k(x, y) \end{aligned}$$

b)

Determine (i.e., give an explicit description of) the images of

i The unit circle $C = \{ \binom{\cos(\theta)}{\sin(\theta)}; 0 \leq \theta < 2\pi \}$

ii The plane $A = \{ \binom{t}{s}; t, s \in \mathbb{R} \}$

under the feature map φ in the feature space $\mathcal{F}$.

---

i)

$$\varphi(C) = \varphi \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta)^2 \\ \sqrt{2}\cos(\theta)\sin(\theta) \\ \sin(\theta)^2 \end{pmatrix}, \qquad \text{where } 0 \leq \theta < 2\pi$$

ii)

$$\varphi(A) = \varphi\begin{pmatrix} t \\ s \end{pmatrix} = \begin{pmatrix} t^2 \\ \sqrt{2}ts \\ s^2 \end{pmatrix}, \qquad \text{where } s, t \in \mathbb{R}$$

c)

The image $\varphi(C)$ lies on a plane H in $\mathbb{R}^3$. *Characterize* that plane.

Recall that $\cos(\theta)^2 + \sin(\theta)^2 = 1$ and $2\sin(\theta)\cos(\theta) = \sin(2\theta)$ and define $v = \cos(\theta)^2$ and $w = \frac{\sqrt{2}}{2}\sin(2\theta)$:

$$\begin{pmatrix} \cos(\theta)^2 \\ \sqrt{2}\cos(\theta)\sin(\theta) \\ \sin(\theta)^2 \end{pmatrix} = \begin{pmatrix} \cos(\theta)^2 \\ \frac{\sqrt{2}}{2}\sin(2\theta) \\ 1 - \cos(\theta)^2 \end{pmatrix} = \begin{pmatrix} v \\ w \\ 1 - v \end{pmatrix}$$

$\Rightarrow$?

d)

Find a point P in $\mathcal{F}$ which is not contained in $\varphi(A)$.

We can show that

$$P = \begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

is not contained in $\varphi(A)$. We chose $P_x = t^2 = 1$, than it holds that $t = 1$ and analog we have given that $P_z = s^2 = 1$, where follows that $s = 1$. Therefore in $\varphi(A)$ it should hold that $P_y = \sqrt{2}st$, but instead $\sqrt{2}st = \sqrt{2} \neq 0 = P_y$.

Exercise 1: The Dual SVM

The primal program for the linear hard margin SVM is

$$\min_{\boldsymbol{w},b} \|\boldsymbol{w}\|^2$$

$$\text{subject to } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \theta) \geq 1, \quad \text{for } 1 \leq i \leq n$$

where $\|\cdot\|$ denotes the Euclidean norm, and the minimization is performed in $\boldsymbol{w} \in \mathbb{R}^d$, $\theta \in \mathbb{R}$, while the data $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ are regarded as fixed constants.

a)

Write down the Lagrange function $\Lambda(\boldsymbol{w}, \theta, \boldsymbol{\alpha})$ for the hard margin SVM, where $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the vector of Lagrange multipliers.

$$\begin{aligned}\Lambda(\boldsymbol{w}, \theta, \boldsymbol{\alpha}) &= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \theta) - 1\right] \\ &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{n} \alpha_i \left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \theta) - 1\right]\end{aligned}$$

b)

Derive the Lagrange dual of the program above. Describe how the solution for the primal program can be obtained from a solution of the dual program. (Hint: The Lagrange dual is a maximization problem in $\boldsymbol{\alpha}$.)

The derivation is given by:

$$\frac{\partial \Lambda(\boldsymbol{w}, \theta, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \tag{1}$$

$$\frac{\partial \Lambda(\boldsymbol{w}, \theta, \boldsymbol{\alpha})}{\partial \theta} = -\sum_{i=1}^{n} \alpha_i y_i \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = \boldsymbol{\alpha}^T\boldsymbol{y} = 0 \tag{2}$$

We can evaluate $\|\boldsymbol{w}\|^2$ using (1) as follows:

$$\begin{aligned}\|\boldsymbol{w}\|^2 = \boldsymbol{w}^T\boldsymbol{w} &= \left(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i\right)^T \left(\sum_{j=1}^{n} \alpha_j y_j \boldsymbol{x}_j\right) \\ &= \left(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i^T\right) \left(\sum_{j=1}^{n} \alpha_j y_j \boldsymbol{x}_j\right) = \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j\end{aligned} \tag{3}$$

Plugging (3) and (1) into the Lagrange function gives:

$$\begin{aligned}
\Lambda(\boldsymbol{\alpha}) &= \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j - \sum_{i=1}^n \alpha_i\left[y_i\left(\left(\sum_{j=1}^n \alpha_j y_j \boldsymbol{x}_j\right)^T \boldsymbol{x}_i + \theta\right) - 1\right] \\
&= \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j - \left(\sum_{j=1}^n \alpha_j y_j \boldsymbol{x}_j^T\right)\left(\sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i\right) - \theta\sum_{i=1}^n \alpha_i y_i + \sum_{i=1}^n \alpha_i \\
&\overset{(2)}{=} \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j - \sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j + \sum_{i=1}^n \alpha_i \\
&= -\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j + \sum_{i=1}^n \alpha_i
\end{aligned}$$

The dual solution is now given by:

$$\max_{\boldsymbol{\alpha}} \Lambda(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}}\left[-\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j + \sum_{i=1}^n \alpha_i\right], \quad \text{s.t. } \alpha_i \geq 0 \;\forall\, i = 1, ..., n$$

c)

Write down the kernelized versions of the primal program and of the dual program.

