Sparsity of SVMs that use the ε-insensitive loss

Ingo Steinwart
Information Sciences Group
Los Alamos National Laboratory
ingo@lanl.gov

Andreas Christmann
Department of Mathematics
University of Bayreuth
andreas.christmann@uni-bayreuth.de

Abstract
We derive lower and upper bounds for the number of support vectors for SVMs based on the ε-insensitive loss function. It turns out that these bounds are asymptotically tight under mild assumptions on the data-generating distribution. Finally, we briefly discuss a trick that can be used to reduce sparsity. For example, in very sparse datasets the SVM is used to estimate the conditional means.

Basic Assumptions and Notations
Assumptions:
- For all ε > 0, the minimizer \(\hat{f}_n\) of \(L_n(f)\) is defined by \(\hat{f}_n(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \eta(x_i)\right)\).
- The loss function \(L_n(f)\) is defined by \(L_n(f) = \int X \cdot L(f(x)) dP(x)\).

ε-Inensitive SVMs that use the ε-insensitive loss \(L_n(f)\).

Infinite sample version: There exists a unique solution \(f_n\) of the optimization problem
\[\text{minimize} \int X \cdot L(f(x)) dP(x) \text{ subject to } \phi(x) \eta(x) \leq 1, \eta(0) = 1\]
where \(\phi(x)\) and \(\eta(x)\) are non-negative. We always assume that the corresponding support level \(f_n = \eta(x)\) is measurable and bounded with \(|f_n| \leq 1\).

Finite vs. Infinite Sample Solutions

A Basic Property of Support Vectors

Lemma: For all \(i, j \geq 0\), and \(D = \{x_1, x_2, \ldots, x_n\}\), we have
\[\{x_i | (\eta_i(x_j) - x_j) > \epsilon\} \subseteq \{x_i | (\eta_i(x_j) + x_j) > \epsilon\}\]

Conditional Minimizers

Definition: For a distribution \(P\) on \(X \times R\) and \(\epsilon > 0\), the set of conditional minimizers is defined by
\[\mathcal{M}^c(\epsilon) = \{f \in L_n(f) : \int X \cdot L(f(x)) dP(x) < \epsilon\}\]

Illustration:
- For the conditional distribution \(P_{\text{grid}}\), we can write \(\mathcal{M}^c(\epsilon) = \{f \in L_n(f) : \int X \cdot L(f(x)) dP(x) < \epsilon\}\)

Tubes around the Infinite Sample Solution

Lemma: With high probability we have \(\hat{f}_n(x) > \epsilon\)

Consequence: For each \(i\), we have
\[\mathcal{M}^c(\epsilon) = \{f \in L_n(f) : \int X \cdot L(f(x)) dP(x) < \epsilon\}\]

The graph illustrates the first inclusion.

First Result

Theorem: For all \(\epsilon > 0\), \(\delta > 0\), and \(\epsilon > 0\) satisfying \(\epsilon \leq \delta < 1\), we have
\[P(|\hat{f}_n(x) - f(x)| > \delta) < \epsilon\]

Corollary: In particular, we have
\[P(\mathcal{M}^c(\epsilon) \cap \mathcal{M}^c(\delta)) \geq P(\mathcal{M}^c(\epsilon)) - \epsilon\]

Question: What happens if \(\epsilon\) changes with \(\delta\)?

Second Result

Theorem: Assume that the condition \(\mathcal{M}^c(\epsilon) \cap \mathcal{M}^c(\delta)\) is dense in \(L_n(f)\). Then, for all \(\epsilon, \delta > 0\), there exists \(a, b > 0\) such that for all \(x \in [0, 1]\), we have
\[P(\hat{f}_n(x)) = \epsilon\]

Corollary: In particular, we have
\[P(\mathcal{M}^c(\epsilon)) \geq P(\mathcal{M}^c(\delta)) - \epsilon\]

Approximation of the Conditional Minimizers

Lemma: Assume that the distribution \(P\) is dense in \(L_n(f)\). Then, for all \(\epsilon > 0\) there exists \(a, b > 0\) such that for all \(x \in [0, 1]\), we have
\[P(\hat{f}_n(x)) = 0\]

Example: Sparsity for \(\epsilon > 0\)

Assumptions:
- The condition \(\epsilon > 0\) holds for the distribution \(P\).
- \(P\) is dense in \(L_n(f)\).

Corollary: In particular, we have
\[P(\mathcal{M}^c(\epsilon)) \geq P(\mathcal{M}^c(\delta)) - \epsilon\]

Observe: \(L_n(f)\) equals the conditional mean and \(P(\mathcal{M}^c(\epsilon))\) and \(P(\mathcal{M}^c(\delta))\) are the measures defined by \(f\).

Example: No Sparsity for the Absolute Loss

Assumptions:
- The condition \(\epsilon = 0\) and \(P\) is dense in \(L_n(f)\).
- \(P\) is a unique solution \(\hat{f}\) for all \(\epsilon > 0\).

Observe: The conditional minimizer is given by \(\mathcal{M}^c(\epsilon) = \{0\}\). For all \(\epsilon > 0 \in \mathbb{R}\), we have
\[\mathcal{M}^c(\epsilon) = \{0\}\]

Consequence: In particular, we have
\[P(\mathcal{M}^c(\epsilon)) \geq P(\mathcal{M}^c(\delta)) - \epsilon\]

Example: In the case \(\epsilon > 0\), we have
\[\mathcal{M}^c(\epsilon) = \{0\}\]

Observe: \(L_n(f)\) equals the conditional mean and \(P(\mathcal{M}^c(\epsilon))\) is the measure defined by \(f\).

Consequence: The SVM solution \(\hat{f}_n\) is no solution for \(\epsilon > 0\). In particular, for all \(\epsilon > 0\), we have
\[\int X \cdot L(f(x)) dP(x) = P(\mathcal{M}^c(\epsilon)) - \epsilon\]