Representer Theorems

Scribe: Megan Stickler

Question: What properties of a loss function are sufficient to imply the existence and uniqueness of an SVM solution?

1 Background Defintions, Lemmas, and Theorems

Definition 1. A loss function $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is *locally Lipschitz (continuous)* if for all $a \geq 0$ there exists a constant c_a such that for $t, t' \in [-a, a]$,

$$\sup_{x \in X, y \in Y} |L(x,y,t) - L(x,y,t')| \leq c_a |t - t'|.$$

- The smallest constant c_a for which this holds is denoted $|\ell|_{a,1}$
- If $\ell_1 = \sup_{a \geq 0} |\ell|_{a,1} < \infty$, then the loss function L is *Lipschitz (continuous)* with Lipschitz constant ℓ_1

Remarks:

1. If Y is finite (as in, for instance, a classification problem) and the supervised loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is convex, then L is automatically locally Lipschitz.
2. A locally Lipschitz loss is also a Nemitski loss, since

$$\begin{aligned} L(x,y,t) &\leq L(x,y,0) + |L(x,y,t) - L(x,y,0)| \\ &\leq L(x,y,0) + |\ell|_{|t|,1} |t|. \end{aligned} \tag{1}$$

In particular, a locally Lipschitz loss is Nemitski p-integrable $\iff R_{L,P}(\cdot) < \infty$. Furthermore, a Lipschitz loss is also a Nemitski loss of order $p = 1$.

Published: April 13, 2020

Lemma 1. *(Lipschitz continuity of Risks) Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be locally Lipschitz, and let P be a distribution on $X \times Y$.*
Then for all $B \geq 0$ and all $f, g \in L^\infty(P_X)$ such that $\|f\|_\infty, \|g\|_\infty \leq B$, we have

$$|R_{L,P}(f) - R_{L,P}(g)| \leq \|\ell\|_{B,1} \|f - g\|_{L^1(P_X)}.$$

Proof. Fixing $B \geq 0$, $\|f\|_\infty$, $\|g\|_\infty \leq B$ gives us that $|f(x)|, |g(x)| \leq B$ for almost every x and for almost every x, $f(x), g(x) \in [-B, B]$.
L is locally Lipschitz, so this gives that for almost every x,

$$|L(x,y,f(x)) - L(x,y,g(x))| \leq \|\ell\|_{B,1} |f(x) - g(x)|.$$

Now

$$\begin{aligned} |R_{L,P}(f) - R_{L,P}(g)| &= \Big| \int_{X\times Y} L(x,y,(f(x))dP(x,y) - \int_{X\times Y} L(x,y,g(x))dP(x,y) \Big| \\ &= \Big| \int_{X\times Y} (L(x,y,f(x)) - L(x,y,g(x)))dP(x,y) \Big| \\ &\leq \int_{X\times Y} |L(x,y,f(x)) - L(x,y,g(x))| dP(x,y) \\ &\leq \int_{X\times Y} \|\ell\|_{B,1} |f(x) - g(x)| dP(x,y) \\ &= \|\ell\|_{B,1} \int_{X\times Y} |f(x) - g(x)| dP(x,y) \\ &= \|\ell\|_{B,1} \|f - g\|_{L^1(P_X)}. \end{aligned}$$

Definition 2. A loss function $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is *differentiable* if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is differentiable for all $x \in X, y \in Y$. $L'(x,y,t)$ denotes the derivative of $L(x,y,t)$, if such a derivative exists.

Proposition 1. *Let P be a distribution on $X \times Y$ and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a differentiable loss function such that both L and $|L'|$ are p-integrable Nemitski losses (recall that L is always positive). Then the risk $R_{L,P} : L^\infty(P_X) \to [0,\infty)$ is Frechét differentiable and its derivative at $f \in L^\infty(P_X)$ is the bounded linear operator $R'_{L,P} : L^\infty(P_X) \to \mathbb{R}$ given by*

$$R'_{L,P}(f)g = \int_{X\times Y} g(x) L'(x,y,f(x)) dP(x,y)$$

for $g \in L^\infty(P_X)$.