Kernelized version of the primal program:

$$\min_{\boldsymbol{w},\theta} \|\boldsymbol{w}\|^2, \quad \text{s.t. } y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + \theta) \geq 1$$

Kernelized version of the dual program:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}}\left[-\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j) + \sum_{i=1}^n \alpha_i\right] \\
&= \max_{\boldsymbol{\alpha}}\left[-\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \underbrace{k(\boldsymbol{x}_i, \boldsymbol{x}_j)}_{k_{ij}} + \sum_{i=1}^n \alpha_i\right], \quad \text{s.t. } \alpha_i \geq 0 \;\forall\, i = 1, ..., n
\end{aligned}$$

Exercise 2: SVMs and Quadratic Programming

We consider the `CVXOPT` Python software for convex optimization. The method `cvxopt.solvers.qp` solves the quadratic optimization problem given in matrix form as:

$$\begin{aligned}
&\min_x \frac{1}{2}\boldsymbol{x}^T P\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} \\
&\text{subject to } G\boldsymbol{x} \preccurlyeq \boldsymbol{h} \\
&\text{and } A\boldsymbol{x} = \boldsymbol{b}
\end{aligned}$$

Here, $\preccurlyeq$ denotes the element-wise inequality: $(\boldsymbol{h} \preccurlyeq \boldsymbol{h}') \Leftrightarrow (\forall i : h_i \leq h'_i)$. Note that the meaning of the variable $\boldsymbol{x}$ is different from that of the same variable in the previous exercise.

a)

We can formulate the quadratic optimization problem as follows:

$$\min_x \left(\frac{1}{2}\boldsymbol{x}^T P \boldsymbol{x} + \boldsymbol{q}^T \boldsymbol{x}\right) = \max_x \left(-\frac{1}{2}\sum_i \sum_j \underbrace{x_i^T}_{\alpha_i^T} \underbrace{P_{i,j}}_{y_i k_{ij} y_j} \underbrace{x_j}_{\alpha_j} + \sum_i \underbrace{q_i}_{-1} \underbrace{x_i}_{\alpha_i}\right)$$

Furthermore we can see that

$$\underbrace{G}_{-I}\underbrace{\boldsymbol{x}}_{\boldsymbol{\alpha}} =\preccurlyeq \underbrace{\boldsymbol{h}}_{\mathbf{0}} \Leftrightarrow -I\boldsymbol{\alpha} \preccurlyeq \mathbf{0} \Leftrightarrow \alpha_i \geq 0 \; \forall \; i = 1, ..., n$$

and

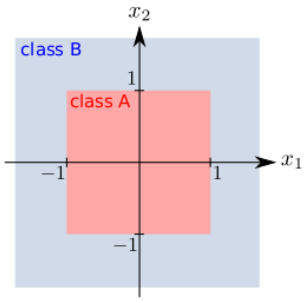
$$\underbrace{A}_{\boldsymbol{y}^T}\underbrace{\boldsymbol{x}}_{\boldsymbol{\alpha}} = \underbrace{\boldsymbol{b}}_{\mathbf{0}} \Leftrightarrow \boldsymbol{y}^T\boldsymbol{\alpha} = \sum_i \alpha_i y_i = 0.$$

Therefore we can identify $\boldsymbol{x} = \boldsymbol{\alpha}$ and

$$\begin{aligned} &P_{ij} = y_i k_{ij} y_j, \; \boldsymbol{q} = -\mathbf{1}, \\ &G = -I, \; \boldsymbol{h} = \mathbf{0}, \\ &A = \boldsymbol{y}^T, \; \boldsymbol{b} = \mathbf{0}. \end{aligned}$$

Exercise 1: Designing a Neural Network

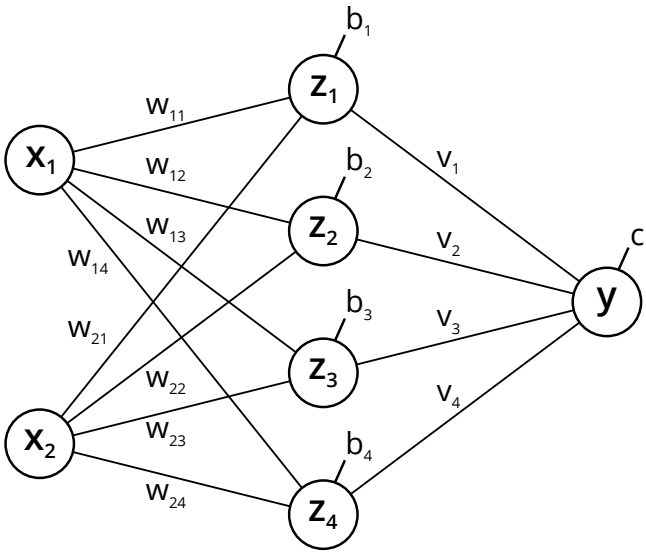
We would like to implement a neural network that classifies data points in $\mathbb{R}^2$ according to the following decision boundary:



Note that the area for class B stretches to the infinity. We consider as an elementary computation the threshold neuron whose relation between inputs $\{z_i\}$ and output z_j is given by

$$z_j = \begin{cases} 1 & \text{if} \quad \sum_i z_i w_{ij} + b_j > 0 \\ 0 & \text{if} \quad \sum_i z_i w_{ij} + b_j \leq 0. \end{cases}$$

In a similar way the XOR problem was solved in the slides, design at hand a neural network that takes x_1 and x_2 as input and produces the output "1" if the input belong to class A, and "0" if the input belongs to class B. Draw the neural network model and write down its corresponding weight and bias parameters.



