Lecture 17

Margin Analysis for Classification

Sasha Rakhlin

Nov 05, 2018

Outline

Last few bits from last lecture

Large Margin Theory for Classification

Constraints on values vs parameters

We discussed $\widehat{\mathscr{R}}_n(G)$ when G is a set with a nice description, such as a ball B_p^n. However, it is not clear whether G has a nice description when $\mathcal{G}$ is a nicely described class of functions.

Examples:

- Let $\mathcal{G} = \{z \mapsto \langle w, z\rangle : w \in B_2^d\}$ and $\|z_i\| \le 1$. We can show that

$$\widehat{\mathscr{R}}_n(\mathcal{G}|_{z_1,\dots,z_n}) \le \frac{1}{\sqrt{n}}$$

  By luck, the upper bound coincides with $\widehat{\mathscr{R}}_n(B_2^n)$, but $\mathcal{G}|_{z_1,\dots,z_n} \neq B_2^n$.
- Let $\mathcal{G} = \{z \mapsto \langle w, z\rangle : w \in B_1^d\}$. Assume $\|z_i\|_\infty \le 1$. We can show that

$$\widehat{\mathscr{R}}_n(\mathcal{G}|_{z_1,\dots,z_n}) \le \sqrt{\frac{\log d}{n}}$$

  Contrast with $\widehat{\mathscr{R}}_n(B_1^n) = 1/n$.

Notation: we shall write $\mathscr{R}(\mathcal{G}) \triangleq \mathbb{E}\widehat{\mathscr{R}}_n(\mathcal{G}|_{Z_1,\dots,Z_n})$.

Outline

Last few bits from last lecture

Large Margin Theory for Classification

Classification with Real-Valued Functions

Typical binary classification methods use sign of a *real-valued function* to make prediction:

$$\mathbb{I}(\mathcal{F}) = \{x \mapsto \text{sign}(f(x)) : f \in \mathcal{F}\}$$

According to previous lecture, sample complexity can be understood by considering sizes of sets G obtained by evaluating functions in $\mathbb{I}(\mathcal{F})$ on data.

Unfortunately, this typically leads to overly pessimistic results since it is difficult to use the structure of $\mathcal{F}$. VC dimension of $\mathbb{I}(\mathcal{F})$ can be very large, yet in practice the methods work well. What is an alternative explanation?

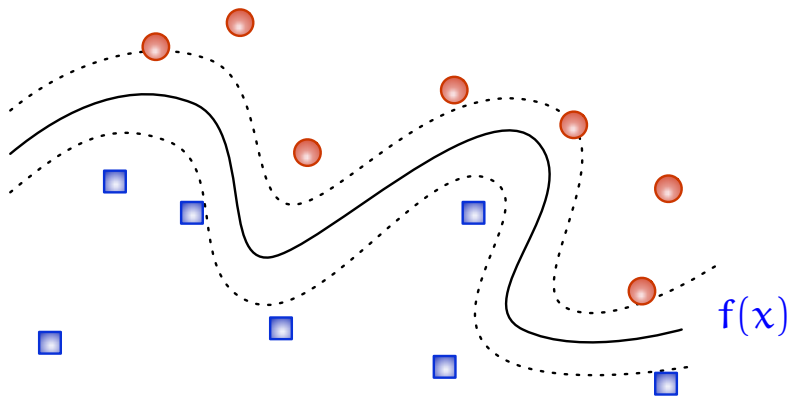
Example: $f(x) = f_w(x) = \langle w, \psi(x) \rangle$ where ψ is a mapping to a high-dimensional feature space. The VC dimension is large (equal to the dimensionality of $\psi(x)$) or infinite, yet the methods perform well!

The conundrum can be addressed by something we already studied – the margin. Just as in Perceptron, we will aim to find a way around dimensionality and use margin + complexity of $\mathcal{F}$ instead.

Margins

Hard margin:

$$\exists f \in \mathcal{F} : \quad \forall i, \quad y_i f(x_i) \geq \gamma$$



Soft margin: we hope to have

$$\exists f \in \mathcal{F} : \quad \frac{\text{card}(\{i : y_i f(x_i) < \gamma\})}{n} \quad \text{is small}$$

Surrogate Loss

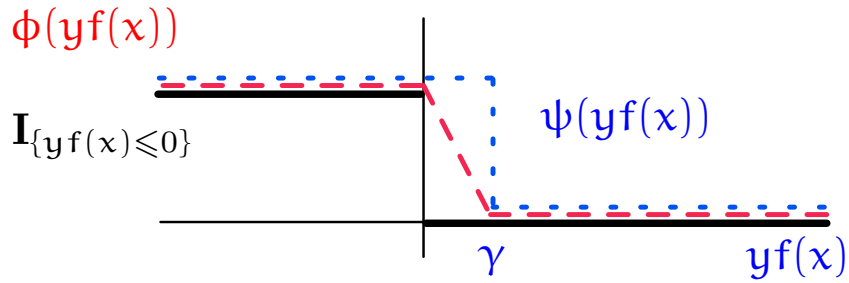
Define

$$\phi(s) = \begin{cases} 1 & \text{if } s \le 0 \\ 1 - s/\gamma & \text{if } 0 < s < \gamma \\ 0 & \text{if } s \ge \gamma \end{cases}$$

Then

$$\mathbf{I}\{yf(x) \le 0\} \le \phi(yf(x)) \le \psi(yf(x)) \triangleq \mathbf{I}\{yf(x) \le \gamma\}$$

