CSC 588: Machine learning theory **Spring 2021**

Lecture 11: Error decomposition in supervised learning & Model selection

Lecturer: Chicheng Zhang *Scribe: Jie Bian*

In previous class, we are dealing with fixed hypothesis class $\mathcal{H}$, in this class, we are talking about how we can choose the classifier helpful for learning from the unfixed hypothesis class $\mathcal{H}$.

1 Error decomposition in supervised learning

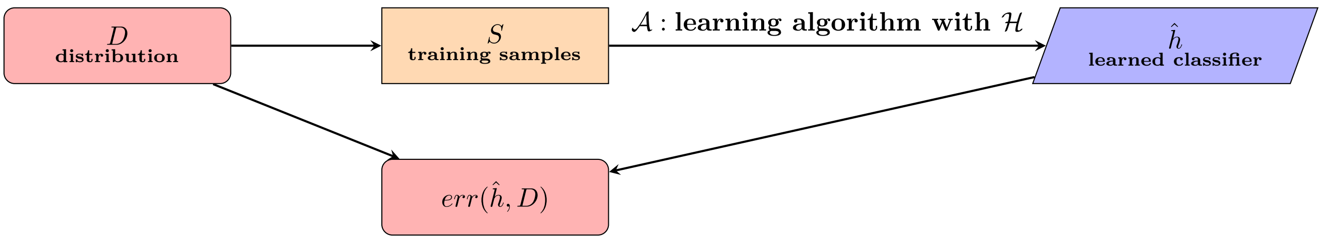

Figure 1: supervised learning pipeline

Q: What are some important factors that contribute to the generalization error of $\hat{h}$?

1. representativeness of training example
2. complexity of $\hat{h}$ ($\mathcal{H}$)
3. optimization accuracy of $\mathcal{A}$
4. expressiveness of $\mathcal{H}$ relative to D

Notation:

$$h' = \operatorname*{argmin}_{h \in \mathcal{H}} err(h, S)$$

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} err(h, D)$$

Theorem 1. *With probability 1-δ,*

$$err(\hat{h}, D) \leq \varepsilon_{gen}{}^{1} + \varepsilon_{opt}{}^{2} + err(h^*, D)^{3} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}{}^{4}$$

where generalization error is defined as $\varepsilon_{gen} = err(\hat{h}, D) - err(\hat{h}, S)$ *, optimization error is defined as* $\varepsilon_{opt} = err(\hat{h}, S) - err(h', S)$

[1] factor 2 is reflected here
[2] factor 3 is reflected here
[3] factor 4 is reflected here
[4] factor 1 is reflected in this inequality

Proof.

$$
\begin{aligned}
err(\hat{h}, D) &= err(\hat{h}, S) + \varepsilon_{gen} \\
&= err(h', S) + \varepsilon_{opt} + \varepsilon_{gen} \\
&= err(h^*, S) + \varepsilon_{opt} + \varepsilon_{gen} + (err(h', S) - err(h^*, S)) \\
&\leq^5 err(h^*, D) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} + \varepsilon_{opt} + \varepsilon_{gen} + (err(h', S) - err(h^*, S)) \\
&\leq^6 err(h^*, D) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} + \varepsilon_{opt} + \varepsilon_{gen}
\end{aligned}
$$

□

Remark:

1. $err(h^*, D)$ is called the bias of $\mathcal{H}$ on D
2. when m is large, $\sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$ can be ignored
3. tightness of the above bound:

Theorem 2. *$err(h^*, S) - err(h', S)$ can be quite large, in this case, at least one of ε_{gen} and ε_{opt} would be large.*

Proof. From error decomposition we have:

$$err(\hat{h}, D) \leq \varepsilon_{gen} + \varepsilon_{opt} + \sqrt{\frac{1}{m}} + (err(h', S) - err(\hat{h}, S) + err(h^*, D)$$

since $err(h^*, D) \leq err(\hat{h}, D)$,

$$err(h^*, S) - err(h', S) \leq \varepsilon_{gen} + \varepsilon_{opt} + \sqrt{\frac{1}{m}}$$