Proof. Let $f \in L^\infty(P_X)$ and let $(f_n) \subset L^\infty(P_X)$ be a sequence such that $f_n \neq 0$, $n \geq 1$, and $\lim_{n\to\infty} \|f_n\|_\infty = 0$. We assume also that $\|f_n\|_\infty \leq 1$ for all $n \geq 1$.
For $x \in X, y \in Y$ we define

$$G_n(x,y) := \begin{cases} \Big|\frac{L(x,y,f(x)+f_n(x)) - L(x,y,f(x))}{f_n(x)} - L'(x,y,f(x))\Big| & f_n(x) \neq 0 \\ 0 & f_n(x) = 0 \end{cases}$$

So then,

$$\left|\frac{R_{L,P}(f+f_n)-R_{L,P}(f)-R'_{L,P}(f)f_n}{\|f_n\|_\infty}\right|$$
$$\leq \int_{X\times Y}\frac{1}{\|f_n\|_\infty}\big|L(x,y,f(x)+f_n(x))-L(x,y,f(x))-f_n(x)L'(x,y,f(x))\big|dP(x,y)$$
$$\leq \int_{X\times Y} G_n(x,y)dP(x,y) \tag{2}$$

Also, by the definitions of G_n and $L'(x,y,\cdot)$, we have

$$\lim_{n\to\infty} G_n(x,y)=0 \tag{3}$$

By the Mean Value Theorem, for $x\in X, y\in Y$ and $n\geq 1$ with $f_n(x)\neq 0$, there exists a $g_n(x,y)$ such that $|g_n(x,y)|\in[0,|f_n(x)|]$ and

$$\frac{L(x,y,f(x)+f_n(x))-L(x,y,f(x))}{f_n(x)}=L'(x,y,f(x)+g_n(x)).$$

Since $|L'|$ is a P-integrable Nemitski loss, there also exist $b: X\times Y\to[0,\infty)$, $b\in L^1(P)$ and increasing function $h:[0,\infty)\to[0,\infty)$ such that

$$|L'(x,y,t)|\leq b(x,y)+h(t).$$

This together with $\|f_n\|_\infty\leq 1$ for $n\geq 1$ gives

$$\left|\frac{L(x,y,f(x)+f_n(x))-L(x,y,f(x))}{f_n(x)}\right|\leq b(x,y)+h(|f(x)+g_n(x,y)|)$$
$$\leq b(x,y)+h(\|f\|_\infty+1).$$

So $G_n(x,y)\leq 2b(x,y)+2h(\|f\|_\infty+1)$. This together with (2), (3), and Lebesgue Dominated Convergence theorem gives us the desired expression for $R'_{L,P}(f)g$.

2 Margin-based losses and Distance-based losses

Motivation: In many problems (most notably SVM), losses are not convex; however, these non-convex loss functions can often be replaced by appropriate convex *'surrogate losses'*.

Definition 3. A supervised loss $L:(Y,\mathbb{R})\to[0,\infty)$ is a *margin-based loss* if there exists a *representing function* $\phi:\mathbb{R}\to[0,\infty)$ such that for $y\in Y, t\in\mathbb{R}$,

$$L(y,t)=\varphi(yt).$$

L is a *distance-based loss* if there exists a representing function $\psi : \mathbb{R} \to [0, \infty)$ with $\Psi(0) = 0$ such that for $y \in Y, t \in \mathbb{R}$,

$$L(y,t) = \psi(y-t).$$

Proposition 2. *Let L be a margin-based loss function with representing function φ. Assume $Y = \{-1, 1\}$ (binary classification problem). Then*

1. *L is (strictly) convex $\iff$ φ is (strictly) convex*
2. *L is continuous $\iff$ φ is continuous*
3. *L is (locally) Lipschitz $\iff$ φ is (locally) Lipschitz*
4. *If L is convex, then it is both Lipschitz and a p-integrable Nemitski loss.*

Examples of Margin-based losses:

- Hinge Loss:
$$L_{hinge}(y,t) = \max\{0, 1 - yt\}$$
  - Convex
  - Lipschitz
  - Hinge loss is a surrogate (convexification) of classification loss.
- Least Squares Loss:
$$\begin{aligned} L_{LS}(y,t) &= (y-t)^2 \\ &= (1-yt)^2 \\ &\text{(since } y = \pm 1) \end{aligned}$$
  - Convex
  - Locally Lipschitz
  - Note that L_{LS} is also an example of a distance-based loss function.
- Truncated Least Squares:
$$L_{Tr}(y,t) = (\max\{0, (1-yt)\})^2$$
  - Convex
  - Locally Lipschitz
- Similar propositions apply in the case of distance-based losses.

