CS281B/Stat241B. Statistical Learning Theory. Lecture 23.

Peter Bartlett

Overview

- Risk bounds for SVMs.
  - Rademacher averages.
- Gradient descent for SVMs.
  - Regret bounds.
  - “Pegasos”

Risk bounds for SVMs

Consider an SVM-like criterion:

$$\min_{f\in\mathcal{H}} \qquad \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{C}{n}\sum_{i=1}^{n}\ell\left(f(x_i), y_i\right).$$

Instead of this regularized empirical risk minimization, we consider a constrained version:

$$\begin{aligned} \min_{f\in\mathcal{H}} \qquad & \frac{1}{n}\sum_{i=1}^{n}\ell\left(f(x_i), y_i\right) \\ \text{s.t.} \qquad & \|f\|_{\mathcal{H}}^2 \leq B^2. \end{aligned}$$

In fact, this is always equivalent, for a suitable choice of the constant B.

Risk bounds for SVMs

Also, notice that choosing $f = 0$ shows that

$$\min_{f \in \mathcal{H}} \qquad \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{C}{n}\sum_{i=1}^{n} \ell\left(f(x_i), y_i\right) \leq \frac{C}{n}\sum_{i=1}^{n} \ell\left(0, y_i\right).$$

Hence, the solution f^* of the regularized problem satisfies

$$\frac{1}{2}\|f\|_{\mathcal{H}}^2 \leq \frac{C}{n}\sum_{i=1}^{n} \ell\left(0, y_i\right).$$

For instance, for hinge loss, the right hand side is C. Thus, we are always restricted to a ball in a RKHS.

Risk bounds for SVMs

We have seen that minimizing the sample average of the loss leads to near minimal expected loss, provided the Rademacher averages of the loss class are small. And if ℓ is 1-Lipschitz in the predictions $f(x_i)$, then for

$$F = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B\},$$
$$\ell(F) = \{(x,y) \mapsto \ell(f(x),y) : f \in F\},$$

we have $\mathbb{E}\|R_n\|_{\ell(F)} \leq 2\mathbb{E}\|R_n\|_F + c/\sqrt{n}$. (Here, c depends on a bound on $\ell(0,y)$.)

Risk bounds for SVMs

Theorem: For an RKHS $\mathcal{H}$ with reproducing kernel k, define $F = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B\}$. For a sample $X_1, \ldots, X_n$,

$$\mathbb{E}\left[\, \|R_n\|_F \,|\, X_1, \ldots, X_n\right] \leq \frac{B}{\sqrt{n}} \sqrt{\frac{\mathrm{tr}(K)}{n}},$$

where $K_{ij} = k(X_i, X_j)$.

Recall that the trace of a matrix is $\mathrm{tr}(K) = \sum_i K_{ii} = \sum_i k(x_i, x_i)$.

Risk bounds for SVMs

Theorem: If $\mathcal{H}$ is a RKHS of functions on compact $\mathcal{X}$ that has a continuous kernel k, P is a probability distribution on $\mathcal{X}$, and λ_j are the eigenvalues of the integral operator

$$T_k f(\cdot) = \int_{\mathcal{X}} k(\cdot, x) f(x) \, dP(x),$$

and $F = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B\}$, then

$$\mathbb{E}\|R_n\|_F \leq B\sqrt{\frac{\mathbb{E}k(X,X)}{n}} = B\sqrt{\frac{\sum_{j=1}^{\infty} \lambda_j}{n}}.$$

Risk bounds for SVMs: Proof

Since k is the reproducing kernel,

$$
\begin{aligned}
\|R_n\|_F &= \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \\
&= \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left\langle k(X_i, \cdot), f \right\rangle \right| \\
&= \sup_{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le B} \left| \left\langle \frac{1}{n} \sum_{i=1}^{n} \epsilon_i k(X_i, \cdot), f \right\rangle \right| \\
&= B \frac{\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i k(X_i, \cdot) \right\|^2}{\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i k(X_i, \cdot) \right\|} \\
&= B \sqrt{\frac{1}{n^2} \sum_{i,j} \epsilon_i \epsilon_j k(X_i, X_j)}.
\end{aligned}
$$

Risk bounds for SVMs: Proof

Applying Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}\left[\,\|R_n\|_F \,|\, X_1,\ldots,X_n\right] &\le B\sqrt{\mathbb{E}\left[\left.\frac{1}{n^2}\sum_{i,j}\epsilon_i\epsilon_j k(X_i,X_j)\right| X_1,\ldots,X_n\right]} \\
&= B\sqrt{\frac{1}{n^2}\sum_i k(X_i,X_i)} \\
&= \frac{B}{\sqrt{n}}\sqrt{\frac{\mathrm{tr}(K)}{n}}.
\end{aligned}
$$

Risk bounds for SVMs: Proof

Applying Jensen's inequality again, we have

$$\mathbb{E}\|R_n\|_F \leq \frac{B}{\sqrt{n}}\sqrt{\mathbb{E}k(X,X)}.$$

Using the decomposition $k(x,y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(u)\psi_j(v)$, we have

$$\begin{aligned}\mathbb{E}\|R_n\|_F &\leq \frac{B}{\sqrt{n}}\sqrt{\mathbb{E}k(X,X)} \\ &= \frac{B}{\sqrt{n}}\sqrt{\sum_j \lambda_j \mathbb{E}\psi_j(X)^2} \\ &= \frac{B}{\sqrt{n}}\sqrt{\sum_j \lambda_j},\end{aligned}$$

because the ψ_j are orthonormal in $L_2(P)$, so $\mathbb{E}\psi_j(X)\psi_i(X) = 1[i=j]$.

