Support Vector machine

Margin theory

Plan

- Motivation
- Confidence margin
- Rademacher complexity
- Margin loss function
- Bounds for classification
- Generalization bound for regression – finite set
- Rademacher complexity of kernel based hypotheses
- Margin theory for kernel based hypotheses

Motivation

We will present a generalization bound based on the notion of margin, which, compared to VC dimension, provides a strong theorical justification for the SVM algorithm.

Confidence margin

The following learning guarantees hold for real-valued functions such as the functions $x \mapsto w.x + b$ returned by SVM, as opposed to classification functions returning +1 or -1 such as $x \mapsto \text{sgn}(w.x + b)$

The confidence margin for a real-valued function h at a point x labeled with y is the quantity $y.h(x)$

We interpret the magnitude of $|h(x)|$ as the confidence of the prediction made by h.

The notion of confidence margin is different from the geometric margin and does not require a linear separability assumption.

Confidence margin

The two notions are related in the separable case as follows : for h : $x \mapsto w.x + b$ with geometric margin ρ_{geom} ,the margin ρ of h for a sample $S = (x_1; \dots.; x_m)$ is at least $\rho_{geom}.||w||$:

$$\rho_{geom} = \frac{y(w.x + b)}{||w||}$$

$$\rho = \min_{1 \le i \le m} \frac{y(w.x_i + b)}{||w||}$$

$$|y.h(x)| \ge \rho_{geom}.||w||$$

VC-dimension generalization bound

the VC–dimension of the family of linear hypotheses in $\mathbb{R}^N$ *is* $N+1.$ *thus, for any* $\delta > 0$ *with probability at least* $1-\delta$*, for any* $h \in \mathcal{H}$:

$$L_D(h) \leq L_s(\mathrm{h}) + \sqrt{\frac{2(N+1)\log\frac{em}{N+1}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

Theorem

Let $S \subseteq \{x : ||x|| < r\}$ be a sample of size m, then, the VC-dimension d of the set of canonical hyperplanes $\mathcal{H} = \left\{x \mapsto \mathrm{sgn}(w.x) : \min_{x \in S}|w.x| \; and \; ||w|| < \Lambda\right\}$ verifies :

$$\mathrm{d} \leq r^2\Lambda^2$$

In linearly separable case, we can take $||w|| = \frac{1}{\rho}$ and set $\Lambda = \frac{1}{\rho}$, the upper bound can then be $\frac{r^2}{\rho^2}$.

Rademarcher complexity bound

We can also, bound the Rademarcher complexity of linear hypotheses with bounded weight vector in a similar way.

$\mathcal{H}$ a hypothesis set.

$\mathcal{G}$ $a\ family\ of\ loss\ functions\ \mathcal{L} : Y \times Y \longmapsto \mathbb{R}$ associated to $\mathcal{H}$, mapping from

$Z = X \times Y\ to\ \mathbb{R}$:

$$\mathcal{G} = \{g: (x; y) \mapsto \mathcal{L}(h(x); y)\ /h \in \mathcal{H}\}.$$

The Rademacher complexity captures the richness of a family of functions by measuring the degree to which a hypothesis set can fit random noise.

Empirical Rademacher complexity

Let $\mathcal{g}$ be a family of functions mapping from $Z = X \times Y\ to\ [a; b]$ and $S = (z_1; \dots.; z_m)$ a fixed sample of size m with elements in Z. The empirical Rademacher complexity of $\mathcal{g}$ with respect to the sample S is defined as :

$$\widehat{\Re}_s(\mathcal{g}) = \mathbb{E}_\sigma[sup_{g\in\mathcal{g}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right)]$$

Where σ=$(\sigma_1; \dots.; \sigma_m)^\intercal$, with σ_i independent uniform random variables called Rademacher variables taking values in {-1;1} .

Rademacher complexity

Let $\mathcal{D}$ denote the distribution according to which samples are drawn. For any integer $m \geq 1$ the Rademacher complexity of $\mathcal{g}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$.

$$\mathfrak{R}_m(\mathcal{g}) = \mathbb{E}_{s \sim \mathcal{D}}[\widehat{\mathfrak{R}}_s(\mathcal{g})]$$

Theorem (generalization bound)

Let $\mathscr{g}$ be a family of functions mapping from $Z = X \times Y\ to\ [0; 1]$ the for any $\delta > 0$, with a probability at least 1- δ over a drawn of i.i.d. sample S of size m, each of the following holds for all g of $\mathscr{g}$:

$$\mathbb{E}[\mathrm{g(z)}] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathscr{g}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$$

And, $\mathbb{E}[\mathrm{g(z)}] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathscr{g}) + 3\sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$

Lemma

Let $\mathcal{H}$ be a family of functions taking values in $\{-1;+1\}$ and let $\mathcal{g}$ be the family of loss function associated to $\mathcal{H}$ for the zero-one loss : $\mathcal{g} = \{(\mathrm{x};\mathrm{y}) \longmapsto 1_{h(x)\neq y} : h \in \mathcal{H}\}$.

