Introduction to Statistical Learning Theory

Lecture 5

A very common and useful ML algorithm we will study is the Support Vector Machine - SVM. It will be a running example and we will see how we can analyse it from various perspectives.

The basic idea of SVM is a large margin linear predictor.

Assume a training set is linearly separable - i.e. there exists some w such that $\forall i : y_i \langle w, x_i \rangle > 0$. This means the ERM has zero loss, but this zero loss is achieved by many vectors. SVM picks the one with the largest margin.

Lemma 1.1

The distance between x and the hyperplane defined by w is $\frac{|\langle w,x \rangle|}{||w||}$.

Algorithm `Hard-SVM`

Input: $(x_1, y_1), ..., (x_m, y_m)$ linearly separable.
Return: $w = \arg\min ||w||^2$
Subject to: $\forall i : y_i \langle w, x_i \rangle \geq 1$

Lemma 1.2

If the data is linearly separable, the Hard-SVM returns the maximal margin vector.

Proof -exercise.

The demand that the data is linearly separable is usually not satisfied, so to solve this we add slack variables.

Algorithm `SVM`

Input: $(x_1, y_1), ..., (x_m, y_m)$, parameter λ
Return: $w = \arg\min_{w,\xi} \left(\lambda||w||^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i\right)$
Subject to: $\forall i : \; y_i \langle w, x_i \rangle \geq 1 - \xi_i$ and $\xi_i \geq 0$.

There is another way to view the SVM objective -

Lemma 1.3

Define $\ell^{hinge}(w, (x, y)) = max\{0, 1 - y \langle w, x \rangle\}$. *Then the SVM returns* $\arg\min(\lambda||w||^2 + L_S^{hinge}(w))$.

This means that we replace the $0 - 1$ loss with the hinge loss, and add a *regularization* that biases towards lower norm.

Lemma 1.4

The hinge loss has the following properties:

- $\ell^{0-1}(w,(x,y)) \leq \ell^{hinge}(w,(x,y))$.
- ℓ^{hinge} *is convex.*
- $\ell^{hinge}(w,(x,y))$ *is* $||x||$*-Lipschitz in* w.

The first two claims make the hinge loss a convex *surrogate loss*, which makes the optimization computationally tractable.

One can show that the hinge loss is the smallest function satisfying all three requirements.

Theorem 1.5 (Representation Theorem)

Let $\bar{w} = \arg\min_{w} \left(\lambda ||w||^2 + \sum_{i=1}^{m} f(\langle w, x_i \rangle, y_i) \right)$ *for some* $\lambda > 0$*, then* $\bar{w} \in span(x_1, ..., x_m)$*, i.e. is a linear combination of the inputs.*

Proof.

Let $\bar{w}$ be the minimizer, then $\bar{w} = w_\perp + w_\parallel$ where $w_\parallel \in span(x_1, ..., x_m)$ and $w_\perp \perp span(x_1, ..., x_m)$. We have $||w||^2 = ||w_\perp||^2 + ||w_\parallel||^2$. If by contradiction $||w_\perp|| > 0$, then $f(\langle \bar{w}, x_i \rangle, y_i) = f(\langle w_\parallel, x_i \rangle, y_i)$ while $||w_\parallel||^2 < ||\bar{w}||^2$ contradiction it being the minimum. □

Theorem 1.6

Let $\bar{w}$ be the minimizer of the SVM objective, then $\bar{w} = \sum \alpha_i y_i x_i$ where $\alpha_i \geq 0$, and $\alpha_i > 0$ iff x_i is on the margin or has a non-zero slack.

These vectors with $\alpha_i > 0$ are the support vectors which give the algorithm its name. The proof is based on the KKT optimality conditions.

We will show how the Rademacher complexity can be used to prove generalization bounds for SVM. We will start with a general linear space:

Theorem 2.1

Define $\mathcal{H}_2 = \{x \to \langle x, w \rangle : ||w||_2 \leq 1\}$ *and let* $S = (x_1, ..., x_m)$ *be vectors in that space. Then*

$$R(\mathcal{H}_2 \circ S) = R(\{\langle w, x_1 \rangle, ..., \langle w, x_m \rangle \ \ ||w||_2 \leq 1\}) \leq \frac{\max_i ||x||_2}{\sqrt{m}}$$

Proof:

$$mR(\mathcal{H}_2 \circ S) = \mathbb{E}_\sigma \left[\sup_{w:\,||w|| \leq 1} \sum_{i=1}^{m} \sigma_i \langle w, x_i \rangle \right] = \mathbb{E}_\sigma \left[\sup_{w:\,||w|| \leq 1} \left\langle w, \sum_{i=1}^{m} \sigma_i x_i \right\rangle \right]$$

Using the Cauchy-Schwartz inequality and the norm bound on w we get

$$mR(\mathcal{H}_2 \circ S) \leq \mathbb{E}_\sigma \left[|| \sum_{i=1}^{m} \sigma_i x_i ||_2 \right] = \mathbb{E}_\sigma \left[\left(|| \sum_{i=1}^{m} \sigma_i x_i ||_2^2 \right)^{1/2} \right]$$

$$\stackrel{1}{\leq} \left(\mathbb{E}_\sigma \left[|| \sum_{i=1}^{m} \sigma_i x_i ||_2^2 \right] \right)^{1/2} = \left(\mathbb{E}_\sigma \left[\sum_{i,j} \sigma_i \sigma_j \langle x_i, x_j \rangle \right] \right)^{1/2}$$

$$\stackrel{2}{=} \left(\sum_{i=1}^{m} ||x_i||^2 \mathbb{E}_\sigma [\sigma_i^2] \right)^{1/2} \leq \sqrt{m} \max_i ||x_i||_2$$

Where (1) is due to the Jensen inequality, and (2) is due to independence. $\square$

Notice that the bound does not depend on the dimension!