The *first layer* contains following parameters

$$\boldsymbol{w} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \end{bmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \end{bmatrix},$$

and the *second layer* contains

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} \quad \text{and} \quad \boldsymbol{c} = \begin{bmatrix} c \end{bmatrix}$$

One possible solution would be:

$$\boldsymbol{w} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} -1 & -1 & -1 & -1 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0.5 \end{bmatrix}$$

Another similar solution could be:

$$\boldsymbol{w} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} -0.5 \end{bmatrix}$$

We want to check some points to test the solution (using the first). Let's consider for $\boldsymbol{x} = (x_1, x_2)^T$:

$$\boldsymbol{x}_a = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \quad \boldsymbol{x}_b = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \boldsymbol{x}_c = \begin{bmatrix} -1 \\ 1.1 \end{bmatrix}, \quad \boldsymbol{x}_d = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad \boldsymbol{x}_e = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

a)

$$\begin{aligned}
z_1: &\quad \sum_i x_i w_{i1} + b_1 = 0.4 \cdot 1 + 0.6 \cdot 0 - 1 = -0.6 \leq 0 \quad \Rightarrow\ z_1 = 0 \\
z_2: &\quad \sum_i x_i w_{i2} + b_2 = 0.4 \cdot (-1) + 0.6 \cdot 0 - 1 = -1.4 \leq 0 \quad \Rightarrow\ z_2 = 0 \\
z_3: &\quad \sum_i x_i w_{i3} + b_3 = 0.4 \cdot 0 + 0.6 \cdot 1 - 1 = -0.4 \leq 0 \quad \Rightarrow\ z_3 = 0 \\
z_4: &\quad \sum_i x_i w_{i4} + b_4 = 0.4 \cdot 0 + 0.6 \cdot (-1) - 1 = -1.6 \leq 0 \quad \Rightarrow\ z_4 = 0
\end{aligned}$$

$$y: \quad \sum_i z_i v_i + c = 4 \cdot (0 \cdot (-1)) + 0.5 = 0.5 > 0 \quad \Rightarrow\ y = 1$$

$\Rightarrow \boldsymbol{x}_a$ is inside

b)

$$z_1: \quad \sum_i x_i w_{i1} + b_1 = -1 \cdot 1 + (-1) \cdot 0 - 1 = -2 \leq 0 \quad \Rightarrow \; z_1 = 0$$
$$z_2: \quad \sum_i x_i w_{i2} + b_2 = -1 \cdot (-1) + (-1) \cdot 0 - 1 = 0 \leq 0 \quad \Rightarrow \; z_2 = 0$$
$$z_3: \quad \sum_i x_i w_{i3} + b_3 = -1 \cdot 0 + (-1) \cdot 1 - 1 = -2 \leq 0 \quad \Rightarrow \; z_3 = 0$$
$$z_4: \quad \sum_i x_i w_{i4} + b_4 = -1 \cdot 0 + (-1) \cdot (-1) - 1 = 0 \leq 0 \quad \Rightarrow \; z_4 = 0$$

$$y: \quad \sum_i z_i v_i + c = 4 \cdot (0 \cdot (-1)) + 0.5 = 0.5 > 0 \quad \Rightarrow \; y = 1$$

$\Rightarrow \boldsymbol{x}_b$ is inside

c)

$$z_1: \quad \sum_i x_i w_{i1} + b_1 = -1 \cdot 1 + 1.1 \cdot 0 - 1 = -2 \leq 0 \quad \Rightarrow \; z_1 = 0$$
$$z_2: \quad \sum_i x_i w_{i2} + b_2 = -1 \cdot (-1) + 1.1 \cdot 0 - 1 = 0 \leq 0 \quad \Rightarrow \; z_2 = 0$$
$$z_3: \quad \sum_i x_i w_{i3} + b_3 = -1 \cdot 0 + 1.1 \cdot 1 - 1 = 0.1 > 0 \quad \Rightarrow \; z_3 = 1$$
$$z_4: \quad \sum_i x_i w_{i4} + b_4 = -1 \cdot 0 + 1.1 \cdot (-1) - 1 = -2.1 \leq 0 \quad \Rightarrow \; z_4 = 0$$

$$y: \quad \sum_i z_i v_i + c = 1 \cdot (-1) + 3 \cdot (0 \cdot (-1)) + 0.5 = -0.5 \leq 0 \quad \Rightarrow \; y = 0$$