3 Existence and Uniqueness of SVM Solutions

Recall: The SVM problem can be formulated as finding the minimizer of

$$R_{L,D,\lambda}(f) = \lambda \|f\|_H^2 + R_{L,D}(f),$$

where $f \in H$ and D are identically distributed data. By the Law of Large Numbers, we expect that $R_{L,D,\lambda}(f)$ is close to

$$R_{L,P,\lambda}(f) = \lambda \|f\|_H + R_{L,P}(f)$$

Question: Does a solution exist? If so, can we represent the solution f in a practical (e.g., computable) form?
We attempt to answer this question with *Representer Theorems*.

Definition 4. Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H a Reproducing Kernel Hilbert Space with measurable kernel k on X, and P a distribution on $X \times Y$. For $\lambda > 0$, a function $f_{P,\lambda,H}$ satisfying

$$\lambda \|f_{P,\lambda,H}\|^2 + R_{L,P}(f_{P,\lambda,H}) = \inf_{f \in H} \lambda \|f\|_H^2 + R_{L,P}(f)$$

is a *general SVM solution*.

Note:

$$\begin{aligned}\lambda \|f_{P,\lambda,H}\|^2 &\leq \lambda \|f_{P,\lambda,H}\|^2 + R_{L,P}(f_{P,\lambda,H}) \\ &\leq R_{L,P}(0).\end{aligned}$$

Hence

$$\|f_{P,\lambda,H}\|_H \leq \sqrt{\frac{1}{\lambda} R_{L,P}(0)}.$$

Theorem 1. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss, P a distribution on $X \times Y$ and H a Reproducing Kernel Hilbert Space of X with a bounded measureable kernel. Then*

1. *If $R_{L,P}(f) < \infty$ for some $f \in H$, then for all $\lambda > 0$ there exists at most one general SVM solution.*
2. *If L is a p-integrable Nemitski loss, then for all $\lambda > 0$ there exists a general SVM solution.*

Proof. 1) Assume that the map $f \to \lambda \|f\|_H^2 + R_{L,P}(f)$ has two minimizers f_1, $f_2 \in H$ such that $f_1 \neq f_2$. Then $\lambda \|f_1\|_H^2 + R_{L,P}(f_1) = \lambda \|f_2\|_H^2 + R_{L,P}(f_2)$. Recalling that $\|\frac{1}{2}(f_1 + f_2)\|_H^2 < \frac{1}{2}\|f_1\|_H^2 + \frac{1}{2}\|f_2\|_H^2$, this with the convexity of $f \to R_{L,P}(f)$ gives that for $f^* := \frac{1}{2}(f_1 + f_2)$,

$$\lambda \|f^*\|_H^2 + R_{L,P}(f^*) < \lambda \|f_1\|_H^2 + R_{L,P}(f_1);$$

that is, f_1 is *not* a minimizer of $f \to \lambda \|f\|_H^2 + R_{L,P}(f)$, and so the assumption that there are two minimizers is false.

2) Since the kernel k is bounded, the map $id : H \to L^\infty(P_X)$ is continuous. The convexity and boundedness of L imply that L is continuous. By prior results, it follows that the map $R_{L,P} : L^\infty(P_X) \to \mathbb{R}$ is a continuous map; hence, $R_{L,P} : H \to \mathbb{R}$

is also continuous. Since L is convex, the map $R_{L,P} : H \to \mathbb{R}$ is also convex. Since $f \to \lambda\|f\|_H^2$ is convex, $f \to \lambda\|f\|_H^2 + R_{L,P}(f)$ is a linear combination of convex functions and is also convex.

Set $A := \{f \in H : \lambda\|f\|_H^2 + R + L, P(f) \leq R_{L,P}(0)\}$. Then $f = 0 \in A$. For $f \in A$, $\lambda\|f\|_H^2 \leq R_{L,P}(0)$, ($R_{L,P} \geq 0$), so $A \subset \left(\sqrt{\frac{1}{\lambda}R_{L,P}(0)}\right)B_H$, where B_H is the closed unit ball on H. By convex analysis, there exists a minimizer $f_{P,\lambda}$ $(= f_{P,\lambda,H})$.

Remark: Convexity of L is not necessary for the existence of a general SVM solution; it was used in the proof, but its absence does not preclude the presence of a solution.

Corollary 1. *Let L be a convex, locally Lipschitz loss, P a distribution on $X \times Y$ with $R_{L,P} < \infty$, and H a measureable Reproducing Kernel Hilbert Space with bounded, measureable kernel k. Then, for all $\lambda > 0$, there exists a unique general SVM solution $f_{P,\lambda,H}$ ($f_{P,\lambda} \in H$).*

Proof. Recall that a locally Lipschitz loss is also a p-integrable Nemitski loss if and only if $R_{L,P}(0) < \infty$. Since $R_{L,P} < \infty$, L is a convex p-integrable Nemitski loss and the hypotheses of the above theorem are satisfied.