The function ϕ is an example of a *surrogate loss function.*



Let

$$\mathbf{L}_\phi(f) = \mathbb{E}\phi(yf(x)) \qquad \text{and} \qquad \widehat{\mathbf{L}}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i f(x_i))$$

Surrogate Loss

Now consider uniform deviations for the surrogate loss:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbf{L}_\phi(f) - \widehat{\mathbf{L}}_\phi(f) \right\}$$

We have shown that this quantity is at most $2\mathscr{R}(\phi(\mathcal{F}))$ for

$$\phi(\mathcal{F}) = \left\{ (x, y) \mapsto \phi(yf(x)) : f \in \mathcal{F} \right\}$$

Observe that in our example ϕ is $1/\gamma$-Lipschitz. Hence,

$$\mathbb{E} \max_{f \in \mathcal{F}} \left\{ \mathbf{L}_\phi(f) - \widehat{\mathbf{L}}_\phi(f) \right\} \leq \frac{2}{\gamma} \mathscr{R}(\mathcal{F})$$

by the "contraction" property from previous lecture.

Margin Bound

Note:

$$\mathbf{L}_{01}(f) \le \mathbf{L}_\phi(f), \quad \widehat{\mathbf{L}}_\phi(f) \le \widehat{\mathbf{L}}_\psi(f)$$

Hence,

$$\mathbb{E} \max_{f \in \mathcal{F}} \left\{ \mathbf{L}_{01}(f) - \widehat{\mathbf{L}}_\psi(f) \right\} \le \frac{2}{\gamma} \mathscr{R}(\mathcal{F})$$

Rewriting, for any algorithm $\widehat{f}_n$ that takes values in $\mathcal{F}$,

$$\mathbb{E}\mathbf{L}_{01}(\widehat{f}_n) \le \mathbb{E}\widehat{\mathbf{L}}_\psi(\widehat{f}_n) + \frac{2}{\gamma} \mathscr{R}(\mathcal{F})$$

Recall: $\widehat{\mathbf{L}}_\psi(f) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{y f(x) \le \gamma\}$ is the fraction of γ-margin errors.

Conclusion: expected zero-one out-of-sample error is controlled by the proportion of margin mistakes made by $\widehat{f}_n$ and Rademacher averages of $\mathcal{F}$, scaled by $1/\gamma$. We avoided the (potentially) large VC dimension of $\mathbb{I}(\mathcal{F})$!

Margin Bound: High Probability

A high probability version of this bound would read: for all $f \in \mathcal{F}$

$$\mathbf{L}_{01}(f) \leq \widehat{\mathbf{L}}_{\psi}(f) + \frac{c}{\gamma}\widehat{\mathscr{R}}_n(\mathcal{F}|_{X_{1:n}}) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Of course, in practice we have in front of us *one* dataset. But, given the above statement, we are $(1-\delta)$-confident that for our data and our method, the expected error is no more than proportion of margin errors + complexity of the model.

NB: we can choose the best γ on the data to optimize the bound (paying extra for a union bound).

Perceptron

We now deduce a "non-separable" analogue for Perceptron.

Indeed, for Perceptron, $\mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\| = 1\}$, and it is easy to show

$$\mathscr{R}(\mathcal{F}) \leq \frac{D}{\sqrt{n}}$$

where $D = \max \|X_i\|$.

If there are no γ-margin mistakes made, our margin bound yields

$$\mathbb{E}\mathbf{L}_{01}(\widehat{f}_n) \leq \frac{2D}{\gamma} \times \frac{1}{\sqrt{n}}$$

This bound does not give the right "rate" (should be square of what we got), but, on the other hand, our margin bound is much more general and does not require separability as in Perceptron.

NB: it is possible to get a margin bound that does reduce to Perceptron with correct rate, but this is more difficult.

Algorithmic consequences

It is fair to ask: what are the practical applications of our bounds? One answer is that complexity notions can be used as regularizers. For instance, suppose the set of linear classifiers in Perceptron example is indexed by radius B:

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : \|w\| \leq B\}.$$

Then $\mathscr{R}(\mathcal{F}) \leq \frac{B}{\sqrt{n}}$ and the margin bound (omitting constants and other factors) reads

$$\mathbf{L}_{01}(w) \lesssim \widehat{\mathbf{L}}_{\psi}(w) + \frac{c}{\gamma} \frac{B}{\sqrt{n}} + \dots$$

With a bit more work (essentially, a union bound), we can turn this into

$$\mathbf{L}_{01}(w) \lesssim \widehat{\mathbf{L}}_{\psi}(w) + \frac{c}{\gamma} \frac{\|w\|}{\sqrt{n}} + \dots$$

Minimizing the upper bound suggests minimizing margin mistakes while keeping small the norm $\|w\|$ of the solution. This is essentially SVM.

Algorithmic consequences

General Prescription: get an upper bound on Rademacher averages of your favorite class (indexed by a "complexity radius"). Turn it into a regularizer. Try your new algorithm on data.

Morally, a sensible regularizer should be related to an upper bound on Rademacher averages in some way, since its role is to control the estimation error. (ok, there are many asterisks here, but it's good to see the connection)

Summary

- Complexity is a subtle notion: margin vs dimensionality of space. VC theory may be too pessimistic here.
- Bound is a-posteriori: if it happens that proportion of γ-margin errors is small, then out-of-sample performance is good
- Empirically-defined bounds can be optimized. Leads to new regularization methods.
- Do neural nets maximize margin in any sense?