$\Rightarrow \boldsymbol{x}_c$ is outside

Example d) and e) are solved using matrix notation:

d)

$$\boldsymbol{x}_d^T \boldsymbol{w} + \boldsymbol{b} = \begin{bmatrix} 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} + \begin{bmatrix} -1 & -1 & -1 & -1 \end{bmatrix} = \begin{bmatrix} -1 & -1 & 1 & -3 \end{bmatrix}$$

$$\Rightarrow \boldsymbol{z} = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}^T$$

$$\boldsymbol{z}^T\boldsymbol{v} + \boldsymbol{c} = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} + 0.5 = -0.5 \quad \Rightarrow y = 0$$

$\Rightarrow \boldsymbol{x}_d$ is outside

e)

$$\boldsymbol{x}_e^T\boldsymbol{w} + \boldsymbol{b} = \begin{bmatrix} -2 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} + \begin{bmatrix} -1 & -1 & -1 & -1 \end{bmatrix} = \begin{bmatrix} -3 & 1 & 1 & -3 \end{bmatrix}$$

$$\Rightarrow \boldsymbol{z} = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^T$$

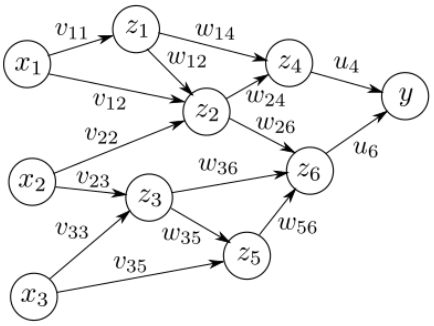
$$\boldsymbol{z}^T\boldsymbol{v} + \boldsymbol{c} = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} + 0.5 = -1.5 \quad \Rightarrow y = 0$$

$\Rightarrow \boldsymbol{x}_e$ is outside

Remark: Since z_j is a step function it is not possible to differentiate it. Therefore the parameters cannot be obtained by backpropagation.

Exercise 2: Backward Computations

We consider a neural network with the following structure:



The elementary computation of this network is the *sigmoid neuron* defined as:

$$g(a_j) = \frac{e^{a_j}}{1 + e^{a_j}} \quad \text{where } a_j = \sum_i z_i w_{ij}$$

Examples of forward computations are:

$$\begin{aligned} z_5 &= g(x_3 \cdot v_{35} + z_3 \cdot w_{35}) \\ z_6 &= g(z_2 \cdot w_{26} + z_3 \cdot w_{36} + z_5 \cdot w_{56}) \\ y &= g(z_4 \cdot u_4 + z_6 \cdot u_6) \end{aligned}$$

Assuming that we have computed the activation of each neuron and the output, and knowing the error gradient $\frac{\partial E}{\partial y}$, *write* the sequence of computations that lead to the evaluation of the error gradient $\frac{\partial E}{\partial w_{12}}$. *Hint*: The derivative of the sigmoid function $g(a_j)$ can be expressed in terms of neuron activations as

$$\frac{\partial}{\partial a_j} \underbrace{g(a_j)}_{z_j} = \underbrace{g(a_j)}_{z_j}(1 - \underbrace{g(a_j)}_{z_j}).$$

We start the backpropagation at the end of the network. The error $E(y) = \frac{1}{2}(y-t)^2$ is given with $y = g(a_y)$, which results in $E(a_y) = \frac{1}{2}(g(a_y) - t)^2$, so we have:

$$\frac{\partial E}{\partial a_y} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial a_y} = \frac{\partial E}{\partial y}\frac{\partial g(a_y)}{\partial a_y} = \frac{\partial E}{\partial y} g(a_y)(1 - g(a_y)) = \frac{\partial E}{\partial y} y(1-y) \tag{1}$$

y depends on two value, namely $z_4 = g(a_4)$ and $z_6 = g(a_6)$. We also know that $y = g(a_y)$, where $a_y = z_4 u_4 + z_6 u_6$, so we can obtain

$$\frac{\partial E}{\partial z_4} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial z_4} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial a_y}\frac{\partial a_y}{z_4} \overset{(1)}{=} \frac{\partial E}{\partial a_y}\frac{\partial}{\partial z_4}(z_4 u_4 + z_6 u_6) = \frac{\partial E}{\partial a_y} u_4 \tag{2}$$

$$\frac{\partial E}{\partial z_6} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial z_6} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial a_y}\frac{\partial a_y}{z_6} \overset{(1)}{=} \frac{\partial E}{\partial a_y}\frac{\partial}{\partial z_6}(z_4 u_4 + z_6 u_6) = \frac{\partial E}{\partial a_y} u_6 \tag{3}$$

where $\frac{\partial E}{\partial a_y}$ was already solved in equation (1). We want to get a solution for $z_2 = g(a_2)$, where we know that $a_4 = z_1 w_{14} + z_2 w_{24}$ and $a_6 = z_2 w_{26} + z_3 w_{36} + z_5 w_{56}$. We can derive

