Rates of Convergence for Cross Validation

Akshay Krishnamurthy
akshaykr@cs.cmu.edu

September 15, 2011

Suppose we have data \( \{X_i, Y_i\}_{i=1}^{2n} \) where \( Y_i = m(X_i) + \epsilon_i \) and we want to do something like Kernel regression to learn the function \( m \). Moreover, suppose that the data points \( X_i \in \mathbb{R}^d \) come from a manifold of intrinsic dimension \( r \) and that \( m \) is \( \beta \)-smooth with respect to the manifold. For kernel regression we know that the MSE decays at a rate of \( n^{-\frac{2\beta}{2\beta+2d}} \), but under the assumption that the data come from the manifold, we would like to bring this rate down to \( n^{-\frac{2\beta}{2\beta+2r}} \). Finally, for simplicity of our argument, suppose that \( m(x) \) is bounded, that is \( m(x) < B \) for all \( x \) and that \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \). In this document we will show that using cross validation to pick the correct bandwidth for kernel regression will get us down to the better rate.

Given our data set (recall that we have \( 2n \) data points), we partition into two groups, a training set \( T \triangleq \{X_i, Y_i\}_{i=1}^{2n} \) and a validation set \( V \triangleq \{X_i, Y_i\}_{i=n+1}^{2n+1} \). We have a set of possible bandwidths \( H \triangleq \{n^{-\frac{2\beta}{2\beta+2r}} : 1 \leq i \leq d\} \). Fairly simple calculations (see for example [1]) that the bias scales like \( O(h^{2\beta}) \) and the variance scales like \( O(\frac{1}{nh^2}) \) since we know that \( m \) is smooth with respect to the manifold. Balancing these expressions reveals that he optimal bandwidth is the \( h^* = n^{-\frac{2\beta}{2\beta+2r}} \) and notice that \( h^* \in H \).

The procedure is to train a kernel density estimate \( h_m^T \) on the training data for each \( h \in H \). We choose \( h \triangleq \arg\min_{h \in H} \hat{R}^V (m_h^T) \) so that \( m_h^T \) is the kernel density estimator that minimizes the empirical risk on the validation data. Here \( \hat{R}^V (m_h^T) \) is the empirical risk of \( m_h^T \), evaluated on the validation data, so \( \hat{R}^V (m_h^T) = \frac{1}{n} \sum_{i=n+1}^{2n} (Y_i - m_h^T(X_i))^2 \).

**Theorem 0.1.**

\[ \mathbb{E}[R(m_h^T)] = O_P(n^{-\frac{2\beta}{2\beta+2r}} + n^{-1/2} \log n) \tag{1} \]

which for \( r > 2\beta \) is \( O_P(n^{-\frac{2\beta}{2\beta+2r}}) \). Here expectations are taken with respect to the training and validation datasets.

**Proof.** First, we’ll show that \( \hat{R}^V (m_h^T) \) concentrates around \( R(m_h^T) \), or, in words, that the cross-validation scores get closer and closer to the true risk of the estimator \( m_h^T \). Since the estimator is trained on an independent data set, we have that \( R(m_h^T) = \mathbb{E}_V [\hat{R}^V (m_h^T)] \), and we can use standard concentration of measure results (such as Hoeffding’s inequality) here. However, before we can use Hoeffding’s we need to show that the random variables (in this case \( Y_i - m_h^T(X_i))^2 \) are bounded, or at least bounded with high probability:

\[ (Y_i - m_h^T(X_i))^2 = (m(X_i) - m_h^T(X_i) + \epsilon_i)^2 \leq 4B^2 + 4B\sigma^2 (\log n + \log(2/\delta_1)) \triangleq C \]

which follows from the fact that the max of \( n \) gaussian random variables is bounded by \( \sigma \sqrt{2 \log n + 2 \log(2/\delta_1)} \) with probability \( \delta_1 \). Note that we bounded some cross terms involving \( \epsilon_i \) by \( Bc^2 \).

