CSC 665: Model Selection

Chicheng Zhang

October 22, 2019

1 Error decomposition in machine learning

Setup:

1. distribution D,
2. training examples S drawn iid from D
3. learning algorithm $\mathcal{A}$ that outputs $\hat{h}$ from hypothesis class $\mathcal{H}$, based on S

Question: what are the factors that contribute to the generalization error of $\hat{h}$?
Denote by $h' \triangleq \arg\min_{h\in\mathcal{H}} \mathrm{err}(h, S)$, $h^\star \triangleq \arg\min_{h\in\mathcal{H}} \mathrm{err}(h, D)$. We have the following theorem.

Theorem 1. *With probability* $1-\delta$*,*

$$\mathrm{err}(\hat{h}, D) \le \epsilon_{gen} + \epsilon_{opt} + \mathrm{err}(h^\star, D) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

where $\epsilon_{gen} = \mathrm{err}(\hat{h}, D) - \mathrm{err}(\hat{h}, S)$ *is called the* generalization gap*,* $\epsilon_{opt} = \mathrm{err}(\hat{h}, S) - \mathrm{err}(h', S)$ *is called the* optimization error*.*

Proof. Observe that

$$\mathrm{err}(\hat{h}, D) = [\mathrm{err}(\hat{h}, D) - \mathrm{err}(\hat{h}, S)] + [\mathrm{err}(\hat{h}, S) - \mathrm{err}(h^\star, S)] + [\mathrm{err}(h^\star, S) - \mathrm{err}(h^\star, D)] + \mathrm{err}(h^\star, D).$$

note that the first term is ϵ_{gen}; the second term is at most ϵ_{opt}, as $\mathrm{err}(h', S) \le \mathrm{err}(h^\star, S)$; the third term is at most $\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$ with probability $1-\delta$ by Hoeffding's inequality. □

Remarks:

1. $\mathrm{err}(h^\star, D)$ is called the *bias* of the hypothesis class $\mathcal{H}$. A more expressive $\mathcal{H}$ gives a smaller bias.
2. When m is reasonably large, then $\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$ can usually be omitted.
3. The bound can be loose: aside from application of Hoeffding's inequality, the only other place we use inequality is bounding $\mathrm{err}(h', S)$ using $\mathrm{err}(h^\star, S)$ - but the gap between them can be large: if the data is highly noisy and $\mathcal{H}$ is too expressive, then $\mathrm{err}(\hat{h}, S)$ can be close to zero, whereas $\mathrm{err}(h^\star, S)$ can be large.

Important special case: ERM. Suppose $\mathcal{A}$ is ERM wrt $\mathcal{H}$. In this case, $\hat{h} = h'$, therefore ϵ_{opt} is zero. Moreover, as we have seen before, we can bound ϵ_{gen} by $\sup_{h \in \mathcal{H}} \text{err}(\hat{h}, D) - \text{err}(\hat{h}, S)$, which in turn can be controlled by $O(\sqrt{\frac{\ln \frac{|\mathcal{H}|}{\delta}}{m}})$ using uniform convergence arguments. We have that with probability $1 - \delta$:

$$\text{err}(\hat{h}, D) \leq \text{err}(h^\star, D) + 2\sqrt{\frac{\ln \frac{2|\mathcal{H}|}{\delta}}{2m}}.$$

There are two possible factors that contribute to $\hat{h}$'s error:

1. The bias of $\mathcal{H}$.
2. The "complexity" of $\mathcal{H}$: an upper bound of the generalization gap of $\hat{h}$.

This is called the bias-complexity tradeoff. Say $\mathcal{H} \subset \mathcal{H}'$, then $\mathcal{H}'$ has a smaller bias, while having a larger complexity.

1. Underfitting: $\mathcal{H}$ is too restricted so that the bias is too large. This can sometimes be caught by observing that $\text{err}(\hat{h}, S)$ is too large, as $\text{err}(\hat{h}, S) \leq \text{err}(h^\star, S) \approx \text{err}(h^\star, D)$.
2. Overfitting: $\mathcal{H}$ is too expressive so that the generalization gap is too large. This cannot be directly caught by monitoring the training error rate of ERM, however, it can be caught by maintaining a separate *validation set*. Suppose we have a fresh validation set V, then $\text{err}(\hat{h}, D) \approx \text{err}(\hat{h}, V)$ by Hoeffding's inequality, so $\epsilon_{\text{gen}} \approx \text{err}(\hat{h}, S) - \text{err}(\hat{h}, V)$.

2 How to choose hypothesis class $\mathcal{H}$ in practice?

1. PAC learning theory deals with learning a fixed hypothesis class $\mathcal{H}$
2. In practice (exploratory data analysis), it is often not the case that analysts "commits" to a fixed learning algorithm - one often tries different learning algorithms for different hypothesis classes (e.g. SVM for training linear classifiers with different regularization parameters, ID3 for training decision trees with different pruning strategies, backprop for training neural nets with different learning rates / weight decay, etc) to see which one performs the best.
3. How shall we choose the learning algorithm to use in practice?
4. For simplicity, let us consider only algorithm that are ERMs over hypothesis classes.