- In the textbook, there are special results for margin-based and distance-based losses.

4 Representer Theorems

There are a number of results in the literature providing representation formulas for the SVM solutions.

Theorem 2. *(Representer Theorem for Empirical SVM Solutions) Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and $D = \{(x_1,y_1)...(x_n,y_n)\} \subset X \times Y$. Let H be a Reproducing Kernel Hilbert Space over X. Then, for all $\lambda > 0$, there exists a unique empirical SVM solution $f_{D,\lambda}$ such that*

$$\|f_{D,\lambda}\|_H^2 + R_{L,D}(f_{D,\lambda}) = \inf_{f \in H} \lambda\|f\|_H^2 + R_{L,D}(f)$$

and there exist $\alpha_1...\alpha_n \in \mathbb{R}$ such that

$$f_{D,\lambda}(x) = \Sigma_{i=1}^n \alpha_i k(x,x_i),\ x \in X.$$

Proof. In this case, the convexity of L implies its continuity. Since convergence in H implies pointwise convergence, the continuity of $R_{L,D} : H \to [0,\infty)$ follows from the continuity of L. The existence and uniqueness of the SVM solution $f_{D,\lambda}$ follow from the same arguments as in Theorem 1.

To derive a representation of $f_{D,\lambda}$, let

$$X' = \{x_1 ... x_n\}$$

and

$$H|_{X'} = span\{k(\cdot, x_i) : x_i \in X'\}.$$

$H|_{X'}$ is a Reproducing Kernel Hilbert Space with kernel $k|_{X' \times X'}$, and there exists an empirical SVM solution $f_{D,\lambda,H|_{X'}} \in H|_{X'}$.
For $f \in$ orthogonal complement $(H|_{X'})^{\perp}$, $f(x_i) = \langle f, k(\cdot, x_i) \rangle = 0$ for $x_i \in X'$. Let $P_{X'}$ be the orthogonal projection of $H \to H|_{X'}$ so that

$$R_{L,D}(P_{X'} f) = R_{L,D}(f)$$

and

$$\|P_{X'} f\|_H \leq \|f\|_H.$$

Then

$$\inf_{f \in H} \lambda \|f\|_H^2 + R_{L,D}(f) \leq \inf_{f \in H|_{X'}} \lambda \|f\|_H^2 + R_{L,D}(f)$$

so that

$$\inf_{f \in H} \lambda \|P_{X'} f\|_H^2 + R_{L,D}(P_{X'} f) \leq \inf_{f \in H} \lambda \|f\|_H^2 + R_{L,D}(f).$$

Uniqueness follows the proof of uniqueness from Theorem 1. Suppose there are two unique solutions f_1, f_2 so that

$$\lambda \|f_1\|_H^2 + R_{L,D}(f_1) = \lambda \|f_2\|_H^2 + R_{L,D}(f_2) = \inf_{f \in H} \lambda \|f\|_H^2 + R_{L,D}(f).$$

Then, letting $f^* := \frac{1}{2}(f_1 + f_2)$, by the convexity of $f \to R_{L,D}(f)$ we have

$$\lambda \|f^*\|_H^2 + R_{L,D}(f^*) < \lambda \|f_1\|_H^2 + R_{L,D}(f_1),$$

so

$$\lambda \|f_1\|_H^2 \neq \inf_{f \in H} \lambda \|f\|_H^2 + R_{L,D}(f)$$

and f_1 is not a solution.

Proposition 3. *(Non-trivial solution). Let L be a convex loss function and P a distribution on $X \times Y$ such that L is a p-integrable Nemitski loss. Assume H is a Reproducing Kernel Hilbert Space with a bounded measureable kernel over X with $R^*_{L,P} < R_{L,P}(0)$. Then, for all $\lambda \geq 0$, $f_{P,\lambda} \neq 0$.*

Proof. By the hypotheses, there exists an $f^* \in H$ such that $R_{L,P}(f^*) < R_{L,P}(0)$. By the convexity of $R_{L,P}$, for $\alpha \in [0,1]$ we have

$$\lambda \|\alpha f^*\|_H^2 + R_{L,P}(\alpha f^*) \leq \lambda \alpha^2 \|f^*\|_H^2 + \alpha R_{L,P}(f^*) + (1 - \alpha) R_{L,P}(0) =: h(\alpha).$$

