Machine Learning I: Foundations
Exercise Sheet 7

Prof. Marius Kloft TA:Billy Joe Franks

17.06.2020
Deadline: 16.06.2020

1) (MANDATORY) 10 Points
Interestingly the linear hard-margin SVM, given by

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t. } 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq 0,\ \forall i \in \{1,\dots,n\}, \tag{1}$$

requires only two (non-equal) training points (with opposite labels) to find a separating hyperplane. Let $X := \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ and $Y := \{y_1,\dots,y_n\}$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, be a dataset. Let $\mathbf{w}^*$ and b^* be the optimal solution to the above optimization problem (1) on X, Y. You may assume $w_1 \neq 0$.

a) Find a minimal dataset (X', Y') with $|X'| = |Y'| = 2$ (consisting only of two data points) with the same hard-margin SVM solution (Eq. (1)) as for the dataset (X, Y) , that is, $\mathbf{w}^*$ and b^*.

If we think about this exercise a bit its easy to figure out that the points we need to choose need to be on either side of the separating hyperplane each with distance margin γ from the separating hyperplane. Additionally the line going through both points needs to be orthogonal to the seperating hyperplane. This is easy to achieve. First we find a point on the hyperplane $H = \{\mathbf{x} \in \mathbb{R}^d | w^{*T}\mathbf{x} + b^* = 0\}$. Since we may assume that $w_1 \neq 0$, we will look for a point $\mathbf{x}$, with $x_1 \in \mathbb{R}$ and $\forall i \neq 1 : x_i = 0$. Lets call this point $\mathbf{x}^*$.

$$0 = w^{*T}\mathbf{x}^* + b^* = w_1^* x_1^* + b^*$$
$$x_1^* = -\frac{b^*}{w_1^*}$$

Now that we have a point on H, we just need to add (subtract) $\gamma \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|}$. Following the transformations done during the lecture we know $\gamma = \frac{1}{\|\mathbf{w}\|}$, for any $\mathbf{w}$ we choose, so the term simplifies to $\frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2}$. As such the points we choose are

- $\mathbf{x}_1^* := x^* + \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2}$ with $y_1 = 1$
- $\mathbf{x}_2^* := x^* - \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2}$ with $y_2 = -1$.

Just to check that $\mathbf{w}^*$ and b^* actually fulfill the constraints for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$.

$$\begin{aligned}
&1 - y_1\left(\mathbf{w}^{*T}\left(\mathbf{x}^* + \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2}\right) + b^*\right) \\
=&1 - \left(w_1^* \frac{-b^*}{w_1^*} + \frac{\|\mathbf{w}^*\|^2}{\|\mathbf{w}^*\|^2} + b^*\right) \\
=&1 - 1 = 0 \leq 0.
\end{aligned}$$

$$\begin{aligned}
&1 - y_2\left(\mathbf{w}^{*T}\left(\mathbf{x}^* - \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2}\right) + b^*\right) \\
=&1 + \left(w_1^* \frac{-b^*}{w_1^*} - \frac{\|\mathbf{w}^*\|^2}{\|\mathbf{w}^*\|^2} + b^*\right) \\
=&1 - 1 = 0 \leq 0
\end{aligned}$$

Since we expect both of these vectors to be support vectors it makes sense that the constraints are fulfilled with an equality to 0.

b) Prove that, for your choice of X' and Y' in a), $\mathbf{w}^*$ and b^* are optimal solutions of (1).

Lets assume the contrary, i.e. there exist $\mathbf{w}'$ and b', such that the constraints of (1) are fulfilled for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$, and $\|\mathbf{w}'\| < \|\mathbf{w}^*\|$. To this end first note the following

$$\frac{\mathbf{w}'^T\mathbf{w}^*}{\|\mathbf{w}^*\|^2} \leq \frac{\|\mathbf{w}'\|\,\|\mathbf{w}^*\|}{\|\mathbf{w}^*\|^2} < \frac{\|\mathbf{w}^*\|\,\|\mathbf{w}^*\|}{\|\mathbf{w}^*\|^2} = 1. \tag{2}$$