If $err(h^*, S) - err(h', S)$ is large and $\sqrt{\frac{1}{m}}$ is small, then $\varepsilon_{gen} + \varepsilon_{opt}$ is large, therefore at least one of ε_{gen} and ε_{opt} would be large. □

Example 1. *$\mathbb{P}(Y = 1|X)$ is shown in figure 3, we also have:*

$$\mathbb{P}_X(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & otherwise \end{cases}$$

$$\mathcal{H} = \{2\mathbb{I}(x \in \cup_{i=1}^{k} [a_i, b_i]) : k \in \mathbb{N}, a_i, b_i \in [0, 1] \ \forall i\}$$

Therefore, the optimal classifier with the minimum generalization error is:

$$h^* = 2\mathbb{I}(x \in [0.5, 1]) - 1$$

The responding error is:

$$err(h^*, D) = \min_{h \in \mathcal{H}} err(h, D) = 0.2$$

[5] from Hoeffding's and is $O(\sqrt{\frac{1}{m}})$ loose

[6] can be quite loose

From Hoeffding's:

$$err(h^*, S) \geq 0.2 - \sqrt{\frac{1}{m}}$$

For any sample set S (as shown in figure 2), if we assign each classifier to each sample point and make sure any two classifiers' intervals not overlap,then we can find such classifier minimizing the training set error:

$$err(h', S) = \min_{h \in \mathcal{H}} err(h, S) = 0$$

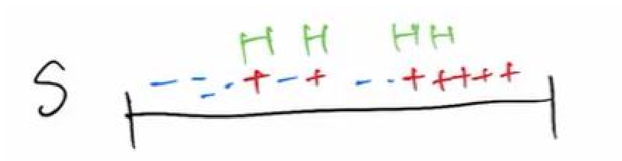

Figure 2: schematic diagram of sample set

Therefore

$$err(h^*, S) - err(h', S) \geq 0.2 - \sqrt{\frac{1}{m}}$$

$err(h^*, S) - err(h', S)$ *is large.*

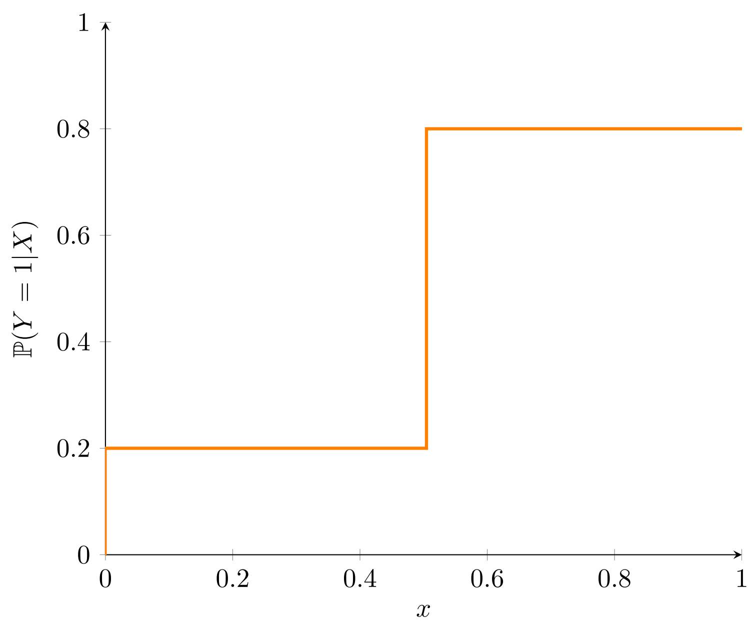

Figure 3: example1

Ways to bring down $\varepsilon_{opt}, \varepsilon_{gen}, err(h^*, D)$:

ε_{opt} $\downarrow$:change the ML optimization algorithm;make $\mathcal{H}$ simple to optimize

ε_{gen} $\downarrow$:choose a less expressive $\mathcal{H}$;collect more samples

$err(h^*, D)$ $\downarrow$:choose a more expressive $\mathcal{H}$

Important special case:$\mathcal{A} = ERM(\mathcal{H})$ Then $\hat{h} = h'(ERM)$, $\varepsilon_{opt=0}$, with $\varepsilon_{gen} \le \sqrt{\frac{\ln \frac{|\mathcal{H}|}{\delta}}{m}}$ and Theorem 1, we have:

$$err(\hat{h}, D) \le err(h^*, D)^7 + 2\sqrt{\frac{\ln \frac{2|\mathcal{H}|}{\delta}}{2m}}^8$$