Since $R_{L,P}(f^*) < R_{L,P}(0)$, there exists some $\alpha^* \in (0,1]$ that minimizes $h : [0,1] \to [0,\infty)$ and so

$$\lambda \|\alpha^* f^*\|_H^2 + R_{L,P}(\alpha^* f^*) \leq h(\alpha^*) < h(0) = \lambda \|0\|_H^2 + R_{L,P}(0).$$

Theorem 3. *Let L be a convex, p-integrable Nemitski loss, P a distribution on $X \times Y$, and k a bounded measureable kernel on X with separable Reproducing Kernel Hilbert Space H and canonical feature map $\Phi : X \to H$. Also, assume the derivative of L, $|L'|$, is a p-integrable Nemitski loss. Then, for $\lambda \geq 0$, the general SVM solution $f_{P,\lambda}$ is*

$$f_{P,\lambda}(x) = \frac{1}{2\lambda} \int_{X\times Y} L'(x', y, f_{P,\lambda}(x'))k(x,x')dP(x'Y);$$

that is,

$$f_{P,\lambda} = \frac{-1}{2\lambda} \mathbb{E}_P[L'\Phi].$$

Note: *If L is not differentiable, one can replace L' with a sub-differential of L, which is included as a case in the more general theorem.*

Proof. Let X be a measureable space. Since L is differentiable, the risk function $R_{L,P} : L^\infty(P_{X'}) \to [0,\infty)$ is Frechét differentiable and

$$R'_{L,P}(f)(g) = \int_{X\times Y} g(x)L'(x,y,f(x))dP(x,y).$$

Let H be a separable Reproducing Kernel Hilbert Space with bounded, measureable kernel k and let $\Phi : X \to H$ be the corresponding canonical feature map.
By prior results, the embedding $id : H \to L^\infty(P_{X'})$ is well-defined and continuous so that for $f_0 \in H$,

$$(R_{L,P} \circ id)'(f_0) = R'_{L,P}(f_0) \circ id.$$

Hence, for $f \in H$,

$$\begin{aligned}(R_{L,P} \circ id)'(f_0)f &= R'_{L,P}(f_0) \circ id(f) \\ &= \int_{X\times Y} f(x)L'(x,y,f_0(x))dP(x,y)\end{aligned}$$

Note: Alternatively, one can think of this as

$$\begin{aligned}(R_{L,P} \circ id)'(f_0) &= \mathbb{E}_{(X,Y)}[L'(x,y,f_0(x))\langle f.\Phi(x)\rangle] \\ &= \langle f, \mathbb{E}_{(X,Y)}[L'(x,y,f_0(x))\Phi]\rangle \\ &= i\,\mathbb{E}_{(X,Y)}[L'(x,y,f_0(x))\Phi(x)]\end{aligned}$$

where $i : H \to H'$ is an isomorphism. In this case, f is an element in H so the final expectation $\mathbb{E}_{(X,Y)}$ is an H-valued expectation.

Let $G : H \to \mathbb{R}$ be given by $G(f) = \|f\|_H^2$. The Frechét derivative of G is $G' f_0 = 2if_0$. Let us consider the *regularized loss* $R_{L,P,\lambda} : H \to \mathbb{R}$

$$R_{L,P,\lambda} = \lambda G + R_{L,P} \circ id.$$

The solution $f_{P,\lambda}$ minimizes $R_{L,P,\lambda}$. Hence,

$$\begin{aligned} 0 &= (\lambda G + R_{L,P} \circ id)'(f_{P,\lambda}) \\ &= i(2\lambda f_{P,\lambda} + \mathbb{E}_{(X,Y)}[L'(x,y,f_{P,\lambda}(x))\Phi(x)]). \end{aligned}$$

Thus,

$$2\lambda f_{P,\lambda} = -\mathbb{E}_{(X,Y)}[L'(x,y,f_{P,\lambda}(x))\Phi(x)].$$

This shows that

$$f_{P,\lambda}(x) = \frac{-1}{2\lambda} \int_{X\times Y} L'(x',y,f_{P,\lambda}(x))k(x,x')dP(x',y).$$

For data $D = \{(x_i,y_i)\}_1^N$ with corresponding empirical distribution, from the above expression we derive

$$f_{D,\lambda}(x) = \frac{-1}{2\lambda N} \Sigma_{i=1}^N L'(x_i,y_i,f_{D,\lambda}(x_i))k(x,x_i),$$

showing that the coefficients α_i from the prior formula have the form

$$\alpha_i = \frac{-1}{2\lambda N} L'(x_i,y_i,f_{D,\lambda}(x_i)).$$

References