The first inequality is due to the Cauchy-Schwarz inequality. Now consider the constraints for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$

$$\begin{aligned}
&1 - y_1\left(\mathbf{w}'^T\left(\mathbf{x}^* + \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2}\right) + b'\right) \\
=&1 - \left(w_1'\frac{-b^*}{w_1^*} + \frac{\mathbf{w}'^T\mathbf{w}^*}{\|\mathbf{w}^*\|^2} + b'\right) \\
=&\left(1 - \frac{\mathbf{w}'^T\mathbf{w}^*}{\|\mathbf{w}^*\|^2}\right) + \left(w_1'\frac{b^*}{w_1^*} - b'\right) \leq 0
\end{aligned}$$

$$\begin{aligned}
&1 - y_2\left(\mathbf{w}'^T\left(\mathbf{x}^* - \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2}\right) + b'\right) \\
=&1 + \left(w_1'\frac{-b^*}{w_1^*} - \frac{\mathbf{w}'^T\mathbf{w}^*}{\|\mathbf{w}^*\|^2} + b'\right) \\
=&\left(1 - \frac{\mathbf{w}'^T\mathbf{w}^*}{\|\mathbf{w}^*\|^2}\right) + \left(-w_1'\frac{b^*}{w_1^*} + b'\right) \leq 0
\end{aligned}$$

Now using (2) we know that $1 - \frac{\mathbf{w}'^T\mathbf{w}^*}{\|\mathbf{w}^*\|^2} > 0$, we will call this quantity ϵ. Also we will call $c := w_1'\frac{b^*}{w_1^*} - b'$. Now simplifying the above inequalities we get

$$\epsilon + c \leq 0, \epsilon - c \leq 0$$

Which we can formulate as

$$\epsilon \leq -c, \epsilon \leq c$$

Since $\epsilon > 0$ this is a contradiction. As such $\mathbf{w}^*$ and b^* must be optimal for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$.

c) Why would it be advantageous to use (X', Y') instead of X, Y during optimization, assuming we had access to both and knew they are equivalent? (Answer this question with at most 5 sentences.)

The complexity of SVM optimization is dependant on the algorithm we choose to optimize the Quadratic Program, however every choice except for gradient descent is at least polynomial in the number of data points. Thus assuming we dont use GD using just 2 datapoints will be more efficient. If however we use GD, we can distinguish two cases. First, the case were we use Batch Gradient Descent, i.e. we use all data points in each iteration. And Second, the case where we use Stochastic Gradient Descent, i.e. we use very few, if not just 1 data point in each iteration. In the first case it is obviously the same argument as above, Batch Gradient Descent is at least linear in the number of data points. In the second case it might seem that there is no difference, as we will always just use 1 data point per iteration. However it can be shown that SGD using $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ will converge faster than any other choice of data set, however this is outside the scope of this exercise and also the lecture.

d) Assume we train the hard-margin SVM with only two (arbitrary) training points (not the optimal data points as above). Consider $d \to \infty$. What can you state regarding overfitting and underfitting here? Explain your answer. (Answer this question with at most 5 sentences.)

The simple answer is, as $d \to \infty$ the SVM is more likely to overfit. However we can explain this in greater detail. Consider the Vapnik-Chervonenkis dimension, this is a measure used to determine the capacity of a space of functions. If you have not heard of VC-dimension, consider reading up on it, as it is a very simple measure of the capacity of a function space `https://en.wikipedia.org/wiki/Vapnik-Chervonenkis_dimension`. Anyway it is simple to show that the linear SVM has a VC-dimension of $d + 1$, where d is the dimension of the data. Knowing this, we are considering a function space that is growing, $d \to \infty$, while the number of training data is staying constant. This is a clear example of overfitting. This should give you an intuition why feature selection is very important when using a linear SVM.