$$\begin{aligned} \frac{\partial E}{\partial z_2} &= \frac{\partial E}{\partial z_4}\frac{\partial z_4}{\partial z_2} + \frac{\partial E}{\partial z_6}\frac{\partial z_6}{\partial z_2} \\ &= \frac{\partial E}{\partial z_4}\frac{\partial z_4}{\partial a_4}\frac{\partial a_4}{\partial z_2} + \frac{\partial E}{\partial z_6}\frac{\partial z_6}{\partial a_6}\frac{\partial a_6}{\partial z_2} \\ &= \frac{\partial E}{\partial z_4} z_4(1-z_4)\frac{\partial}{\partial z_2}(z_1 w_{14} + z_2 w_{24}) + \frac{\partial E}{\partial z_6} z_6(1-z_6)\frac{\partial}{\partial z_2}(z_2 w_{26} + z_3 w_{36} + z_5 w_{56}) \\ &= \frac{\partial E}{\partial z_4} z_4(1-z_4) w_{24} + \frac{\partial E}{\partial z_6} z_6(1-z_6) w_{26}, \end{aligned} \tag{4}$$

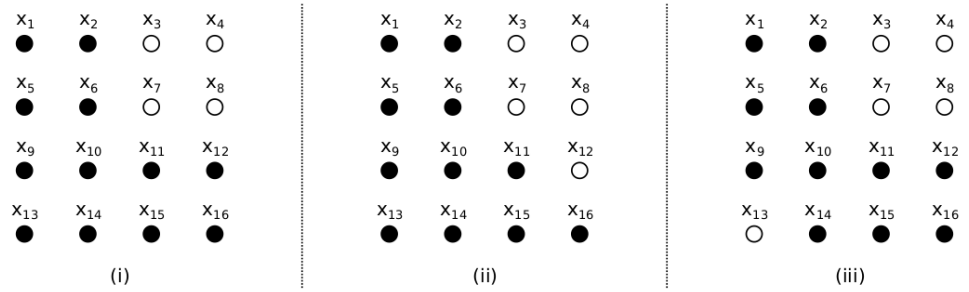
where $\frac{\partial E}{\partial z_4}$ is given by (2) and $\frac{\partial E}{\partial z_6}$ is given by (3). We need $a_2 = z_1 w_{12} + x_1 v_1 + x_2 v_2$ from $z_2 = g(a_2)$ and can finally obtain

$$\frac{\partial E}{\partial w_{12}} = \frac{\partial E}{\partial z_2}\frac{\partial z_2}{\partial w_{12}} = \frac{\partial E}{\partial z_2}\frac{\partial z_2}{\partial a_2}\frac{\partial a_2}{\partial w_{12}} = \frac{\partial E}{\partial z_2} z_2(1-z_2) z_1, \tag{5}$$

where we know $\frac{\partial E}{\partial z_2}$ from (4). We now have derived the parameter w_{12} by backpropagation.

Exercise 1: Weighting the Data

We consider the following three two-dimensional binary classification datasets composed of 16 samples each:



Black circles denote the first class (-1) and white circles denote the second class ($+1$). We decide to use a boosted classifier with a linear soft-margin SVM as weak learner. The boosted classifier is given by the discriminant function

$$f(x) = \alpha_0 + \sum_{t=1}^{T} \alpha_t h_t(x)$$

where $\alpha_0, ..., \alpha_T \in \mathbb{R}$, and where the function

$$h_t(x) = \text{sign}(w_t^T x_i + b_t)$$

returns the classification result (-1 or $+1$) of the t-th weak classifier. It is trained, under the weighting of the data $p_{t,1}, ..., p_{t,16}$, to minimize the SVM objective

$$\min_{w_t, \xi_t, b_t} \frac{1}{2}\|w_t\|^2 + C \sum_{i=1}^{16} p_{t,i}\xi_{t,i}$$

under the constraints

$$\forall_{i=1}^{16} \; : \quad y_i(w_t^T x_i + b_t) \geq 1 - \xi_{t,i}, \quad \xi_{t,i} \geq 0.$$

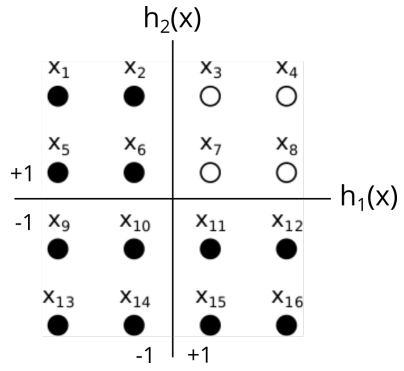
We also assume the parameter C to be large (e.g. $C = 100$).

Determine at hand and for each dataset a possible boosted classifier that classifies the data perfectly. Draw the decision boundary learned by each individual weak learner, and the final decision boundary. Write down the coefficients $\alpha_0, ..., \alpha_T$, and the weighting terms $p_{t,i}$ for each weak learner $1 \leq t \leq T$ and data point $1 \leq i \leq 16$.

