SVMs: basics, representer theorem

Administrative/Meta:

- **New room!** Siebel 1214.
- Office hours **cancelled** on Monday, 9/26. Instead, we voted to have office hours Saturday, 9/24, during 1-3pm. We will also add extra office hours next weekend.
- No class next wednesday.
- Schedule change: SVMs, since continues with duality theme, and the introduction of risk minimization.
- A very thorough reference for SVMs is Steinwart and Christmann (2008).
- I have an obsession with the minus signs in Fenchel duality, in particular with convex ERM (they've changed in this lecture, versus the last one...).

SVM basics

This lecture serves two purposes: giving a concrete instance of the abstract risk minimization setup from last time, and introducing some key theory of the SVM (support vector machine).

Remark. It's worth asking: now that everyone talks about neural nets all day, why bring up SVMs?

- We have few insights into non-linear function classes, SVMs are one.
- There's still lots of interesting research going on; for instance see "kernel mean embedding" literature.

Let's motivate the form of the SVM optimization problem geometrically; a later lecture will justify our abstract convex risk minimization setup more systematically (lecture "consistency of convex risk minimization"). In particular, this derivation is **just a heuristic**, but still illuminative.

Let data $(x_i)_{i=1}^d$ with $x_i \in \mathbb{R}^d$, and labels $(y_i)_{i=1}^n$ with $y_i \in \{-1, +1\}$ be given. Within the goal of linear classification (finding $w \in \mathbb{R}^d$ and predicting $x \mapsto \mathbf{1}[w^\top x \geq 0]$), a reasonable idea is to be correct with some *margin*, meaning we seek a feasible point to the following problem:

$$\text{find } w \in \mathbb{R}^d \centerdot \|w\|_2 = 1, \forall i \centerdot y_i \langle w, x_i \rangle \geq 1.$$

Geometrically, all points x_i have distance at least 1 from the hyperplane $\{x \in \mathbb{R}^d : \langle w, x \rangle = 0\}$, and the side they fall on depends on y_i. *(Picture drawn in class.)*

It may happen that a hyperplane separating the points exists, but only by some positive distance less than one. An easier objective is

$$\min \left\{ \|w\|_2^2 : w \in \mathbb{R}^d, \forall i \centerdot y_i \langle w, x_i \rangle \geq 1 \right\}.$$

An optimal value $\bar{w}$ will necessarily have examples at distance $\|\bar{w}\|_2$ from $\{x \in \mathbb{R}^d : \langle w, x \rangle = 0\}$.

It's still possible the constraint can fail, due to inputs that are not (strictly) linearly separable (e.g., the "xor" from lecture 2). Thus for every example i, introduce variable ϵ_i for some slack on that example:

$$\min \left\{ \frac{\lambda}{2} \|w\|_2^2 + \sum_i \epsilon_i : w \in \mathbb{R}^d, \epsilon \in \mathbb{R}_+^n, \forall i \centerdot y_i \langle w, x_i \rangle \geq 1 - \epsilon_i \right\}.$$

The parameter $\lambda \geq 0$ lets us trade off between the competing goals of having $\|w\|_2$ small and having $\sum_i \epsilon_i$ small:

- λ large means we have more points on the wrong side of the hyperplane, but also more that are correct and far from the hyperplane.
- λ small means we have more points on the right side of the hyperplane, but little guarantee that many are far from the hyperplane.

As a final simplification, the optimal $\bar{\epsilon}$ satisfies $\bar{\epsilon}_i := \max\{0, 1 - y_i \langle \bar{w}, x_i \rangle\}$. Thus we end up with the regularized ERM problem

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^n \max\{0, 1 - y_i \langle w, x_i \rangle\}.$$

Remarks.

- The (non-convex!) problem

$$\min \left\{ \sum_{i=1}^n \max\{0, 1 - y_i \langle w, x_i \rangle\} : w \in \mathbb{R}^d, \|w\|_2 = 1 \right\}$$

  has a pleasing interpretation: the objective value, for any w with $\|w\|_2 = 1$, is the distance we must move all the points to be distance at least 1 from (and on the correct side of) hyperplane $\{x \in \mathbb{R}^d : \langle w, x \rangle = 0\}$.

- Another convention is to introduce a parameter $b \geq 0$, and learn a predictor $x \mapsto 2 \cdot \mathbf{1}[\langle w, x \rangle \geq b] - 1$. This has all sorts of weird consequences on the learning problem, particularly with the kernel, also b is not regularized, etc.

SVM Duality

Let's return to the optimization setup from last time. Define **hinge loss** ℓ (and per-coordinate variant ℓ_i) as

$$\ell(z) := \max\{0, 1 + z\} \qquad \ell_i(v) := \ell(v_i).$$

Collect all examples $((x_i, y_i))_{i=1}^n$ as the first n rows of matrix $A \in \mathbb{R}^{m \times d}$, meaning $A_{ij} = y_i (x_i)_j$. For now, leave the rows $i \in \{n+1, \ldots, m\}$ undefined; the case $m > n$ will be useful shortly.