This is called the bias-complexity tradeoff.

underfitting:this occurs when bias is too large

Example 2.

$$\mathcal{H} = \{linear\ classifier\}$$

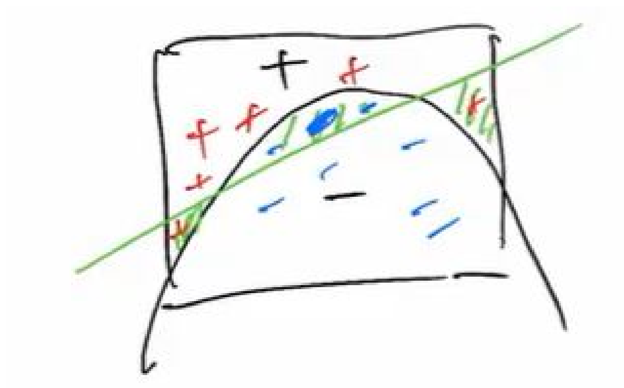

Figure 4: linear classifier on unlinear samples

Sometimes, underfitting can be caught by seeing $err(\hat{h}, S)$ is too large.Reason of this is $err(\hat{h}, S)$ is too large$\Rightarrow$ $err(h^*, S)$ is also large, and with hoeffding's, $err(h^*, S) \approx err(h^*, D)$,then $err(\hat{h}, S)$ is too large$\Rightarrow$ $err(h^*, D)$ is also large.

overfitting:this occurs when $|\mathcal{H}|$ is too large so that complexity term is too large, this also means the generalization error $\varepsilon_{gen} = err(\hat{h}, D) - err(\hat{h}, S)$ is large

Example 3.

$$err(\hat{h}, S) = 0$$

[7]bias

[8]complexity of $\mathcal{H}$

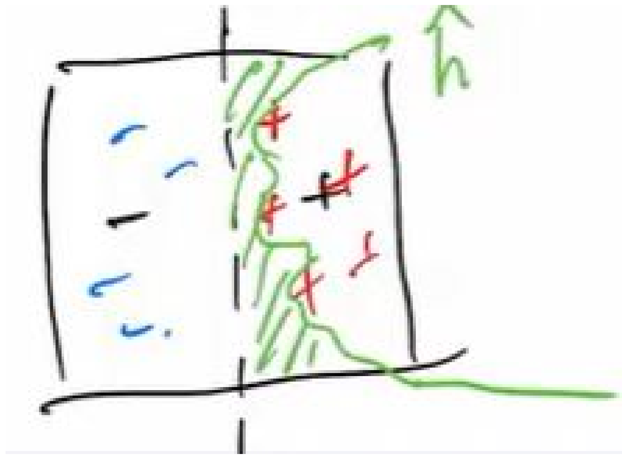

Figure 5: nonlinear classifier on linear samples

we can detect overfitting by using fresh validation set V ,that is because by Hoeffding's, $err(\hat{h}, D) \approx err(\hat{h}, V)$, therefore

$$\varepsilon_{gen} \approx err(\hat{h}, V) - err(\hat{h}, S)$$

By checking if $err(\hat{h}, V) - err(\hat{h}, S)$ is large we can detect overfitting.

2 Model Selection

- **How can we choose a good learning algorithm in practice?**
- **To make it simple,we only consider ERM over hypothesis classes.**

Setup: $\mathcal{H}_1, ..., \mathcal{H}_k (\mathcal{H}_i = \{\text{decision tree with depth} \leq i\})$

$$h_i^* = \underset{h \in \mathcal{H}_i}{\operatorname{argmin}}\, err(h, D)$$

$$\hat{h}_i = \underset{h \in \mathcal{H}_i}{\operatorname{argmin}}\, err(h, S)$$