We want to classify a point as "white" if $f(x) \geq 0$ and as "black" if $f(x) < 0$ and we define

$$E = \frac{1}{2}\|w_t\|^2 + C \sum_{i=1}^{16} p_{t,i}\xi_{t,i}$$

i)



We define two classifier $h_1(x)$ and $h_2(x)$. Since we use SVM with soft margin, we use slack variables $\xi_{t,i}$. The weights $p_{t,i}$ for the two classifier are chosen as

$$p_{1,i} = \begin{cases} 1 & i = 7 \\ 1 & i = 11 \\ 0 & \text{else} \end{cases} \qquad p_{2,i} = \begin{cases} 1 & i = 6 \\ 1 & i = 7 \\ 0 & \text{else} \end{cases},$$

so have to minimize the following functions:

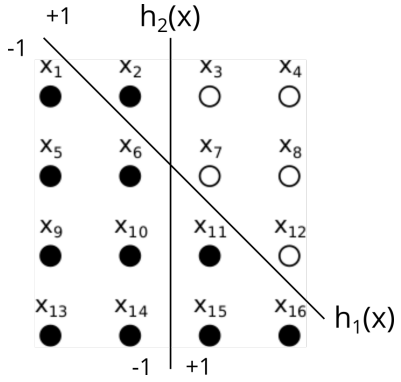
$$E_1 = \frac{1}{2}\|w_1\|^2 + C(\xi_{1,7} + \xi_{1,11})$$
$$E_2 = \frac{1}{2}\|w_2\|^2 + C(\xi_{2,6} + \xi_{2,7})$$

Assuming that we get functions $h_1(x)$ and $h_2(x)$ we get $+1$ for values on one side and -1 for values on the other side, as shown in the figure. Choosing $\alpha_0 = -0.5$, $\alpha_1 = 1$ and $\alpha_2 = 1$, we can check if our classifier will work correctly:

- Top right: $f(x) = \alpha_0 + \sum_{t=1}^{2} \alpha_t h_t(x) = -0.5 + 1 \cdot 1 + 1 \cdot 1 = 1.5 \geq 0 \Rightarrow$ "white"
- Top left: $f(x) = -0.5 + 1 \cdot 1 + 1 \cdot (-1) = -0.5 < 0 \Rightarrow$ "black"
- Bottom right: $f(x) = -0.5 + 1 \cdot (-1) + 1 \cdot 1 = -0.5 < 0 \Rightarrow$ "black"
- Bottom left: $f(x) = -0.5 + 1 \cdot (-1) + 1 \cdot (-1) = -2.5 < 0 \Rightarrow$ "black"

We can conclude that the classification is correct.

ii)



Again, we define two classifier $h_1(x)$ and $h_2(x)$. The weights $p_{t,i}$ for the two classifier are chosen as

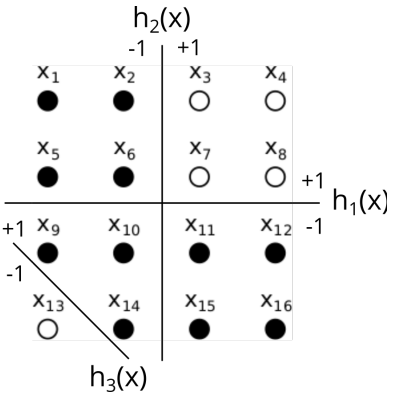
$$p_{1,i} = \begin{cases} 1 & i = 6 \\ 1 & i = 7 \\ 1 & i = 11 \\ 0 & \text{else} \end{cases} \qquad p_{2,i} = \begin{cases} 1 & i = 6 \\ 1 & i = 7 \\ 0 & \text{else} \end{cases},$$

so have to minimize the following functions:

$$E_1 = \frac{1}{2}\|w_1\|^2 + C(\xi_{1,6} + \xi_{1,7} + \xi_{1,11})$$
$$E_2 = \frac{1}{2}\|w_2\|^2 + C(\xi_{2,6} + \xi_{2,7})$$

If we choose $\alpha_0 = 0.5$, $\alpha_1 = 1$ and $\alpha_2 = 1$, as above, we will achieve a correct classification (can be shown analog to i).

iii)



In this task, we define three classifier $h_1(x)$, $h_2(x)$ and $h_3(x)$. The weights $p_{t,i}$ for the classifier are chosen as

$$p_{1,i} = \begin{cases} 1 & i = 7 \\ 1 & i = 11 \\ 0 & \text{else} \end{cases} \qquad p_{2,i} = \begin{cases} 1 & i = 6 \\ 1 & i = 7 \\ 0 & \text{else} \end{cases} \qquad p_{3,i} = \begin{cases} 1 & i = 9 \\ 1 & i = 13 \\ 1 & i = 14 \\ 0 & \text{else} \end{cases},$$

so have to minimize the following functions:

$$\begin{aligned} E_1 &= \frac{1}{2}\|w_1\|^2 + C(\xi_{1,7} + \xi_{1,11}) \\ E_2 &= \frac{1}{2}\|w_2\|^2 + C(\xi_{2,6} + \xi_{2,7}) \\ E_3 &= \frac{1}{2}\|w_3\|^2 + C(\xi_{3,9} + \xi_{3,13} + \xi_{3,14}) \end{aligned}$$

We choose $\alpha_0 = 0.5$, $\alpha_1 = 1$, $\alpha_2 = 1$ and $\alpha_3 = 2$ and we get a perfect classification (which can be shown for all 5 regions, analog to i).