Now we can use Hoeffding’s inequality. Specifically:

\[ \mathbb{P} \left( |\hat{R}^V (m_h^T) - R(m_h^T)| > \epsilon \right) = \mathbb{P} \left( \frac{1}{n} \sum_{i=n+1}^{2n} (Y_i - m_h^T(X_i))^2 - R(m_h^T) > \epsilon \right) \leq 2 \exp \left( -\frac{2\epsilon^2}{C^2} \right) \triangleq \delta \]

Using a union bound over all of the estimators \( m_h^T \) and flipping this around, we get that with probability \( \geq 1 - \delta \):

\[ |R(m_h^T) - \hat{R}^V (m_h^T)| \leq \sqrt{\frac{C^2 \log(2|H|/\delta)}{2n}} = O_P(n^{-1/2} \log n) \]
This expression holds for all $h \in \mathcal{H}$, because of the union bound. The $\log n$ term comes from the fact that $C = O_P(\log n)$. We only care about the dependence on $n$, so from now on will just use the $O_P$ notation. Now we look at the minimizer of the CV score.

$$R(m_h^T) \leq \hat{R}^V(m_h^T) + O_P(n^{-1/2}\log n) \leq \hat{R}^V(m_{h^*}^T) + O_P(n^{-1/2}\log n)$$

By the fact that $\hat{h}$ is the bandwidth that minimizes the cross validation risk. Next we take expectations, first with respect to the validation data $V$, and then with respect to the training data $T$.

$$\mathbb{E}_V[R(m_h^T)] = R(m_h^T) \leq \mathbb{E}_V[\hat{R}^V(m_h^T)] + O_P(n^{-1/2}\log n) = R(m_h^T) + O_P(n^{-1/2}\log n)$$

$$\mathbb{E}[R(m_h^T)] \leq \mathbb{E}_T[R(m_h^T) + O_P(n^{-1/2}\log n)] \leq \mathbb{E}_T[R(m_{h^*}^T) + O_P(n^{-1/2}\log n)] = O_P(n^{-\frac{2\alpha}{2\alpha + \beta}}) + O_P(n^{-1/2}\log n)$$

In the first line we used that in expectation the empirical risk on the validation data is equal to the true risk (by independence). We also used the concentration result twice, first as we did before to relate the true risk of $m_h^T$ to the empirical risk of $m_h^T$, and again to related the empirical risk of $m_h^T$ to the true risk of $m_h^T$. When we take expectations with respect to the training data $T$, we get that the expected risk of our estimator is upper bounded by the risk of the estimator with the correctly chosen bandwidth, plus some additive term. Noticing that the estimator $m_{h^*}^T$ has expected risk that converges at rate $O_P(n^{-\frac{2\alpha}{2\alpha + \beta}})$ gives us the result.

The problem with this result is that if $r$ is too small in comparison to $\beta$, then we are left with a convergence at rate $O_P(n^{-1/2})$. This limit arises because we use Hoeffding’s inequality to show obtain convergence between the empirical and true risk at this rate, and we must carry this through our analysis. A better rate can be derive if we use one of Bernstein’s inequalities instead of Hoeffding’s inequality, although the analysis is slightly more complicated. We prove this tighter rate here:

**Theorem 0.2.**

$$\mathbb{E}(R(m_h^T)) = O_P(n^{-\frac{2\alpha}{2\alpha + \beta}})$$

where expectation is taken with respect to both the training data and the validation data.

**Proof.** For this result we use a version of Berstein’s inequality, called the Craig-Bernstein Inequality:

**Lemma 0.3.** Let $X_1, \ldots, X_n$ be random variables and suppose that:

$$\mathbb{E}(|X_i - \mathbb{E}[X_i]|^k) \leq \frac{\text{var}(X_i)}{2} k! r^{k-2}$$

for some $r > 0$. Then with probability $> 1 - \delta$:

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \leq \frac{\log(1/\delta)}{nt} + \frac{t\text{var}(X_i)}{2(1-c)}$$

for $0 \leq tr \leq c < 1$.

