Homework 5

AM 218 Machine Learning

Due on May 14 at 11:59PM on Canvas

1 SVM (15 points)

We have a set of six labeled samples in the two-dimensional space, $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_6, y_6)\}$ where $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i \in \{-1, 1\}$, $i = 1, 2, \dots, 6$ listed as follows,

i	x_1	x_2	y
1	-1.2	1.6	1
2	-1.6	2	1
3	4	1	-1
4	-3	0	1
5	3	-0.8	-1
6	2	0	-1

(i) (3 pts) Find an linear classifier defined by $(\mathbf{w}, b)$ such that training samples are positive if and only if $\mathbf{w}^\top \mathbf{x} + b \geq 0$. In other words, find a hyperplane defined by $(\mathbf{w}, b)$ that can separate the the positive and negative samples.

(ii) (6 pts) In order to find a maximum margin classifier, we define the following SVM optimization problem with hard margin

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 \\ s.t. \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, i = 1, 2, \dots, 6. \end{aligned}$$

Can you find the solution to the above optimization problem without solving it? Please give the optimal solution $(\mathbf{w}^*, b^*)$ and describe how you derive the solution. (Hint: You can plot the training samples and find the solution using geometry.)

(iii) (6 pts) We also discuss in the class that we can solve the primal problem using the solution to the dual problem. Specifically, given the optimal solution to the dual problem, α_i^*, the solution $\mathbf{w}^*$ is given as follows

$$\mathbf{w}^* = \sum_{i=1}^{6} \alpha_i^* y_i \mathbf{x}_i.$$

Please find $\alpha_i^*, i = 1, 2, \ldots, 6$ without solving the dual optimization problem.

2 Kernels (10 pts)

To capture the non-linear relationship using SVM, we introduce the kernel function, which is the inner product of two vectors that are mapped into another high-dimensional space. Specifically, $K(\mathbf{x}, \mathbf{z})$ is a kernel function if it can be written as $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$ where $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ and the feature map $\phi(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$.

(i) (5 pts) Please show that if $K_1(\mathbf{x}, \mathbf{z})$ and $K_2(\mathbf{x}, \mathbf{z})$ are kernel functions, with positive α and β, then $K(\mathbf{x}, \mathbf{z}) = \alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$ is also a kernel function.

(ii) (5 pts) Please show that $K(\mathbf{x}, \mathbf{z}) = 400(\mathbf{x}^\top \mathbf{z})^2 + 100\mathbf{x}^\top \mathbf{z}$ is a kernel function.

(Hint: the definition of Kernel function)

3 Decision Tree (15 pts)

We plan to train a decision tree model on the following dataset to predict whether an email is a spam or not. We have reprocessed the data and the binary-valued features indicate whether I know the author, whether the email is long, and whether it contain certain key words. Finally, the last column indicates whether the email is determined as a spam. ($y = +1$ for ”spam” and $y = -1$ for ”ham”)

x_1 know author?	x_2 is long?	x_3 has "research"	x_4 has "grade"	x_5 has "lottery"	y spam?
0	0	1	1	0	+1
1	1	0	1	0	+1
0	1	1	1	1	+1
0	1	0	0	0	+1
0	1	0	0	0	+1
1	0	1	1	1	-1
0	0	1	0	0	-1
1	0	0	0	0	-1
1	0	1	1	0	-1
1	1	1	1	1	+1

(i) Which feature should we choose first to spilt the data using the information gain as the splitting criterion?

(ii) Grow the decision tree to the fullest (i.e., no more feature/samples left or no more information gain for spliting) manually using the information gain as spliting criterion. Draw the decision tree and the predicted labels for each leaf node.

4 Boosting (15 pts)

We will apply the AdaBoost algorithm on the following dataset with the weak learners (e.g. decision stumps) of the form (i) "$x \geq \theta_x$" **or** (ii) "$y \geq \theta_y$" for some integers θ_x and θ_y (either one of the two forms), i.e.,

$$\text{label} = \begin{cases} + & \text{if } x \geq \theta_x \\ - & \text{otherwise} \end{cases} \quad \textbf{or} \; \text{label} = \begin{cases} + & \text{if } y \geq \theta_y \\ - & \text{otherwise} \end{cases}$$

(i) Start the first round with a uniform distribution D_1 over the data. Find the weak learner h_1 that can minimize the weighted misclassification rate and predict the data samples using h_1.

(ii) Update the weight of each data sample, denoted by D_2, based on the results in (1). Find the

i	x	y	Label
1	1	10	$-$
2	4	4	$-$
3	8	7	$+$
4	5	6	$-$
5	3	16	$-$
6	7	7	$+$
7	10	14	$+$
8	4	2	$-$
9	4	10	$+$
10	8	8	$-$

weak learner h_2 that can minimize the weighted misclassification rate with D_2, and predict the data samples using h_2.

(iii) Write the form of the final classifier obtained by the two-round AdaBoost.