UCB CS189: Homework 1

Dang Truong

October 2024

Theory of Hard-Margin Support Vector Machines

A *decision rule* (or *classifier*) is a function $r : \mathbb{R}^d \to \pm 1$ that maps a feature vector (test point) to $+1$ (“in class") or -1 (“not in class"). The decision rule for linear SVMs is of the form

$$r(x) = \begin{cases} +1 \text{ if } w \cdot x + \alpha \geq 0, \\ -1 \text{ otherwise,} \end{cases} \tag{1}$$

where $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ are the parameters of the SVM. The primal hard-margin SVM optimization problem (which chooses the parameters) is

$$\min_{w,\alpha} \|w\|^2 \quad \text{subject to} \quad y_i(X_i \cdot w + \alpha) \geq 1, \quad \forall i \in \{1, \ldots, n\}, \tag{2}$$

where $\|w\| = \sqrt{w \cdot w}$.
We can rewrite this optimization problem by using Lagrange multipliers to eliminate the constraints. We thereby obtain the equivalent optimization problem

$$\max_{\lambda_i \geq 0} \min \|w\|^2 - \sum_{i=1}^{n} \lambda_i (y_i(X_i \cdot w + \alpha) - 1). \tag{3}$$

Note: λ_i must be greater than or equal to 0.

(a) Show that the equation (3) can be rewritten as the *dual optimization problem*

$$\max_{\lambda_i \geq 0} \sum_{i=1}^{n} \lambda_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j X_i \cdot X_j \quad \text{subject to} \quad \sum_{i=1}^{n} \lambda_i y_i = 0. \tag{4}$$

$$\begin{aligned} L(w, \alpha) &= \|w\|^2 - \sum_{i=1}^{n} \lambda_i (y_i(X_i \cdot w + \alpha) - 1) \\ &= w^\top \cdot w - \sum_{i=1}^{n} \lambda_i y_i X_i \cdot w - \sum_{i=1}^{n} \lambda_i y_i \alpha + \sum_{i=1}^{n} \lambda_i \end{aligned}$$

Taking the partial derivatives of L with respect to w and α, we obtain

$$\begin{aligned}\frac{\partial L}{\partial w} &= 2w - \sum_{i=1}^{n} \lambda_i y_i X_i \\ \frac{\partial L}{\partial \alpha} &= \sum_{i=1}^{n} \lambda_i y_i\end{aligned}$$

To minimize $L(w, \alpha)$, we need $\frac{\partial L}{\partial w} = 0$ and $\frac{\partial L}{\partial \alpha} = 0$. Hence, we obtain $w = \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_i X_i$ and $\sum_{i=1}^{n} \lambda_i y_i = 0$. Substituting these new constraints back into $L(w, \alpha)$, we obtain

$$\begin{aligned}L(w,\alpha) &= \sum_{i=1}^{n} \frac{\lambda_i y_i X_i^\top}{2} \cdot \sum_{j=1}^{n} \frac{\lambda_j y_j X_j}{2} - \sum_{i=1}^{n} \lambda_i y_i X_i^\top \cdot \sum_{j=1}^{n} \frac{\lambda_j y_j X_j}{2} - 0 + \sum_{i=1}^{n} \lambda_i \\ &= \sum_{i=1}^{n} \lambda_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j X_i \cdot X_j\end{aligned}$$

Therefore, we can rewrite the equation (3) as

$$\max_{\lambda_i \geq 0} \sum_{i=1}^{n} \lambda_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j X_i \cdot X_j \text{ subject to } \sum_{i=1}^{n} \lambda_i y_i = 0.$$

(b) Suppose we know the values λ_i^* and α^* that optimize equation (3). Show that the decision rule specified by equation (1) can be written

$$r(x) = \begin{cases} +1 & \text{if } \alpha^* + \frac{1}{2}\sum_{i=1}^{n} \lambda_i^* y_i X_i \cdot x \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

Since $w = \frac{1}{2}\sum_{i=1}^{n} \lambda_i^* y_i X_i$ and $\alpha = \alpha^*$, we have $w \cdot x + \alpha = \frac{1}{2}\sum_{i=1}^{n} \lambda_i^* y_i X_i \cdot x + \alpha^*$.

(c) Apply Karush-Kuhn-Tucker (KKT) conditions, any pair of optimal primal and dual solutions w^*, α^*, λ^* for a linear, hard-margin SVM must satisfy the following condition:

$$\lambda^*(y_i(X_i \cdot w^* + \alpha^*) - 1) = 0 \quad \forall i \in \{1, \ldots, n\}$$

This condition is called *complementary slackness*. Explain what this implies for points corresponding to $\lambda_i^* > 0$.