Let $U_i^h \triangleq (Y_i - m_h^T(X_i))^2 - (Y_i - m(X_i))^2$. Then it is easy to see that $\sum_{i=n+1}^{2n} U_i^h = \hat{R}^V(m_h^T) - \hat{R}^V(m)$, and moreover we have that $\mathbb{E}[U_i^h] = R(m_h^T) - R(m)$. We would like to use the Craig-Bernstein inequality to show that $U_i^h$ concentrates around $\mathbb{E}[U_i^h]$, and then a union bound to show this for all choices of $h$. To do this, we first need to calculate the variance of $U_i^h$.

$$\text{var}(U_i^h) \leq \mathbb{E}[(U_i^h)^2] = \mathbb{E}[(Y_i - m_h^T(X_i))^2 - (Y_i - m(X_i))^2]^2$$

$$= \mathbb{E}[(m(X_i) - m_h^T(X_i)] + 2\epsilon_i(m(X_i) - m_h^T(X_i) - \epsilon_i)^2$$

$$= \mathbb{E}[(m(X_i) - m_h^T(X_i)^4 + 4\epsilon_i (m(X_i) - m_h^T(X_i))^3 + 4\epsilon^2 (m(X_i) - m_h^T(X_i))]$$

$$\leq 4B^2 \mathbb{E}[(m(X_i) - m_h^T(X_i)^2] + 4\sigma^2 \mathbb{E}[(m(X_i) - m_h^T(X_i))]$$

$$= 4(B^2 + \sigma^2) \mathbb{E}[U_i^h]$$
With this bound on the variance, we now apply the Craig-Bernstein Inequality (our random variables $U_i$ is a sum of bounded and gaussian random variables, and it is easy to show that the moment inequality is satisfied for them):

$$
\frac{1}{n} \sum_{i=n+1}^{2n} U_i - \mathbb{E}[U_i] \leq \frac{\log(1/\delta)}{nt} + \frac{2t(B^2 + \sigma^2)\mathbb{E}[U_i]}{1 - c}
$$

We need to ensure that $c < 1$. To do this, let $c = tr = 8t(B^2 + \sigma^2)/15$ and let $t < 15/(8(B^2 + \sigma^2))$, then it’s easy to see that $c < 1$. Now grouping terms we get:

$$(1 - a)(R(m^T_h) - R(m)) - (\hat{R}^V(m^T_h) - \hat{R}^V(m)) \leq \frac{\log(1/\delta)}{nt}$$

where $a = 2t(B^2 + \sigma^2)/(1 - c) < 1$, with probability $> 1 - \delta$. Now taking a union bound over the $d$ choices for $h$, we have that for all $h$:

$$R(m^T_h) - R(m) \leq \frac{1}{1 - a} \left[ \hat{R}^V(m^T_h) - \hat{R}^V(m) + \frac{\log(d/\delta)}{nt} \right]$$

With this uniform bound established, we can now do the same trick as we did before. By the fact that $\hat{h}$ is the bandwidth that minimizes the risk on the validation dataset, we have that:

$$R(m^T_{\hat{h}}) - R(m) \leq \frac{1}{1 - a} \left[ \hat{R}^V(m^T_{\hat{h}}) - \hat{R}^V(m) + \frac{\log(d/\delta)}{nt} \right]$$

(5)

where $h^*$ is the best choice of bandwidth. Now taking expectations with respect to the validation dataset and then with respect to the training dataset:

$$\mathbb{E}_V[R(m^T_{\hat{h}}) - R(m)] \leq \frac{1}{1 - a} \left[ \mathbb{E}_V[\hat{R}^V(m^T_{\hat{h}}) - \hat{R}^V(m)] + \frac{\log(d/\delta)}{nt} \right] = \frac{1}{1 - a} (R(m^T_{\hat{h}}) - R(m)) + O_P(1/n)$$

(7)

$$\mathbb{E}[R(m^T_{\hat{h}}) - R(m)] \leq \frac{1}{1 - a} \mathbb{E}_T[R(m^T_{\hat{h}}) - R(m)] + O_P(1/n) \leq O_P(n^{-\frac{2}{d+\gamma}} + n^{-1}) = O_P(n^{-\frac{2}{d+\gamma}})$$

(8)

Here we first used the uniform bounds we derived above, and that $m^T_{\hat{h}}$ is the estimator that minimizes the cross-validation risk. Then we used that the true risk of the optimal estimator decays at rate $O_P(n^{-\frac{2}{d+\gamma}})$ which we know from standard analysis of kernel regression. Finally, we used that both $t$, $a$ and $d$ are constants that don’t depend on $n$, so they do not appear in our final rate.

References