Introduction to Machine Learning **Fall Semester**

Homework 4: December 7, 2022

Due: December 21, 2022

Theory Questions

1. **(20 points) SVM with multiple classes.** One limitation of the standard SVM is that it can only handle binary classification. Here is one extension to handle multiple classes. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and now let $y_1, \ldots, y_n \in [K]$, where $[K] = \{1, 2, \ldots, K\}$. We will find a separate classifier $\mathbf{w}_j$ for each one of the classes $j \in [K]$, and we will focus on the case of no bias ($b = 0$). Define the following loss function (known as the *multiclass hinge-loss*):

$$\ell(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i) = \max_{j \in [K]} \left(\mathbf{w}_j \cdot \mathbf{x}_i - \mathbf{w}_{y_i} \cdot \mathbf{x}_i + \mathbb{1}(j \neq y_i)\right),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Define the following multiclass SVM problem:

$$f(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i)$$

After learning all the $\mathbf{w}_j$, $j \in [K]$, classification of a new point $\mathbf{x}$ is done by $\arg\max_{j \in [K]} \mathbf{w}_j \cdot \mathbf{x}$. The rationale of the loss function is that we want the "score" of the true label, $\mathbf{w}_{y_i} \cdot \mathbf{x}_i$, to be larger by at least 1 than the "score" of each other label, $\mathbf{w}_j \cdot \mathbf{x}_i$. Therefore, we pay a loss if $\mathbf{w}_{y_i} \cdot \mathbf{x}_i - \mathbf{w}_j \cdot \mathbf{x}_i \leq 1$, for $j \neq y_i$.

Consider the case where the data is linearly separable. Namely, there exists $\mathbf{w}_1^*, ..., \mathbf{w}_K^*$ such that $y_i = argmax_y \mathbf{w}_y^* \cdot \mathbf{x}_i$ for all i. Show that any minimizer of $f(\mathbf{w}_1, ..., \mathbf{w}_K)$ will have zero classification error.

2. **(15 points) Solving hard SVM.** Consider two distinct points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ with labels $y_1 = 1$ and $y_2 = -1$. Compute the hyperplane that Hard SVM will return on this data, i.e., give explicit expressions for $\boldsymbol{w}$ and b as functions of $\mathbf{x}_1$, $\mathbf{x}_2$.
(Hint: Solve the dual problem by transforming it to an optimization problem in a single variable. Use your solution to the dual to obtain the primal solution).

3. **(15 points) ℓ^2 penalty.** Consider the following problem:

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2} \sum_{i=1}^{n} \xi_i^2$$
$$\text{s.t. } y_i(\boldsymbol{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n$$

   (a) Show that a constraint of the form $\xi_i \geq 0$ will not change the problem. Meaning, show that these non-negativity constraints can be removed. That is, show that the optimal value of the objective will be the same whether or not these constraints are present.

   (b) What is the Lagrangian of this problem?

   (c) Minimize the Lagrangian with respect to $\boldsymbol{w}, b, \boldsymbol{\xi}$ by setting the derivative with respect to these variables to 0.

(d) What is the dual problem?

4. **(15 points) Soft SVM on separable data.** Consider the soft-SVM problem with linearly separable data (assume no bias for simplicity):

$$\min_{\mathbf{w},\xi} 0.5\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$s.t\ \forall i: \quad y_i \mathbf{w}\cdot\mathbf{x}_i \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

Let $\mathbf{w}^\star$ be the solution of **hard SVM**. Show that if $C \geq \|\mathbf{w}^\star\|^2$ then the solution of soft SVM separates the data. (Hint: Show that in the optimal solution $(\mathbf{w}', \boldsymbol{\xi}')$ of the soft SVM problem, the sum of ξ_i''s is bounded by some constant smaller than 1).

5. **(15 points) Separability using polynomial kernel.** Let $x_1, ..., x_n \in \mathbb{R}$ be distinct real numbers, and let $q \geq n$ be an integer. Show that when using a polynomial kernel, $K(x, x') = (1 + xx')^q$, hard SVM achieves zero training error. Use the following fact: Given distinct values $\alpha_1, ..., \alpha_n$, the Vandermonde matrix defined by,

$$\begin{pmatrix} 1 & \alpha_1^1 & ... & \alpha_1^q \\ 1 & \alpha_2^1 & ... & \alpha_2^q \\ \vdots & & & \\ 1 & \alpha_n^1 & ... & \alpha_n^q \end{pmatrix},$$

is of rank n. (Hint: use the lemma from slide 7 in recitation 7).

Programming Assignment

Submission guidelines:

- Download the supplied files from Moodle. Written solutions, plots and any other non-code parts should be included in the written solution submission.
- Your code should be written in Python 3.
- Your code submission should include a single python file: `svm.py`.

1. **(20 points) Kernel SVM.** In this exercise, we will explore different polynomial kernel degrees for SVM. We will use an existing implementation of SVM: the SVC class from sklearn.svm. This class solves the soft-margin SVM problem. In the file `svm.py` you will find the method `plot_results` which gets following as an input:

   - A list of fitted estimators. That is, each element of the list is a return value of the fit method of the SVM model.
   - A list of names that corresponds to the models above.
   - Data points in $\mathbb{R}^2$: a numpy array of shape $n \times 2$, where n is the number of data points.
   - A numpy array of lables in $\{-1, 1\}$ for the above data points.

   The method plots the data points and the classifiers' prediction.

   (a) **(5 points)** The code you are given generates 200 samples (100 samples for each class) which are classified by a circle centered around the origin. Train three soft SVM models with regularization parameter $C = 10$, using linear kernel, **homogeneous** polynomial kernel of degree 2 and **homogeneous** polynomial kernel of degree 3. Plot your results using the methods above. Which of the models fits the data well? Explain the phenomena you see in the plots.
   (Hint: See the `coef0` parameter of the SVC class which determines whether or not the polynomial kernel is homogeneous or not).

   (b) **(5 points)** Repeat clause (a) above but now with **non-homogeneous** polynomial kernel. Do the results change? Explain.

   (c) **(10 points)** Perturb the labels in the following manner: Change each negative label to a positive one with probability 0.1. Train a soft-SVM model with a polynomial kernel of degree 2 (non-homogeneous), and another model using RBF kernel with $\gamma = 10$. Which of the two models seems to generalize better on the noisy data? What happens if you change γ? Submit your plots and explain the phenomena you see.

   (**Remark:** the parameter γ roughly corresponds to the $\frac{1}{\sigma}$ parameter of the RBF kernel. In your implementation, you may use `gamma` = 'auto' for polynomial kernels).