Machine Learning for EDS

Tutorial week 6

2023/2024

Problem 1 (SVM in a simple setting)

Consider a feature space $\mathcal{X} = \mathbb{R}$ and let $x_1 = 3, x_2 = 5, x_3 = 6.5$ and $y_1 = -1,\ y_2 = 1, y_3 = 1$. We consider classifiers of the form $f_{w,w_0}(x) = \text{sgn}(wx + w_0)$ for $x \in \mathbb{R}$, so the model $S = \{f_{w,w_0} : w, w_0 \in \mathbb{R}\}$. We use hard margin SVM to select a suitable classifier from this model for the given sample.

1. Draw an x and y-axis and 'plot' the points x_1, x_2, x_3 on the x-axis. Give the points a different color depending on their label y_i.

2. Say we select an element of S using hard margin SVM for the given sample. Without solving the SVM optimization problem, determine where the discriminating 'hyperplane' (i.e. the decision boundary) should be. Draw this point on the x-axis of your plot.

3. Indicate the corresponding support vectors and the width of the margin.

4. Give the set of possible values for (w, w_0) which lead to a classifier $f_{w,w_0}(x)$ that correctly classifies x_1, x_2, x_3 and has the decision boundary that you found in 2. Verify that indeed $y_i(wx_i + w_0) > 0$ for any choice of w and w_0 from this set. For two such combinations of w and w_0, draw the function $h(x) = wx + w_0$ in your plot.

It is clear from question 4 that the parameters w and w_0 are not uniquely identified: multiple combinations lead to the same decision boundary. Consider the normalization proposed in the lectures: let w and w_0 be such that $y_i(wx_i + w_0) \geq 1$ for all i, define the margin as the distance between the hyperplanes $wx + w_0 = 0$ and $wx + w_0 = 1$ and maximize the margin under these restrictions.

5. Without using previous results, determine what the value of w and w_0 should be based on the conditions mentioned above. Plot the resulting function $h(x) = wx + w_0$. Also plot horizontal lines at $y = 1$ and $y = -1$ and show where $h(x)$ intersects these lines. Are these points of intersection indeed the support vectors you found in 3? Is the width of the margin the same as in 3?

6. Say point x_3 would have had label $y_3 = -1$ instead of 1. Without going in detail, what approach could you take to still be able to apply hard margin SVM to this sample?

Problem 2 (VC dimension of kernel SVM classifiers)

In this problem we will consider the VC dimension of the models associated to SVM classifiers based on different kernel functions. Assume you have a sample of feature vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ where $\boldsymbol{x}_i \in \mathbb{R}^p$, with corresponding labels $y_1, \ldots, y_n$ with $y_i \in \{-1, 1\}$.

1. Consider the linear kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$. Based on the results of week 5, give the VC dimension of the model of classifiers corresponding to the SVM based on this kernel, i.e. $S = \{\boldsymbol{x} \mapsto \operatorname{sgn}(\boldsymbol{w}^\top \boldsymbol{x} + w_0) : \boldsymbol{w} \in \mathbb{R}^p, w_0 \in \mathbb{R}\}$.

2. Consider the polynomial kernel function with degree d: $K(\boldsymbol{x}, \boldsymbol{x}') = (c + \boldsymbol{x}^\top \boldsymbol{x}')^d$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$ and some $c \geq 0$. This kernel function can be decomposed $K(\boldsymbol{x}, \boldsymbol{x}') = \varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}')$, where $\varphi(\boldsymbol{x})$ is a function from $\mathbb{R}^p$ to some other space.

   (a) For $p = 2$, give the form of $\varphi(\boldsymbol{x})$ for the polynomial kernel function with degree $d = 3$ and $c = 1$. Use the notation $\boldsymbol{x} = (x_1, x_2)$.

   (b) Give an upper bound of the VC dimension of the model of classifiers considered by the SVM based on the polynomial kernel with degree $d = 3$ and $c = 1$ for $p = 2$ using the results of week 5.

   (c) Answer questions (a) and (b) again for the case $c = 0$.

For the RBF kernel with parameter $\gamma > 0$: $K(\boldsymbol{x}, \boldsymbol{x}^\top) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|)$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$, the implied feature space is infinite dimensional. As we saw in the lecture, we can derive a bound on the VC dimension of classifiers corresponding to SVMs that does not depend on the dimension of the feature space. We will prove a simplified version of this result here.

Let $R \subseteq \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq r\}$ for some $r > 0$. We will prove that

$$S = \left\{ \boldsymbol{x} \mapsto \operatorname{sgn}(\boldsymbol{w}^\top \boldsymbol{x}) : \min_{\boldsymbol{x} \in R} |\boldsymbol{w}^\top \boldsymbol{x}| = 1 \text{ and } \|\boldsymbol{w}\| \leq \Lambda \right\},$$

has

$$\operatorname{VCdim}(S) \leq r^2 \Lambda^2,$$

where $\|\cdot\|$ denotes the Euclidean norm, i.e. $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$.

3. (a) Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$ be a set of features with $\boldsymbol{x}_i \in R$ that can be shattered by S. Show that for any choice of $\{y_1, \ldots, y_d\} \in \{-1, 1\}^d$, we have $d \leq \Lambda \| \sum_{i=1}^d y_i \boldsymbol{x}_i \|$.

Hint: use the Cauchy-Schwarz inequality, which says that $|\boldsymbol{a}^\top \boldsymbol{b}| \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\|$ for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ of the same dimension.

(b) Say that each y_i is independently drawn from a uniform distribution over $\{-1, 1\}$. Show that for each $i = 1, \ldots, n$, $\mathbb{E}[y_i] = 0$, $\mathbb{E}[y_i^2] = 1$ and $\mathbb{E}[y_i y_j] = 0$ for $j \neq i$.

(c) The inequality derived in (a) holds for any combination of y_i's, so also in expectation over $\{y_1, \ldots, y_d\}$ drawn independently according to the distribution described in (b). Use this to show that $d \leq \Lambda\sqrt{\sum_{i=1}^d \|\boldsymbol{x}_i\|^2}$.

Hint: use that $\mathbb{E}|x| \leq (\mathbb{E}|x|^2)^{1/2}$ by Jensen's inequality.

(d) Conclude that $d \leq r^2\Lambda^2$ and that therefore $\mathrm{VCdim}(S) \leq r^2\Lambda^2$.