2) Consider the kernel ridge regression optimization problem (Lecture 8, Slide 39). Let $\alpha^* \in \mathbb{R}^d$ be the vector that minimizes the loss function. Show that:

$$\alpha^* = \left(K + \frac{1}{2C}\mathrm{I}_{n\times n}\right)^{-1} y$$

$$\begin{aligned}
\frac{\partial \frac{1}{2}\alpha^\top K\alpha + C||y - K\alpha||^2}{\partial \alpha} &= \frac{\partial \frac{1}{2}\alpha^\top K\alpha}{\partial \alpha} + \frac{\partial C||y - K\alpha||^2}{\partial \alpha} = \mathbf{0} \\
\frac{1}{2}2K\alpha + \frac{\partial C(y - K\alpha)^\top (y - K\alpha)}{\partial \alpha} &= \mathbf{0} \\
K\alpha - 2CK^\top (y - K\alpha) &= \mathbf{0} \\
K\alpha - 2CK(y - K\alpha) &= \mathbf{0} \\
\frac{1}{2C}K\alpha - K(y - K\alpha) &= \mathbf{0} \\
\frac{1}{2C}K\alpha - Ky + KK\alpha &= \mathbf{0} \\
\frac{1}{2C}K\alpha + KK\alpha &= Ky \\
K\left(\frac{1}{2C}\mathrm{I}_n\alpha + K\alpha\right) &= Ky \\
K^{-1}K\left(\frac{1}{2C}\mathrm{I}_n\alpha + K\alpha\right) &= K^{-1}Ky \\
\frac{1}{2C}\mathrm{I}_n\alpha + K\alpha &= y \\
\left(\frac{1}{2C}\mathrm{I}_n + K\right)\alpha &= y \\
\alpha = \left(\frac{1}{2C}\mathrm{I}_n + K\right)^{-1} y &= \alpha^*
\end{aligned}$$

3) In the lecture we found a closed form solution for linear ridge regression and we incorporated b afterwards by simply changing the dataset slightly. This however means that b is regularized during optimization. What would happen if we introduce b in a different way? Consider linear ridge regression with offset

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\left\|\mathbf{y} - (X^T\mathbf{w} + \hat{\mathbf{b}})\right\|^2 \tag{3}$$

where $\forall i : \hat{b}_i = b$. $\hat{\mathbf{b}}$ simply copies b into each component. Alternatively the norm could be written as a sum incorporating only b, as follows

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \left(y_i - (\mathbf{x}_i^T\mathbf{w} + b)\right)^2 \tag{4}$$

(3) and (4) have the same closed-form solution. Find this solution. Thereby choose the version from the above two that you prefer ((3) or (4)).

So we can rewrite (3) as

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\left\|\mathbf{y} - (X^T\mathbf{w} + b\mathbf{1}_{n\times 1})\right\|^2$$

And then we can rewrite it to a more familiar form as follows

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\mathbf{w}^T I\mathbf{w} + C\left\|\mathbf{y} - \left(\begin{bmatrix} X \\ \mathbf{1}_{1\times n}\end{bmatrix}^T \begin{bmatrix}\mathbf{w}\\ b\end{bmatrix}\right)\right\|^2$$

Now we can replace $\tilde{\mathbf{w}} := \begin{bmatrix}\mathbf{w}\\ b\end{bmatrix}$ and $\tilde{X} := \begin{bmatrix} X \\ \mathbf{1}_{1\times n}\end{bmatrix}$ and set

$$\tilde{I} := \begin{bmatrix} I_d & \mathbf{0}_{d\times 1} \\ \mathbf{0}_{1\times d} & 0\end{bmatrix}.$$

Then the problem becomes

$$\min_{\tilde{\mathbf{w}}\in\mathbb{R}^d} \frac{1}{2}\tilde{\mathbf{w}}^T\tilde{I}\tilde{\mathbf{w}} + C\left\|\mathbf{y} - \left(\tilde{X}^T\tilde{\mathbf{w}}\right)\right\|^2$$

Now this is exactly what we expect the optimization problem to look like, except for $\tilde{I}$. However, following the proof from the lecture, we easily observe the solution to be

$$\mathbf{w}_{RRwo} = \left(\tilde{X}\tilde{X}^T + \frac{1}{2C}\tilde{I}\right)^{-1}\tilde{X}y.$$

Interestingly the only difference from the proposition in the lecture, i.e. just appending 1 to every datapoint, is $\tilde{I}$ instead of I, as such we see adding a proper offset is incredibly simple. We just have to change a 1 to a 0.

4) Solve programming task 7.