For any sample $S = ((x_1;y_1);\ldots;(x_m;y_m))$ of elements in $X \times \{-1;+1\}$, let S_X denote its projection over X: $S_X = (x_1;\ldots.;x_m)$, then the following relation holds between the empirical Rademacher complexities of $\mathcal{g}$ and $\mathcal{H}$:

$$\widehat{\mathfrak{R}}_s(\mathcal{g}) = \frac{1}{2}\widehat{\mathfrak{R}}_{S_X}(\mathcal{H})$$

Rademacher complexity bounds – binary classification

Let $\mathcal{H}$ be a family of functions taking values in $\{-1;+1\}$ and let D be the distribution over the input space X. Then for any $\delta > 0$, with a probability at least $1-\delta$ over a sample S of size m drawn according to D, each of the following holds for all $h \in \mathcal{H}$:

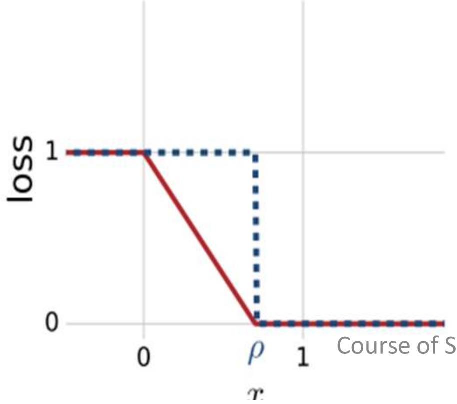
$$L_D\,(h) \leq L_S(\mathrm{h}) + \Re_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$L_D\,(h) \leq L_S(\mathrm{h}) + \widehat{\Re}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Margin loss function

For any ρ>0 , the ρ –margin loss function is he function L_ρ $\mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_+$ defined for all $y, y' \in \mathbb{R}$ by $L_\rho(y, y') = \ \Phi_\rho(yy')$ with,

$$\Phi_\rho(\mathrm{x}) = \min(1; \max\left(0; 1 - \frac{x}{\rho}\right)) = \begin{cases} 1 & if\ x \leq 0 \\ 1 - \frac{x}{\rho} & if\ 0 < x \leq \rho \\ 0 & if\ \rho < x \end{cases}$$



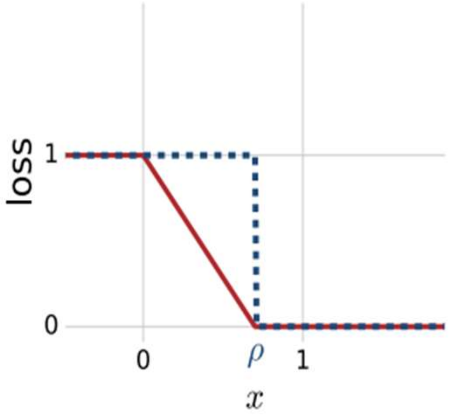
Empirical margin loss

Given a sample$S = (x_1; \dots ; \mathrm{x}_m)$ and a hypothesis h, the empirical margin loss is defined by,

$$L_{s,\rho}(\mathrm{h}) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i))$$

As for any I we have $\Phi_\rho(y_i h(x_i)) \leq 1_{y_i h(x_i) \leq \rho}$,

$$L_{s,\rho}(\mathrm{h}) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) \leq \rho}$$



Φ_ρ is $\frac{1}{\rho}$-Lipschitz

Tragland’s lemma

Let $\Phi_1,\ldots,\ \Phi_m$ be l-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$ and $\sigma_1;\ldots.;\sigma_m$ be Rademacher random variables. Then for any hypothesis set $\mathcal{H}$ of real valued functions, the following inequality holds :

$$\frac{1}{m}\mathbb{E}_\sigma\left[sup_{h\in\mathcal{H}}\left(\sum_{i=1}^{m}\sigma_i(\Phi_i\circ h)(x_i)\right)\right]\leq\frac{1}{m}\mathbb{E}_\sigma\left[sup_{h\in\mathcal{H}}\left(\sum_{i=1}^{m}\sigma_i h(x_i)\right)\right]=l\widehat{\mathfrak{R}}_S(\mathcal{H})$$