We will show a generalization bound for Hard-SVM, if the data is linearly separable.

Theorem 2.2

Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{\pm 1\}$ such that there exists some w^* with $P_{\mathcal{D}}(y \langle w^*, x \rangle \geq 1) = 1$ and $||x||_2 \leq R$ with probability 1. Let w_S be the output of the Hard-SVM, then with probability greater or equal to $1 - \delta$ we have

$$P_{\mathcal{D}}(y \neq sign(\langle w_S, x \rangle)) = L_{\mathcal{D}}^{0-1}(w_S) \leq \frac{2R||w^*||}{\sqrt{m}} + (1 + R||w^*||)\sqrt{\frac{2\ln(2/\delta)}{m}}$$

Proof: As the hinge loss bounds the $0-1$ loss we note that $L_{\mathcal{D}}^{0-1}(w_S) \leq L_{\mathcal{D}}^{hinge}(w_S)$. Also note that $L_S^{hinge}(w_S) = 0$.

Define $\phi(\langle w, x \rangle, y) = \max\{0, 1 - y\langle w, x \rangle\}$. Note that ϕ is $R-$Lipschitz on our domain.

Define $\mathcal{H}_2 = \{w : ||w||_2 \leq ||w^*||_2\}$, we know that for any sample $w_S \in \mathcal{H}_2$ so it is enough to bound
$R(\mathcal{F} \circ S) = \{(\phi(\langle w, x \rangle, y), ..., \phi(\langle w, x \rangle, y)) : w \in \mathcal{H}_2\}$. From theorem 2.1 and the concentration lemma we get that $R(\mathcal{F} \circ S) \leq \frac{R||w^*||}{\sqrt{m}}$.
From the generalization theorem on Rademacher complexity, with probability greater or equal to $1 - \delta$ for all $w \in \mathcal{H}_2$
$L_{\mathcal{D}}(h) - L_S(h) \leq 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) + c\sqrt{\frac{2\ln(2/\delta)}{m}}$, where c is the maximal loss which in our case is $1 + R||w^*||$ finishing the proof. $\square$

There is one drawback to our proof - we do not know $||w^*||$. We will now show a data-dependent bound.

Theorem 2.3

Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{\pm 1\}$ such that there exists some w^* with $P_{\mathcal{D}}(y \langle w^*, x \rangle \geq 1) = 1$ and $||x||_2 \leq R$ with probability 1. Let w_S be the output of the Hard-SVM, then with probability greater or equal to $1 - \delta$ we have

$$P_{\mathcal{D}}(y \neq sign(\langle w_S, x \rangle)) \leq \frac{4R||w_S||}{\sqrt{m}} + (1 + 2R||w_S||)\sqrt{\frac{2\ln(4||w_S||/\delta)}{m}}$$

Proof - Define $\mathcal{H}_i = \{w : ||w|| \leq 2^i\}$ and $\delta_i = \delta/2^i$. Note that $\sum_{i=1}^{\infty} \delta_i = \delta$. For each i we have (similar to previous theorem) that for all $h \in \mathcal{H}_i$ with probability greater then $1 - \delta_i$,

$$L_{\mathcal{D}}(w) \leq L_S(w) + \frac{2R2^i}{\sqrt{m}} + (1 + R2^i)\sqrt{\frac{2\ln(2/\delta_i)}{m}}$$

From the union bound, we get that with probability greater then $1 - \delta$ this holds for all $\mathcal{H}_i$. This means that for all $w \in \mathcal{H}$ we have for $i = \lceil \log(||w||) \rceil \leq \log(||w||) + 1$

$$L_{\mathcal{D}}(w) \leq L_S(w) + \frac{4R||w||}{\sqrt{m}} + (1 + 2R||w||)\sqrt{\frac{2\ln(4||w||/\delta)}{m}}$$

Plugging $w = w_S$, remembering $L_S(w_S) = 0$ finishes the proof. $\square$

We notice that the last proof can be adjusted easily to work for ”soft” SVM

Theorem 2.4

Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{\pm 1\}$ such that $||x||_2 \leq R$ with probability 1. Let w_S be the output of the SVM algorithm, then with probability greater or equal to $1 - \delta$ we have

$$L_{\mathcal{D}}^{0-1}(w_S) \leq L_S^{hinge}(w_S) + \frac{4R||w_S||}{\sqrt{m}} + (1 + 2R||w_S||)\sqrt{\frac{2\ln(4||w_S||/\delta)}{m}}$$