CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 27**

Model Selection and Oracle Inequalities

Lecturer: Peter Bartlett *Scribe: Nate Coehlo*

1 The Setup

Often one fits a model by empirical risk minimization: Given a loss function L and a class of functions $\mathcal{F}_\lambda$, one chooses $\hat{f}_\lambda = \arg\min_{f\in\mathcal{F}_\lambda}(\hat{L}(f))$. In the context of linear regression with square error loss, this corresponds to fitting by least squares.

If you are asked to choose between several models, empirical risk minimization is not a good idea.

If $\mathcal{F}_{\lambda_1} \subseteq \mathcal{F}_{\lambda_2}$, $\quad \hat{f}_{\lambda_1} = \arg\min_{f\in\mathcal{F}_{\lambda_1}}(\hat{L}(f))$, $\quad \hat{f}_{\lambda_2} = \arg\min_{f\in\mathcal{F}_{\lambda_2}}(\hat{L}(f))$, Then $\hat{L}(\hat{f}_{\lambda_2}) \leq \hat{L}(\hat{f}_{\lambda_1})$

That is, empirical risk minimization likes complex models because they provide better in-sample fits. To prevent this over-fitting, regularization is often used.

Our model selection framework is as follows:

1. Let $\mathcal{M} = \{\mathcal{F}_\lambda : \lambda \in \Omega\}$ be our model space where Ω is some ordered set and $(\lambda_1 \leq \lambda_2 \Rightarrow \mathcal{F}_{\lambda_1} \subseteq \mathcal{F}_{\lambda_2})$
2. Let $C_n(\lambda)$ be a regularization term that is non-decreasing in λ.
3. Choose $\hat{f} = \hat{f}_{\hat{\lambda}} = \arg\min_{\lambda\in\Omega}(\hat{L}(\hat{f}_\lambda) + C_n(\lambda))$

Intuitively, the regularization term helps regulate the bias / variance tradeoff.

Example. Ridge Regression and the Lasso

Consider the case of linear regression on q variables.
Let $\mathcal{F}_\lambda = \{(1, x_1, x_2, ..., x_q)^T\beta : ||\beta||_p \leq \lambda\}$
Let $C_n(\lambda) = c\lambda$ for some $c \geq 0$.
In words, we regularize proportional to the L_p norm of the coefficients.
Now let $p = 2$, $p = 1$ to see that Ridge Regression and the Lasso fall in this model selection framework.

In general, $\hat{f}$ in 3 above may not be well defined. However, it is well defined for most practical choices of Model, Loss Function, and Regularization Term. When 3 is well defined, $\mathcal{F}_{\hat{\lambda}}$ is our model of choice, and $\hat{f}_{\hat{\lambda}}$ is the empirical risk minimization estimate within that model.

2 An Oracle Inequality

Consider the case with only countably many models: $\Omega = \{1, 2, 3, ...\}$

Definition. The $n^{-1/2}$ Risk Estimation Condition

A risk estimate $r_{n,k}$ satisfies the $n^{-1/2}$ risk estimation condition if there exists constants b and c such that for all choices of $n \in \{1, 2, ...\}$ and $\epsilon \geq 0$ we have

$$P(L(\hat{f}_k) > r_{n,k} + \epsilon) \leq ce^{-2bn\epsilon^2}$$

Theorem 2.1. Suppose $r_{n,k}$ satisfies the $n^{-1/2}$ risk condition with constants b and c. Define

$$C_n(k) = r_{n,k} - \hat{L}(\hat{f}_k) + \sqrt{\frac{\log(k)}{nb}}$$

Then

$$\mathbb{E}(L(\hat{f}_{\hat{k}})) - L^* \leq \inf_k \left(\inf_{f \in \mathcal{F}_k} (L(f) - L^*) + \mathbb{E}(C_n(k)) \right) + \sqrt{\frac{\log(2ce)}{2nb}}$$

Proof. (Sketch) $L^* \equiv \inf_f (L(f)) = \inf_k \left(\inf_{f \in \mathcal{F}_k} (L(f)) \right) \equiv \inf_k (L^*_k)$, so it suffices to show that for any k,

$$\mathbb{E}(L(\hat{f}_{\hat{k}})) - L^*_k \leq \inf_{f \in \mathcal{F}_k} (L(f) - L^*) + \mathbb{E}(C_n(k)) + \sqrt{\frac{\log(2ce)}{2nb}} \ ,$$

where $L^*_k = \inf_{f \in \mathcal{F}_k} L(f)$. We define

$$\tilde{r}_{n,k} = \hat{L}(\hat{f}_k) + C_n(k) = r_{n,k} + \sqrt{\frac{\log(k)}{nb}}$$