Exercise 2: Boosted Regressors

We consider the boosted regressor

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

where $h_t(x) = w_t x$ is the real-valued prediction produced by the t-th weak regressor and $x \in \mathbb{R}^d$. Assuming a labeled dataset $\{(x_1, y_1), ..., (x_n, y_n)\}$, the boosted regressor is trained to minimize the mean squared error

$$\sum_{i=1}^{n} (y_i - f(x_i))^2, \tag{1}$$

where each weak regressor minimizes the following weighted objective function

$$\sum_{i=1}^{n} p_{t,i}(y_i - h_t(x_i))^2.$$

Show that a single weak regressor can be made as accurate as the boosted regressor when using an appropriate weighting $\{p_1, ..., p_n\}$. Write down one possible weighting for the single regressor that leads to the same accuracy as the optimal regressor in the sense of Equation (1).

Using the weak regressors $h_t(x) = w_t^T x$, the boosted regressor can be written as

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) = \underbrace{\sum_{t=1}^{T} \alpha_t w_t^T}_{v_t} x = v_t x,$$

where we can use (1) to obtain the error:

$$E(v) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - v^T x_i)^2.$$

Now we have a look at the error of a single weak regressor:

$$E_t(w) = \sum_{i=1}^{n} p_{t,i}(y_i - h_t(x_i))^2 = \sum_{i=1}^{n} p_{t,i}(y_i - w_t^T x_i)^2$$

If we identify $v = w$ and if we choose $p_{t,i} = 1$ for all t and i, we can see that $E(v) = E_t(w)$. It means that in this case a single weak regressor $h_t(x)$ is as accurate as the boosted regressor $f(x)$.

Exercise 1: Kernel Ridge Regression

In the lecture, ridge regression was stated as the regularized quadratic program

$$\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2,$$

where $w \in \mathbb{R}^d$ is optimized and $x_1, ..., x_n \in \mathbb{R}^d$ are the data points, and $y_1, ..., y_n \in \mathbb{R}$ are the labels.

a)

Following the strategy outlined in the lecture slides, *give* an explicit formula for the solution of the above program.

We start defining the Q as the part we want to minimize and we transform as follows (where vectors are bold and matrices are big letters):

$$\begin{aligned}
Q &= (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 + \lambda \|\boldsymbol{w}\|_2^2 \\
&= (X\boldsymbol{w} - \boldsymbol{y})^2 + \lambda \|\boldsymbol{w}\|_2^2 \\
&= (X\boldsymbol{w} - \boldsymbol{y})^T (X\boldsymbol{w} - \boldsymbol{y}) + \lambda \boldsymbol{w}^T \boldsymbol{w} \\
&= (\boldsymbol{w}^T X^T - \boldsymbol{y}^T)(X\boldsymbol{w} - \boldsymbol{y}) + \lambda \boldsymbol{w}^T \boldsymbol{w} \\
&= \boldsymbol{w}^T X^T X \boldsymbol{w} - \boldsymbol{w}^T X^T \boldsymbol{y} - \boldsymbol{y} X \boldsymbol{w} + \boldsymbol{y}^T \boldsymbol{y} + \lambda \boldsymbol{w}^T \boldsymbol{w} \\
&= \boldsymbol{w}^T X^T X \boldsymbol{w} - 2\boldsymbol{w}^T X^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{y} + \lambda \boldsymbol{w}^T \boldsymbol{w}
\end{aligned}$$

We now can obtain the minimum like:

$$\begin{aligned}
\min_{\boldsymbol{w}} Q = \frac{\partial Q}{\partial \boldsymbol{w}} &= \frac{\partial}{\partial \boldsymbol{w}} \boldsymbol{w}^T X^T X \boldsymbol{w} - 2\boldsymbol{w}^T X^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{y} + \lambda \boldsymbol{w}^T \boldsymbol{w} \\
&= 2X^T X \boldsymbol{w} - 2X^T \boldsymbol{y} + 2\lambda \boldsymbol{w} \stackrel{!}{=} 0 \\
\Rightarrow \quad & (X^T X + \lambda I_d)\boldsymbol{w} = X^T \boldsymbol{y} \\
\Rightarrow \quad & \hat{\boldsymbol{w}} = (X^T X + \lambda I_d)^{-1} X^T \boldsymbol{y}
\end{aligned}$$

b)

Kernelize the ridge regression model, following the strategy outlined in the lecture slides.

We replace X by the feature map $\Phi = \Phi(X)$ and X^T by $\Phi^T = \Phi^T(X)$. So we obtain

$$\hat{\boldsymbol{w}} = (\Phi^T \Phi + \lambda I_d)^{-1} \Phi^T \boldsymbol{y} \tag{1}$$

First we show an equation we will use below:

$$(\Phi^T \Phi + \lambda I_d)\Phi^T = \Phi^T \Phi \Phi^T + \lambda \Phi^T = \Phi^T (\Phi \Phi^T + \lambda I_n)$$

Multiplying $(\Phi^T\Phi + \lambda I_d)^{-1}$ at the left results in

$$\Phi^T = (\Phi^T\Phi + \lambda I_d)^{-1}\Phi^T(\Phi\Phi^T + \lambda I_n)$$

and multiplying $(\Phi\Phi^T + \lambda I_n)^{-1}\boldsymbol{y}$ at the right gives

$$\Phi^T(\Phi\Phi^T + \lambda I_n)^{-1}\boldsymbol{y} = (\Phi^T\Phi + \lambda I_d)^{-1}\Phi^T\boldsymbol{y}. \tag{2}$$