Regret bounds for hinge loss

Consider the online convex optimization problem with hinge loss and a Euclidean norm constraint on the parameter vector θ:

$$\begin{aligned}
\ell_t(\theta) &= \left(1 - y_t\theta^T x_t\right)_+, \\
\|x_t\| &\le R, \\
y_t &\in \{-1, 1\}, \\
\theta &\in \Theta, \\
\Theta &= \{\theta : \|\theta\| \le B\}.
\end{aligned}$$

Regret bounds for hinge loss

We have seen (see Lecture 15) that projected gradient descent gives $\sqrt{n}$ regret

Theorem: For $G = \max_t \|\nabla \ell_t(\theta_t)\|$ and $D = \mathrm{diam}(\Theta)$, the gradient strategy:

$$\theta_t := \Pi_\Theta(\theta_t - \eta \nabla \ell_t(\theta_t)),$$

with $\eta = D/(G\sqrt{n})$ has regret satisfying

$$\hat{L}_n - L_n^* \le GD\sqrt{n}.$$

Thus,

$$\frac{1}{n}\sum_{t=1}^{n}\left(1 - y_t\theta_t^T x_t\right)_+ - \min_{\theta\in\Theta}\frac{1}{n}\sum_{t=1}^{n}\left(1 - y_t\theta^T x_t\right)_+ \le \frac{GD}{\sqrt{n}} = \frac{2RB}{\sqrt{n}}.$$

Regret bounds for SVMs

Suppose we augment the loss function with a regularization term:

$$\begin{aligned} \ell_t(\theta) &= \frac{\lambda}{2}\|\theta\|^2 + \left(1 - y_t\theta^T x_t\right)_+, \\ \|x_t\| &\leq R, \\ y_t &\in \{-1, 1\}, \\ \theta &\in \mathbb{R}^d. \end{aligned}$$

Regret bounds for SVMs

Since ℓ_t is λ-strongly convex wrt squared Euclidean distance, we can use gradient descent (mirror descent with squared Euclidean regularizer) with $\eta_t = 2/(t\lambda)$ to show that

$$\frac{1}{n}\sum_{t=1}^{n} \ell_t(\theta_t) \leq \min_{\theta} \frac{1}{n}\sum_{t=1}^{n} \ell_t(\theta) + O\left(\frac{G^2 \log n}{\lambda n}\right).$$

That is,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} \left(1 - y_t \theta_t^T x_t\right)_+ + \frac{\lambda}{2n}\sum_{t=1}^{n} \|\theta_t\|^2 \\
&\quad \leq \min_{\theta} \left(\frac{1}{n}\sum_{t=1}^{n} \left(1 - y_t \theta^T x_t\right)_+ + \frac{\lambda}{2}\|\theta\|^2\right) + O\left(\frac{G^2 \log n}{\lambda n}\right).
\end{aligned}$$

PEGASOS

(PEGASOS=Primal Estimated sub-GrAdient SOlver for SVM)

We can use an online convex optimization method like online gradient descent to design a fast approximate solver for the SVM QP:
The regret bound holds for any sequence of (x_t, y_t) pairs. (Given a fixed sample, we can, for example, choose a sequence uniformly at random from the sample.) Since the ℓ_t are convex, we can take

$$\overline{\theta} = \frac{1}{n} \sum_{t=1}^{n} \theta_t$$

and we have a good approximate solution to the SVM QP:

PEGASOS

$$\frac{1}{n}\sum_{t=1}^{n}\left(1 - y_t\overline{\theta}^T x_t\right)_+ + \frac{\lambda}{2}\|\overline{\theta}\|^2$$
$$\leq \min_\theta \left(\frac{1}{n}\sum_{t=1}^{n}(1 - y_t\theta^T x_t)_+ + \frac{\lambda}{2}\|\theta\|^2\right) + O\left(\frac{G^2\log n}{\lambda n}\right).$$

(And it’s possible to use concentration to show that a uniform random choice gives a similar—only a constant factor worse—bound on the solution to the original QP over the full sample.)

Kernel version of PEGASOS

The representer theorem tells us that, for data $(x_1, y_1), \ldots, (x_n, y_n)$, we can write the SVM QP as

$$\min_{\theta} \qquad \frac{1}{n}\sum_{t=1}^{n}(1 - y_t(K\alpha)_t)_+ + \frac{\lambda}{2}\alpha^T K\alpha,$$

where $K_{ij} = k(x_i, x_j)$. And we can use a similar stochastic gradient approach: choose an (x_t, y_t) pair uniformly, compute the gradient of the corresponding loss ℓ_t, and use it to update the α vector.