Given hypothesis classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$. For every i, define $h_i^\star = \arg\min_{h \in \mathcal{H}} \text{err}(h, D)$ as the optimal classifier on $\mathcal{H}_i$; $\hat{h}_i = \arg\min_{h \in \mathcal{H}_i} \text{err}(h, S)$ is the output of ERM over $\mathcal{H}_i$. How do we select which $\hat{h}_i$ to pick to have low generalization error? Can we certain model selection criteria via the error decomposition theorem?

Idea 1: validation. As discussed before, a fresh validation set can help us provide good evaluation on the trained classifiers. Suppose V is a validation set of size n. Let $\hat{\mathcal{H}} = \left\{\hat{h}_1, \ldots, \hat{h}_k\right\}$ be the set of ERMs. Define $\hat{h} = \arg\min_{h \in \hat{\mathcal{H}}} \text{err}(h, V)$ as the ERM over the ERMs. We have the following simple theorem:

Theorem 2. *With probability* $1 - \delta$,

$$\begin{aligned}
\text{err}(\hat{h}, D) &\leq \min_{i \in \{1,\ldots,k\}} \text{err}(\hat{h}_i, D) + 2\sqrt{\frac{\ln \frac{4}{\delta}}{2n}} \\
&\leq \min_{i \in \{1,\ldots,k\}} \left(\text{err}(h_i^\star, D) + 2\sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{2m}} + 2\sqrt{\frac{\ln \frac{4}{\delta}}{2n}} \right)
\end{aligned}$$

The proof of this theorem follows from simple analysis of ERMs, which is left as an exercise.

Suppose $n = \Theta(m)$, then the third term is dominated by the second term (complexity of $\mathcal{H}_i$), implying that $\hat{h}$'s error upper bound is almost the same as the error upper bound of doing ERM over $\mathcal{H}_i$ (had we known the "best" i - the one that has the best bias-complexity tradeoff).

Idea 2: structural risk minimization (penalized ERM). There is an alternative approach (inspired by theory) that achieves roughly the same type of error guarantee as validation. Note that selecting i that minimizes $\text{err}(\hat{h}_i, S)$ may be a terrible idea, as $\hat{h}_i$ may overfit. However, we can do the following fix: we add penalty that depends on $\mathcal{H}_i$'s complexity, that is,

$$\hat{i} = \arg \min_{i \in \{1,\ldots,k\}} \text{err}(\hat{h}_i, S) + 2\sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{2m}},$$

and define the final output as $\hat{h} = \hat{h}_{\hat{i}}$. Note that similar to SVM, this can be interpreted as regularized loss minimzation - for different classifiers, in addition to minimizing its empirical error, we also add a penalty that depends on which hypothesis class it lies in.

As we will see next, this approach implicitly minimizes the generalization error bounds on all $\hat{h}_i$'s.

Theorem 3. *With probability* $1 - \delta$,

$$\begin{aligned}
\text{err}(\hat{h}, D) &\leq \min_{i \in \{1,\ldots,k\}} \left(\text{err}(\hat{h}_i, D) + 2\sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{2m}} \right) \\
&\leq \min_{i \in \{1,\ldots,k\}} \left(\text{err}(h_i^\star, D) + 4\sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{2m}} \right).
\end{aligned}$$

The first inequality is called an *oracle inequality*, in that it relates the performance of a learned classifier to a classifier output by some ideal learning algorithm (that relies on information unavailable in reality). To see this, note that we can define i_0 that minimizes $\text{err}(\hat{h}_i, D) + 2\sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{2m}}$, which is unavailable as $\text{err}(\hat{h}_i, D)$ cannot be exactly computed. The theorem tries to relate the generalization error of $\hat{h}$ to that of $\hat{h}_{i_0}$.

Proof. By Hoeffding's inequality and union bound, with probability $1 - \delta$,

$$|\text{err}(h, S) - \text{err}(h_i, D)| \leq \sqrt{\frac{\ln \frac{2k|\mathcal{H}_i|}{\delta}}{2m}}, \forall h \in \mathcal{H}_i. \tag{1}$$

Note that this is a *non-uniform convergence* statement: classifiers in larger hypothesis classes's error concentration are controlled more loosely.

Therefore, for every i in $\{1, \ldots, k\}$,

$$\begin{aligned}
\text{err}(\hat{h}, D) = \text{err}(\hat{h}_{\hat{i}}, D) &\leq \text{err}(\hat{h}_{\hat{i}}, S) + \sqrt{\frac{\ln \frac{4k|\mathcal{H}_{\hat{i}}|}{\delta}}{2m}} \\
&\leq \text{err}(\hat{h}_i, S) + \sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{2m}} \\
&\leq \text{err}(\hat{h}_i, D) + 2\sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{m}} \\
&\leq \text{err}(h_i^\star, D) + 4\sqrt{\frac{\ln \frac{4k|\mathcal{H}_i|}{\delta}}{m}}.
\end{aligned}$$

where the first inequality is by error concentration in $\mathcal{H}_{\hat{i}}$; the second inequality is by the optimality of $\hat{i}$; the third inequality is from by error conentration in $\mathcal{H}_i$; where the last step is by the familiar analysis of ERM on $\mathcal{H}_i$ given uniform convergence (1) holds. The theroem is concluded by noting that the above holds for all i. $\square$