In particular, if Φ_i= Φ for all i then,

$$\widehat{\mathfrak{R}}_S(\Phi\circ\mathcal{H})\leq l\widehat{\mathfrak{R}}_S(\mathcal{H})$$

Margin bound for binary classification

Let $\mathcal{H}$ be a set of real-valued functions. Fix $\rho > 0$. Then for any $\delta > 0$, with a probability at least $1 - \delta$, each of the following holds for all h of $\mathcal{H}$:

$$L_D\,(h) \leq L_{s,\rho}(\mathrm{h}) + \frac{2}{\rho}\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$L_D(h) \leq L_{s,\rho}(\mathrm{h}) + \frac{2}{\rho}\widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Theorem

Let $\mathcal{H}$ be a set of real-valued functions. Fix r > 0. Then for any $\delta > 0$, with a probability at least $1 - \delta$, each of the following holds for all $h \in \mathcal{H}$ and ρ $\in]0; 1]$:

$$L_D\ (h) \leq L_{s,\rho}(\mathrm{h}) + \frac{4}{\rho}\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

$$L_D\ (h) \leq L_{s,\rho}(\mathrm{h}) + \frac{4}{\rho}\widehat{\mathfrak{R}}_S(\mathcal{H}) + \sqrt{\frac{\log log_2 \frac{2r}{\rho}}{m}} + 3\sqrt{\frac{\log \frac{4}{\delta}}{2m}}$$

Theorem

Let $S \subseteq \{x : ||x|| < r\}$ be a sample of size m and let $\mathcal{H} = \{x \mapsto w.x \ : \quad ||w|| < \Lambda\}$. Then the empirical Rademacher complexity of $\mathcal{H}$ can be bounded as follows,

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{m}}$$

Corollary

Let $\mathcal{H} = \{x \mapsto w.x \ : \ ||w|| < \Lambda\}$ and assume $\mathrm{X} \subseteq \{x : \left||x|\right| < r\}$. Fix $\rho > 0$. Then for any $\delta > 0$, with a probability at least $1 - \delta$ over the choice of a sample S of size m, the following holds for all $h \ \in \mathcal{H}$:

$$L_D\ (h) \leq L_{s,\rho}(\mathrm{h}) + 2\sqrt{\frac{r^2\Lambda^2}{\rho^2 m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

Corollary

Chosing $\Lambda = 1$, and by generalization of the last corollary to a uniform bound over $\rho \in]0;r]$, for any $\delta > 0$, with probabilty at least $1 - \delta$ the following holds for all $h \in \{x \mapsto w.x : ||w|| < 1\}\, and\ \rho \in]0;r]$:

$$L_D\ (h) \leq L_{s,\rho}(\mathrm{h}) + 4\sqrt{\frac{r^2}{\rho^2 m}} + \sqrt{\frac{\log log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Corollary

For any $\rho > 0$, the ρ-margin loss function is upper bounded by the ρ-hinge loss :

$$\forall x \in \mathbb{R} \quad \Phi_{\rho}(\mathrm{x}) = \min(1; \max\left(0; 1-\frac{x}{\rho}\right)) \leq \max\left(0; 1-\frac{x}{\rho}\right)$$

Thus, with probabilty at least $1-\delta$ the following holds for all $h \in \left\{x \mapsto w.x \ : \left||w|\right| < 1\right\}$ $and\ all\ \ \rho > 0$:

$$L_D\ (h) \leq \frac{1}{m}\sum_{i=1}^{m} \max\left(0; 1-\frac{y_i(w.x_i)}{\rho}\right) + 4\sqrt{\frac{r^2}{\rho^2 m}} + \sqrt{\frac{\log log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Corollary

Since for any $\rho > 0, \frac{h}{\rho}$ admits the same generalization error as h, with probability at least $1 - \delta$, the following holds for all $h \in \{x \mapsto w.x \; : \left||w|\right| < 1\}$ and all $\rho > 0$:

$$L_D\,(h) \leq \frac{1}{m}\sum_{i=1}^{m} \max(0; 1 - y_i(w.x_i)) + 4\sqrt{\frac{r^2}{\rho^2 m}} + \sqrt{\frac{\log log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

ρ is left as a free parameter of the algorithm, typically determined by cross-validation.