When $\lambda_i^* > 0$, we must have that $y_i(X_i \cdot w^* + \alpha^*) - 1 = 0$ or $y_i(X_i \cdot w^* + \alpha^*) = 1$. This implies that $X_i \cdot w^* + \alpha^* = 1$ if $y_i = 1$ and $X_i \cdot w^* + \alpha^* = -1$ if $y_i = -1$. Therefore, points corresponding to $\lambda_i^* > 0$ lie on either the positive margin or the negative margin.

(d) The training points X_i for which $\lambda_i^* > 0$ are called the *support vectors*. In practice, we frequently encounter training data sets for which the support vectors are a small minority of the training points, especially when the number of training points is much larger than the number of features. Explain why the support vectors are the only training points needed to evaluate the decision rule.

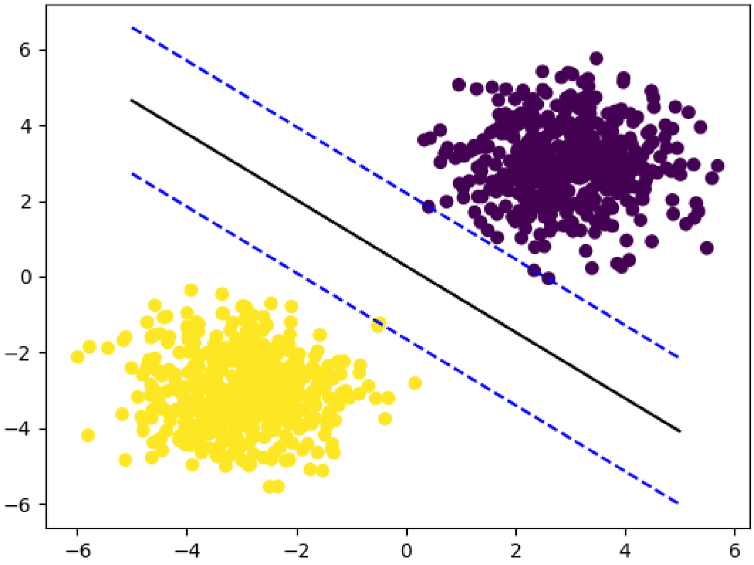
The support vectors are the only training points needed to evaluate the decision rule since they are the only points that directly determine the position of the decision boundary.

(e) The obtained parameters when fitting the linear SVM to the 2D synthetic dataset found in **toy-data.npz** approximately correspond to

$$w = \begin{bmatrix} -0.4528 \\ -0.5190 \end{bmatrix} \quad \text{and} \quad \alpha = 0.1471.$$

Using only matplotlib basic plotting functions, in your write-up, produce a plot of

- the data points,
- the decision boundary,
- the margins, defined as $\{x \in \mathbb{R}^2 : w \cdot x + \alpha = \pm 1\}$.



There are four support vectors in the plot above: one yellow dot and three purple dots lying on the two blue dashed lines.

(f) Assume the training points X_i and labels y_i are linearly separable. Using the original SVM formulation (not the dual) prove that there is at least one support vector for each class, +1 and -1.