Q: How to use $\mathcal{H}_1, ..., \mathcal{H}_k$ to find a good $\hat{h}$ with low error? $\hat{h} = \operatorname{argmin}_{h \in \cup_i \mathcal{H}_i} err(h, S)$ is not a good idea since $\hat{h}_k$ may not be the best among $\{\hat{h}_1, ..., \hat{h}_k\}$.
Idea 1:Validation:

$$\hat{\mathcal{H}} = \{\hat{h}_1, ..., \hat{h}_k\}$$

$\hat{h} = \operatorname{argmin}_{h \in \hat{\mathcal{H}}} err(h, V)$, where V is a fresh validation sample set.
Analysis:

Claim 3. *With probability 1-$\frac{\delta}{2}$, $\forall i$*

$$err(\hat{h}_i, D) \leq err(h_i^*, D) + 2\sqrt{\frac{\ln \frac{k|\mathcal{H}_i|}{\delta}}{2m}}$$

(from standard ERM analysis + union bound over all i)

Claim 4. *With probability 1-δ,*

$$err(\hat{h}, D) \leq \min_i err(\hat{h}_i, D) + 2\sqrt{\frac{\ln \frac{4}{\delta}}{|V|}}$$

Claim 5. *From claim 3 and 4,we can show with probability 1-δ:*

$$err(\hat{h}, D) \leq \min_i (err(h_i^*, D) + 2\sqrt{\frac{\ln \frac{k|\mathcal{H}_i|}{\delta}}{2m}}) + 2\sqrt{\frac{\ln \frac{4}{\delta}}{|V|}}$$

where $|V| = \Theta(m)$

Claim 5 shows in this case $\hat{h}$ has the best bias-complexity tradeoff.

Idea 2: Structural risk minimization(penalized ERM)

$$\hat{i} = \underset{i \in \{1,...,k\}}{\operatorname{argmin}} \; err(\hat{h}_i, S) + \sqrt{\frac{\ln \frac{2k|\mathcal{H}_i|}{\delta}}{2m}})^{9}$$

Output $\hat{h} = \hat{h}_{\hat{i}}$

Example 4.

$$\mathcal{H}_1 \leq \mathcal{H}_2 \leq ... \leq \mathcal{H}_k$$

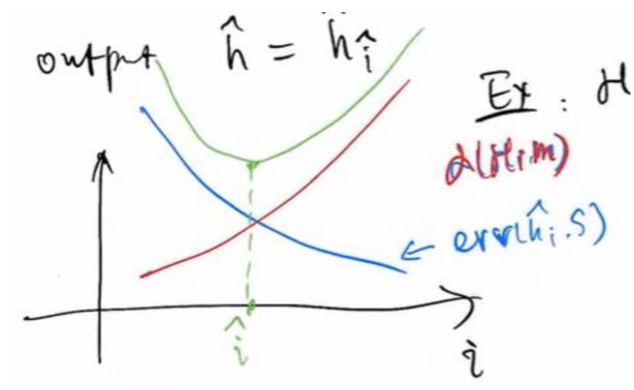

Figure 6: penalized ERM

In next class, we are going to show

$$err(\hat{h}, D) \leq \min_{i \in \{1,...,k\}} err(h_i^*, D) + 4\alpha(\mathcal{H}_i, m)$$

This proves our output also achieves near-optimal bias-complexity tradeoff.

[9] penalty for complexity term, define it as $\alpha(\mathcal{H}_i, m)$

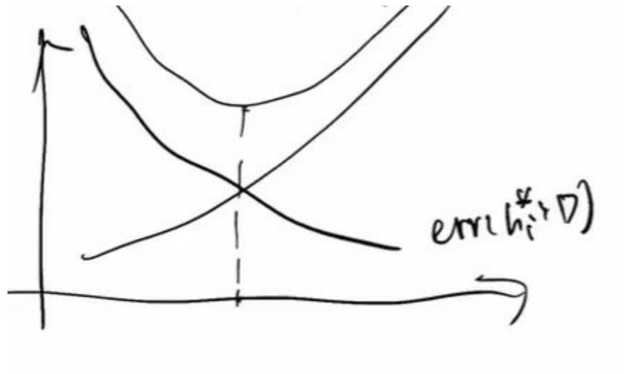

Figure 7: upper bound of error for penalized ERM