With this notation, the SVM primal and dual are as follows.

Theorem (Baby Representer Theorem)**.** Suppose $\lambda > 0$. Then

$$\min \left\{ \sum_{i=1}^n \ell_i(-Aw) + \frac{\lambda}{2}\|w\|_2^2 : w \in \mathbb{R}^d \right\} = \max \left\{ -\sum_{i=1}^n s_i - \frac{1}{2\lambda}\|A^\top s\|_2^2 : s \in [0,1]^n \times \{0\}^{m-n} \right\}.$$

Primal-dual optimal pairs $(\bar{w}, \bar{s})$ always exist. $\bar{s}$ is optimal iff it has the following form:

$$\bar{s} \in \begin{cases} \{0\} & i > n, \\ \{0\} & i \leq n, (A\bar{w})_i > 1, \\ [0,1] & i \leq n, (A\bar{w})_i = 1, \\ \{1\} & i \leq n, (A\bar{w})_i < 1. \end{cases}$$

Lastly, $\bar{w}$ is unique, and has the form $\bar{w} = A^\top \bar{s}/\lambda$.

Remarks.

- **Support vectors.** There may be more than one dual optimum, but all of them are zero on all coordinates except those for which $i \le n$ and $(A\bar{w})_i \le 1$. Expanding the definition of A, then $\bar{s}_i > 0$ implies $y_i \langle \bar{w}, x_i \rangle < 1$, which is consistent with the original geometric development. Such vectors are called **support vectors**.

  These support vectors are consistent with our earlier geometric picture. We were finding a vector w so that $y_i \langle w, x_i \rangle \ge 1$ for all (x_i, y_i). In particular, this means that for any pair with $y_j \langle \bar{w}, x_j \rangle > 1$ (where $\bar{w}$ is optimal), we can wiggle x_j a little and it doesn't affect the optimal choice for $\bar{w}$. Similarly, $\bar{s}$ is completely determined by those eamples which have $(A\bar{w})_i \le 1$.

- **Kernel trick.** Define **gram matrix** $G := AA^\top$; by definition of A, $G_{ij} = y_i y_j \langle x_i, x_j \rangle$. Note that $\|A^\top s\|_2^2 = s^\top G s$, meaning the dual can be written without A. So if we solve the problem in the dual, it seems we never need the full matrices A or G, and can just do pairwise computations on the fly.

  We still seem to need a copy of A at prediction time: our prediction on an example i is $(A\bar{w})_i$. This is the purpose of having $m > n$: let's pack unlabeled examples we care about at indices $i > n$, meaining $A_{ij} = (x_i)_j$ where $i > n$ and we don't know a corresponding y_i. But $\bar{w} = A^\top \bar{s}/\lambda$, thus $(A\bar{w})_i = (AA^T \top \bar{s}/\lambda)_i = (G\bar{s}/\lambda)_i$, so once again at evaluation time we just need G, or rather pairwise evaluations.

  Together, this gives the **kernel trick**: by optimizing SVMs in the dual and then predicting with $A\bar{w}$, we can avoid explicit vector representations of examples and work only with inner products $\langle x_i, x_j \rangle$, which we can define as we wish. **This will be discussed properly in the next lecture,** the presentation I gave threw many people off.

- **Dual optimization.** For a long time, optimizing SVMs was done in the dual, namely via **dual coordinate ascent**, something we'll discuss in a few lectures. There are all sorts of ridiculous tricks here. *[Todo to future matus: tell us some tricks please.]*

- **Uniqueness in the dual.** If G is invertible (namely if it is symmetric positive definite, as it is symmetric positive semi-definite by construction), then the dual is strongly convex and also has a unique optimum.

  On the other hand, it may seem weird that $\bar{w}$ is unique but satisfies $\bar{w} = A^\top \bar{s}/\lambda$, where $\bar{s}$ is *any* dual optimum. The resolution is related to the preceding point on uniqueness in the dual: the degree of freedom in $\bar{s}$ is washed out by A, meaning $A^\top \bar{s}$ is still unique.

- **Representer theorem.** This theorem is usually presented as follows. If we make the matrix A bigger and bigger (while holding the training sample fixed), we can few it as a mapping into an infinite dimensional space of functions; namely Aw is now a function, and $(Aw)_i$ is its value on point i, where we have infinitely many options for i. As we'll see shortly, the theorem still goes through, and the essential bit is that we can still write $(A\bar{w}) = (G\bar{s}/\lambda)_i$, meaning we can **represent** the optimum by just the nonzero entries (and where they appear) of $\bar{s}$.

Note: some more material was presented after this, but the presentation wasn't great; see the next lecture for a cleaned-up version.

[matus notes to future self: can prove infdim attainment via double-duality. should include something about primal-dual error gaps for approximate optima.]

References

Steinwart, Ingo, and Andreas Christmann. 2008. *Support Vector Machines.* 1st ed. Springer.