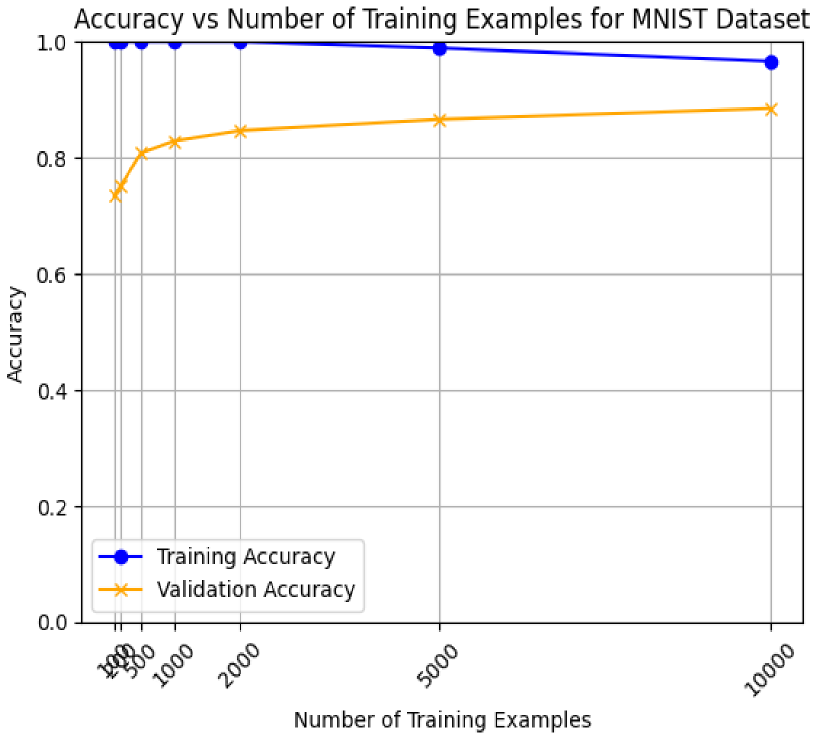
Proof. Assume for contradiction that w and α are the optimal parameters for the SVM and there are no support vectors for either class. Since there are no support vectors, we know that $y_i(X_i \cdot w + \alpha) > 1$ for every training point X_i. For each point X_i, let define $\varepsilon_i = y_i(X_i \cdot w + \alpha) - 1 > 0$. Then, we let ε be the minimum of all ε_i across both classes $\varepsilon = \min\{\varepsilon_i\} > 0$. Now, construct a new weight vector $w' = w/(1+\varepsilon/2)$ and corresponding bias $\alpha' = \alpha/(1+\varepsilon/2)$. Given a point X_i, we have

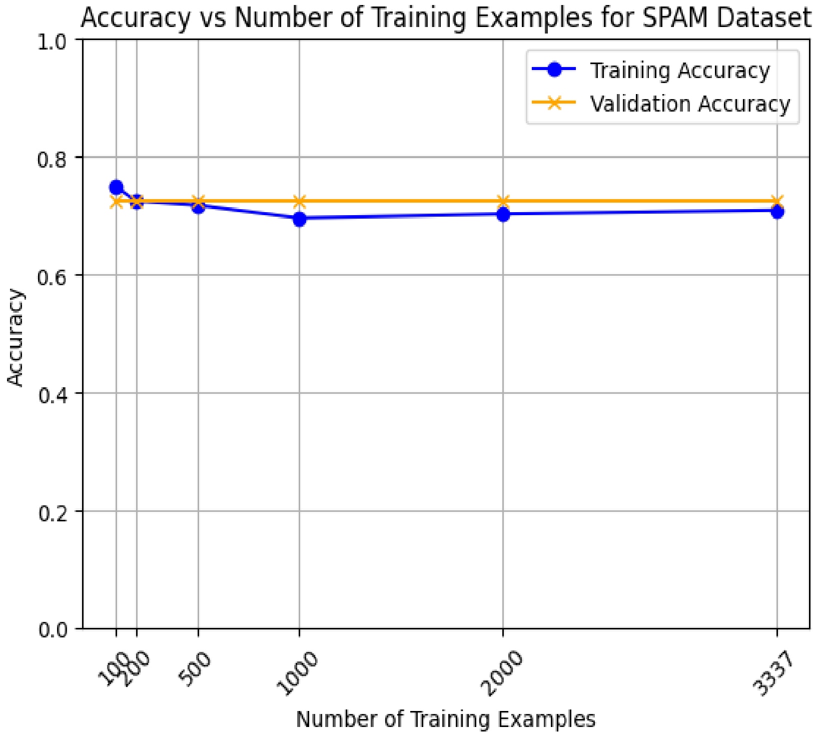
$$y_i(X_i \cdot w' + \alpha') = y_i\left(\frac{X_i \cdot w}{1+\varepsilon/2} + \frac{\alpha}{1+\varepsilon/2}\right) = \frac{y_i(X_i \cdot w + \alpha)}{1+\varepsilon/2} = \frac{1+\varepsilon_i}{1+\varepsilon/2} \geq \frac{1+\varepsilon}{1+\varepsilon/2} > 1$$

□

Hence, the decision boundary determined by the weight w' and bias α' correctly classifies every training point X_i. However, this decision boundary has a smaller margin since $\|w'\| = \|w\|/(1+\varepsilon/2) < \|w\|$. This contradicts the assumption that we have found the optimal SVM solution. Therefore, there is at least one support vector for each class.

Support Vector Machines: Coding


Accuracy vs Number of Training Examples for MNIST Dataset
Accuracy
Number of Training Examples
Training Accuracy
Validation Accuracy


Accuracy vs Number of Training Examples for SPAM Dataset
Accuracy
Number of Training Examples
Training Accuracy
Validation Accuracy