We can now predict a new point $\hat{y}_* = \hat{\boldsymbol{w}}^T\Phi(\boldsymbol{x}_*) = \Phi(\boldsymbol{x})^T\hat{\boldsymbol{w}}$, where $\Phi(\boldsymbol{x}_*)$ is a feature map of a new feature to predict and $\hat{y}$ is the prediction. Equation (1) gives us the estimated parameters $\hat{\boldsymbol{w}}$. We can derive the prediction starting with

$$\begin{aligned}
\hat{y}_* &= \Phi(\boldsymbol{x}_*)^T\hat{\boldsymbol{w}} \\
&\overset{(1)}{=} \Phi(\boldsymbol{x}_*)^T(\Phi^T\Phi + \lambda I_d)^{-1}\Phi^T\boldsymbol{y} \\
&\overset{(2)}{=} \Phi(\boldsymbol{x}_*)^T\Phi^T(\Phi\Phi^T + \lambda I_n)^{-1}\boldsymbol{y} \\
&= \kappa^T(K + \lambda I_d)^{-1}\boldsymbol{y},
\end{aligned}$$

where $K = \Phi\Phi^T$ is a kernel matrix of size $n \times n$, measuring the distances between all given points, and $\kappa = \Phi\Phi(\boldsymbol{x}_*)$ is a kernel vector of size n, measuring the distances between the new feature $\Phi(\boldsymbol{x}_*)$ and the given features Φ.

If we define $\alpha_i = (K + \lambda I_d)^{-1}y_i$, we can use an alternative form

$$\hat{y}_* = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_*, \boldsymbol{x}_i),$$

where $k(\boldsymbol{x}_*, \boldsymbol{x}_i)$ is a kernel measuring the distance between the new feature $\boldsymbol{x}_*$ and a given feature $\boldsymbol{x}_i$.

Exercise 2: Lagrange Multipliers

Consider the slightly modified quadratic program

$$\begin{aligned}
&\min_{\xi,w} \sum_{i=1}^{n} \xi_i^2 \\
&\text{subject to } \xi_i = w^T x_i - y_i \text{ for } 1 \le i \le n \text{ and } \|w\|_2^2 \le C,
\end{aligned}$$

where C is a regularization constant.

a)

Calculate the Lagrange dual of this program, and its solution.

The constraint $\|w\|_2^2 \leq C$ is equal to $\|w\|_2^2 - C \leq 0$. Using this condition, the Lagrange function can be obtained:

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}, \eta) &= \sum_{i=1}^{n} \xi_i^2 + \eta(\|\boldsymbol{w}\|_2^2 - C) = \sum_{i=1}^{n} (\boldsymbol{w}\boldsymbol{x}_i - y_i)^2 + \eta(\|\boldsymbol{w}\|_2^2 - C) \\ &= (X\boldsymbol{w} - \boldsymbol{y})^2 + \eta(\|\boldsymbol{w}\|_2^2 - C) \\ &= (X\boldsymbol{w} - \boldsymbol{y})^T (X\boldsymbol{w} - \boldsymbol{y}) + \eta(\boldsymbol{w}^T\boldsymbol{w} - C) \\ &= (\boldsymbol{w}^T X^T - \boldsymbol{y}^T)(X\boldsymbol{w} - \boldsymbol{y}) + \eta(\boldsymbol{w}^T\boldsymbol{w} - C) \\ &= \boldsymbol{w}^T X^T X\boldsymbol{w} - 2\boldsymbol{w}^T X^T \boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}) + \eta(\boldsymbol{w}^T\boldsymbol{w} - C)\end{aligned}$$

The dual program is given by $\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \eta)$. The derivation of the Lagrange function $\mathcal{L}$ regarding $\boldsymbol{w}$ is obtained by

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} &= 2X^T X\boldsymbol{w} - 2X^T\boldsymbol{y} + 2\eta\boldsymbol{w} \stackrel{!}{=} 0 \\ &\Rightarrow (X^T X + \eta I_d)\boldsymbol{w} = X^T\boldsymbol{y} \\ &\Rightarrow \hat{\boldsymbol{w}} = (X^T X + \eta I_d)^{-1} X^T\boldsymbol{y}.\end{aligned}$$

b)

Describe how a solution for the primal program can be found from a solution of the dual.

The dual problem is given by

$$\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \eta) = 2X^T X\boldsymbol{w} - 2X^T\boldsymbol{y} + 2\eta\boldsymbol{w},$$

which solution is given by (a) $\hat{\boldsymbol{w}} = (X^T X + \eta I_d)^{-1} X^T\boldsymbol{y}$. To solve the primal problem

$$\min_{\xi, \boldsymbol{w}} \sum_{i=1}^{n} \xi_i^2,$$

we can plug $\boldsymbol{w}$ into the condition, such that

$$\xi_i = \boldsymbol{w}^T x_i - y_i = ((X^T X + \eta I_d)^{-1} X^T\boldsymbol{y})^T x_i - y_i.$$

c)

Explain how the solutions relate to the original quadratic program and its solutions in ridge regression and kernel ridge regression.

The result of $\hat{\boldsymbol{w}}$ is the same as we had in exercise 1, if we identify $\eta = \lambda$.