For any $\rho > 0$ the bound suggest selecting $\boldsymbol{w}$ as the solution of the following optimization problem :

$$\min_{\left||\boldsymbol{w}|\right|^2 \leq \frac{\mathbf{1}}{\boldsymbol{\rho}^2}} \frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \max(0; 1 - y_i(w.x_i))$$

Corollary

ρ is left as a free parameter of the algorithm, typically determined by cross-validation.

For any $\rho > 0$ the bound suggest selecting $\boldsymbol{w}$ as the solution of the following optimization problem :

$$\min_{\left|\left|\boldsymbol{w}\right|\right|^2 \leq \frac{\boldsymbol{1}}{\boldsymbol{\rho}^2}} \frac{\boldsymbol{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \max(0; 1 - y_i(w.x_i))$$

Introducing the Lagrange variable $\lambda \geq 0$, this optimization problem can be written as :

$$\min_{\boldsymbol{w}} \boldsymbol{\lambda}\left|\left|\boldsymbol{w}\right|\right|^2 + \frac{\boldsymbol{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \max(0; 1 - y_i(w.x_i))$$

Generalization bound for regression – finite set

Let L be a loss function upper bounded by M. Assume that the hypothesis set $\mathcal{H}$ is finite. Then for any $\delta > 0$, with probabilty at least $1 - \delta$ the following inequality holds for all $h \in \mathcal{H}$:

$$L_D\ (h) \leq L_s(\mathrm{h}) + M\sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$$

Notes: An alternative objective function would be based on the empirical margin loss instead of the ρ-hinge loss, however, the latter is advantageous as it is convexe while the margin loss is not.

The bounds just discussed do not depend on directly depend on the dimension of the feature space and guarantee good generalization with a favorable margin.

PDS Kernel – Learning guarantees

In this section we present general learning guarantees for hypothesis sets for SVM combined with PDS kernels**.**

Let $\mathcal{H}$ be a hypothesis set with bounded norm $\mathcal{H} = \{h \in \mathbb{H}: \quad ||h||_{\mathbb{H}} \leq \Lambda\}$ for some $\Lambda \geq 0$ where $\mathbb{H}$ is the RKHS associated to a kernel K.

Any $h \in \mathcal{H}$ is of the form $x \longmapsto < h, K(x,.) > = < h, \Phi(x) >$ with $||h||_{\mathbb{H}} \leq \Lambda$, where Φ is a feature mapping associated to K, that is of the form $x \longmapsto < w, \Phi(x) >$ with $||w||_{\mathbb{H}} \leq \Lambda$

Rademacher complexity of kernel based hypotheses

Let $K : \boldsymbol{\chi} \times \boldsymbol{\chi} \longrightarrow \mathbb{R}$ be a PDS kernel, and let $\Phi : \boldsymbol{\chi} \longrightarrow \mathbb{H}$ be a feature mapping associated to K. Let $S \subseteq \{x : K(x,x) \leq r^2\}$ be a sample of size m, and let $\mathcal{H} = \{x \longmapsto < w, \Phi(x) > : ||w||_{\mathbb{H}} \leq \Lambda\}$ for some $\Lambda \geq 0$. then,

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\Lambda\sqrt{Tr[K]}}{m} \leq \sqrt{\frac{r^2\Lambda^2}{m}}$$

If $K(x,x) \leq r^2$ for all $x \in \boldsymbol{\chi}$, the inequalities above hold for all samples S.

Margin theory for kernel based hypotheses

Let $K : \boldsymbol{\chi} \times \boldsymbol{\chi} \longrightarrow \mathbb{R}$ be a PDS kernel with $r^2 = sup_{x \in \boldsymbol{\chi}} K(x, x)$. Let $\Phi : \boldsymbol{\chi} \longrightarrow \mathbb{H}$ be a feature mapping associated to K. Let $\mathcal{H} = \{x \longrightarrow w.\Phi(x): \quad ||w||_{\mathbb{H}} \leq \Lambda\}$ for some $\Lambda \geq 0$. for a fixed $\rho > 0$, for any $\delta > 0$, with probability at least $1 - \delta$ for any h $\in \mathcal{H}$

$$R(h) \leq \hat{R}_{S,\rho}(x) + 2\sqrt{\frac{r^2\Lambda^2}{\rho^2 m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$$

$$R(h) \leq \hat{R}_{S,\rho}(x) + 2\sqrt{\frac{Tr[K]\Lambda^2}{\rho^2 m}} + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$

Note: The trace of the kernel matrix is an important quantity for controlling the complexity of hypotheses sets based on kernels.