From the $n^{-1/2}$ risk estimation condition, we get

$$\begin{aligned}
\Pr\left(L(\hat{f}_{\hat{k}}) - \tilde{r}_{n,\hat{k}} > \epsilon\right) &\leq \Pr\left(\sup_k \left(L(\hat{f}_k) - \tilde{r}_{n,k}\right) > \epsilon\right) \\
&\leq \sum_{k=1}^{\infty} \Pr\left(L(\hat{f}_k) > r_{n,k} + \epsilon + \sqrt{\frac{\log k}{bn}}\right) \\
&\leq \sum_k c \exp\left(-2bn\left(\epsilon + \sqrt{\frac{\log k}{bn}}\right)^2\right) \\
&\leq 2ce^{-2bn\epsilon^2} \qquad (\because \sum_k \frac{1}{k^2} \leq 2)
\end{aligned}$$

Then decompose as follows:

$$\mathbb{E}(L(\hat{f}_{\hat{k}})) - L^*_k = \left(\mathbb{E}\left(L(\hat{f}_{\hat{k}}) - \inf_j(\tilde{r}_{n,j})\right)\right) + \left(\mathbb{E}\left(\inf_j(\tilde{r}_{n,j}) - L^*_k\right)\right)$$

First term in RHS is bounded by $\sqrt{\log(2ce)/2bn}$ from the exponential tail bound for $L(\hat{f}_{\hat{k}}) - \tilde{r}_{n,k}$ proved above. For the second term,

$$\mathbb{E}\Big(\inf_j(\tilde{r}_{n,j}) - L_k^*\Big) \leq \mathbb{E}\Big((\tilde{r}_{n,k}) - L_k^*\Big) = \mathbb{E}\Big(\hat{L}(\hat{f}_k) + C_n(k) - L_k^*\Big)$$
$$\leq \mathbb{E}\Big(\hat{L}(f_k^*) + C_n(k) - L_k^*\Big) = L(f_k^*) + \mathbb{E}\Big(C_n(k)\Big) - L_k^* = \mathbb{E}\Big(C_n(k)\Big)$$

□

3 Examples

3.1 Hold Out Set

Suppose we have the following i.i.d. copies of (X, Y)

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), \quad (X_1^{'}, Y_1^{'}), (X_2^{'}, Y_2^{'}), ..., (X_{nb}^{'}, Y_{nb}^{'})$$

Suppose that we train on the (X_i, Y_i) then test on the $(X_i^{'}, Y_i^{'})$

Let $r_{n,k} = \frac{1}{nb}\sum_{i=1}^{nb} 1_{[\hat{f}_k(X_i^{\prime}) \neq Y_i^{\prime}]}$, i.e. the empirical risk of $\hat{f}_k$ on the validation set.

Now note that $1_{[\hat{f}_k(X_i^{\prime}) \neq Y_i^{\prime}]}$ are i.i.d. in $[0, 1]$ with expectation $L(\hat{f}_k)$ so Hoeffding implies that

$$P(L(\hat{f}_k) > r_{n,k} + \epsilon) \leq e^{-2bn\epsilon^2}$$

Thus, $r_{n,k}$ satisfies the $n^{-1/2}$ Risk Estimation Condition with $c = 1$ and b related to fraction of the observations in the test set. In five fold cross validation we would have $b = 1/4$.

3.2 Structural Risk Minimization

If $\mathcal{F} \subseteq \{-1, 1\}^{\mathcal{X}}$ and 0/1 loss is being used, then

$$r_{n,k} = \hat{L}(\hat{f}_k) + \sqrt{\frac{d_{VC}(F_k)\log(2n+1) + \log(4)}{n}}$$

satisfies the $n^{-1/2}$ Risk Estimation Condition with $b = 1/2$ and $c = 1$.

4 Universal Consistency

Suppose $Y = \{-1, 1\}$ and we are using 0/1 loss.

Definition. A method $f_n : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \longrightarrow \mathcal{Y}$ is Universally Consistent if $\forall P$ on $\mathcal{X} \times \mathcal{Y}$ we have

$$L(f_n) \xrightarrow{P} L^* \ .$$

Consider an f_n that chooses from $\mathcal{F} = \cup_{\lambda>0}\mathcal{F}_\lambda$. Suppose that

1. $\forall P$, $\lim_{\lambda\to\infty} \inf_{f\in\mathcal{F}_\lambda} L_\phi(f) = L^*_\phi(f)$ (Universal Approximation)
2. We have an oracle inequality:

$$\lim_{n\to\infty} \mathbb{E}(L_\phi(f_n) - L^*) \leq \lim_{n\to\infty} \inf_{\substack{\lambda>0 \\ f\in\mathcal{F}}} \left(L_\phi(f) - L^*_\phi + \mathbb{E}(C_n(\lambda)\right)$$

3. The complexity penalty decreases with n. That is, there exists a sequence $\lambda_n \to \infty$ such that $\mathbb{E}(C_n(\lambda_n)) \to 0$
4. ϕ is Classification Calibrated

Then

$$L(f_n) \xrightarrow{P} L^*$$