Exercise 3

Question 1

Relation between ε and noise variance: We assume that the error in the data is normally distributed with mean μ and variance σ^2. Then choosing $\varepsilon \approx \sigma^2$ is the optimal choice, because if we take a closer look at the formula on slide 16 (lecture 5), we can see that our due to our choise of ε the loss function will accept results which differences are smaller than σ^2. This makes sense, because this is the variance (e.g. the standard deviation).

The assumption of a normal distributed error in the data is, considering a practical case, relatively good as we can see using the central limit theorem or the (strong) law of large numbers.

ε vs. over-/underfitting: We differ the cases

- ε small: The decision boundary is quite hard and the model does not accept a lot of noise in the data, so our model has a tendency to overfitting.
- ε big: The model accepts way more noise and in the edge case (ε being so big, that the real margin is smaller than epsilon) the model accepts the whole data as one class. This a tendency to underfitting.

Question 2

We define $L(w, b) := \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \hat{\xi}^{(i)} + \xi^{(i)}$ and we want to minimize L with respect to the following constraints

$$\begin{aligned} y^{(i)} - (w\Phi(x^{(i)} + b) &\leq \varepsilon + \hat{\xi}^{(i)} \\ (w\Phi(x^{(i)}) + b) - y^{(i)} &\leq \varepsilon + \xi^{(i)} \\ \hat{\xi}^{(i)} &\geq 0; \xi^{(i)} \geq 0, \forall i \in \{1, \dots, m\} \end{aligned}$$

We want to find the dual form of this minimization problem using Lagrange multipliers. Therefore we define for $i = 1, \dots, m$ the function

$$f^{(i)}(w, b) := \begin{pmatrix} y^{(i)} - (w\Phi(x^{(i)}) + b) - \varepsilon - \hat{\xi}^{(i)} \\ (w\Phi(x^{(i)}) + b) - y^{(i)} - \varepsilon - \xi^{(i)} \\ -\hat{\xi}^{(i)} \\ -\xi^{(i)} \end{pmatrix}$$

which, regarding our constraints, should be less or equal to zero for each component. Regarding the partial derivatives, we get

$$\nabla_w f^{(i)} = \begin{pmatrix} -\Phi(x^{(i)}) \\ \Phi(x^{(i)}) \\ 0 \\ 0 \end{pmatrix}, \nabla_b f^{(i)} = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \nabla_{\hat{\xi}^{(j)}} f^{(i)} = \delta_{ij} \cdot (0, -1, 0, 1)^T, \nabla_{\xi^{(j)}} f^{(i)} = (-1, 0, -1, 0)^T.$$

Now, we use Lagrange multipliers to obtain the ansatz

$$F(w, b, ((\lambda_k^{(i)})_{\substack{k=1,\dots,4 \\ i=1,\dots,m}}) = L(w, b) + \sum_{i=1}^m \sum_{k=1}^4 \lambda_k^{(i)} f_k^{(i)}(w, b).$$

We now compute the partial derivatives, set them to zero and therefore obtain

$$
\begin{aligned}
0 &= \nabla_w F = w + \sum_{i=1}^{m} \sum_{k=1}^{4} \lambda_k^{(i)} \nabla_w f_k^{(i)}(w,b) = w + \sum_{i=1}^{m} -\lambda_1^{(i)} \Phi(x^{(i)}) + \lambda_2^{(i)} \Phi(x^{(i)}) \\
0 &= \nabla_b F = \sum_{i=1}^{m} -\lambda_1^{(i)} + \lambda_2^{(i)} \\
0 &= \nabla_{\hat{\xi}^{(j)}} F = C - \lambda_1^{(j)} - \lambda_3^{(j)} \\
0 &= \nabla_{\xi^{(j)}} F = C - \lambda_2^{(j)} - \lambda_4^{(j)}.
\end{aligned}
$$

If we put this back into F, we obtain the dual form. Using the first of the upper equations, we have

$$
\begin{aligned}
w &= \sum_{i=1}^{m} \lambda_1^{(i)} \Phi(x^{(i)}) - \lambda_2^{(i)} \Phi(x^{(i)}) \\
\Rightarrow \frac{1}{2} \|w\|^2 &= \frac{1}{2} w^T w = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\lambda_1^{(i)} - \lambda_2^{(i)})(\lambda_1^{(j)} - \lambda_2^{(j)}) \Phi(x^{(i)})^T \Phi(x^{(j)})
\end{aligned}
$$

Now we infer

$$
\begin{aligned}
F(w, b, ((\lambda_k^{(i)})_{\substack{k=1,\dots,4 \\ i=1,\dots,m}}) &= \frac{1}{2}\|w\|^2 + \underbrace{C \sum_{i=1}^{m} (\hat{\xi}^{(i)} \xi^{(i)}) + \sum_{i=1}^{m} \lambda_3^{(i)} (-\hat{\xi}^{(i)}) + \sum_{i=1}^{m} \lambda_4^{(i)} (-\xi^{(i)})}_{=\sum_{i=1}^m (C - \lambda_3^{(i)}) \hat{\xi}^{(i)} + (C - \lambda_4^{(i)}) \xi^{(i)} = \sum_{i=1}^m \lambda_1^{(i)} \hat{\xi}^{(i)} + \lambda_2^{(i)} \xi^{(i)}} \\
&\quad + \sum_{i=1}^{m} \lambda_1^{(i)} f_1^{(i)}(w,b) + \sum_{i=2}^{m} \lambda_1^{(i)} f_2^{(i)}(w,b) \\
&= \frac{1}{2}\|w\|^2 + \underbrace{\sum_{i=1}^{m} \lambda_1^{(i)} (f_1^{(i)}(w,b) + \hat{\xi}^{(i)})}_{\sum_{i=1}^m \lambda_1^{(i)} (y^{(i)} - (w\Phi(x^{(i)}) + b) - \varepsilon)} + \underbrace{\sum_{i=1}^{m} \lambda_2^{(i)} (f_2^{(i)}(w,b) + \xi^{(i)}}_{\sum_{i=1}^m \lambda_2^{(i)} ((w\Phi(x^{(i)}) + b) - y^{(i)} - \varepsilon)} \\
&= \frac{1}{2}\|w\|^2 + \underbrace{\sum_{i=1}^{m} (-\lambda_1^{(i)} + \lambda_2^{(i)})\, b}_{=0} + (-\varepsilon) \sum_{i=1}^{m} \lambda_1^{(i)} + \lambda_2^{(i)} + \sum_{i=1}^{m} (-\lambda_1^{(i)} + \lambda_2^{(i)}) y^{(i)} \\
&\quad + \underbrace{\sum_{i=1}^{m} (-\lambda_1^{(i)} + \lambda_2^{(i)}) w \Phi(x^{(i)})}_{=-w^T w,\text{ using the first constraint}}
\end{aligned}
$$

Therefore we get the desired result for the dual form

$$
\begin{aligned}
F(w, b, ((\lambda_k^{(i)})_{\substack{k=1,\dots,4 \\ i=1,\dots,m}}) &= -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\lambda_1^{(i)} - \lambda_2^{(i)})(\lambda_1^{(j)} - \lambda_2^{(j)}) \Phi(x^{(i)})^T \Phi(x^{(j)}) + (-\varepsilon) \sum_{i=1}^{m} \lambda_1^{(i)} + \lambda_2^{(i)} \\
&\quad + \sum_{i=1}^{m} (-\lambda_1^{(i)} + \lambda_2^{(i)}) y^{(i)}.
\end{aligned}
$$

This was our claim.