Model Selection

Mathematical Analysis of Machine Learning Algorithms
(Chapter 8)

Model Selection Problem

Model

A model is a learning algorithm $\mathcal{A}(\theta, \mathcal{S}_n)$ that maps the training data $\mathcal{S}_n$ to a prediction function $f \in \mathcal{F}(\theta) = \{f(w, x) : w \in \Omega(\theta)\} \subset \mathcal{F}$, indexed by a hyperparameter $\theta \in \Theta$. For simplicity, we take $\mathcal{F} = \cup \mathcal{F}(\theta)$.

Model Selection

The goal of model selection is to find the best model hyperparameter θ so that the corresponding learning algorithm $\mathcal{A}(\theta, \cdot)$ achieves a small test error.

We also let

$$\phi(f, Z) = L(f(X), Y) \quad \phi(w, Z) = L(f(w, X), Y)$$

$$\phi(f, \mathcal{D}) = \mathbb{E}_{Z \sim \mathcal{D}}\, \phi(f, Z), \quad \phi(f, \mathcal{S}_n) = \frac{1}{n} \sum_{Z \in \mathcal{S}_n} \phi(f, Z).$$

Definition of Model Selection

Definition 1 (Def 8.1)

Consider a loss function $\phi(f, z) : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$, and a model family $\{\mathcal{A}(\theta, \mathcal{S}_n) : \Theta \times \mathcal{Z}^n \to \mathcal{F}, n \geq 0\}$. Consider $N \geq n \geq 0$, and iid dataset $\mathcal{S}_n \subset \mathcal{S}_N \sim \mathcal{D}^N$. A model selection algorithm $\bar{A}$ maps $\mathcal{S}_N$ to $\hat{\theta} = \hat{\theta}(\mathcal{S}_N) \in \Theta$, and then train a model $\hat{f} = \mathcal{A}(\hat{\theta}(\mathcal{S}_N), \mathcal{S}_n) = \bar{A}(\mathcal{S}_N)$. It satisfies an $\epsilon_{n,N}(\cdot, \cdot)$ oracle inequality if there exists $\epsilon_{n,N}(\theta, \delta)$, such that for all $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $\mathcal{S}_N$:

$$\phi(\mathcal{A}(\hat{\theta}(\mathcal{S}_N), \mathcal{S}_n), \mathcal{D}) \leq \inf_{\theta \in \Theta} \left[\mathbb{E}_{\mathcal{S}_n} \phi(\mathcal{A}(\theta, \mathcal{S}_n), \mathcal{D}) + \epsilon_{n,N}(\theta, \delta) \right].$$

More generally, a learning algorithm $\bar{A} : \mathcal{S}_N \to \mathcal{F}$ is $\epsilon_{n,N}(\cdot, \cdot)$ adaptive to the model family $\{\mathcal{A}(\theta, \cdot) : \theta \in \Theta\}$ if there exists $\epsilon_{n,N}(\theta, \delta)$, such that for all $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $\mathcal{S}_N$:

$$\phi(\bar{A}(\mathcal{S}_N), \mathcal{D}) \leq \inf_{\theta \in \Theta} \left[\mathbb{E}_{\mathcal{S}_n} \phi(\mathcal{A}(\theta, \mathcal{S}_n), \mathcal{D}) + \epsilon_{n,N}(\theta, \delta) \right].$$

Model Selection Example: Hyperparameter Tuning

Consider ridge regression algorithm indexed by the regularization parameter $\lambda > 0$:

$$\hat{w}(\lambda) = \arg\min_{w \in \mathbb{R}^d} \left[\sum_{i=1}^{n} (w^\top X_i - Y_i)^2 + \lambda \|w\|_2^2 \right],$$

where $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are training data. For this problem, we have

$$\mathcal{F} = \{w^\top x : w \in \mathbb{R}^d, x \in \mathbb{R}^d\}.$$

The goal is to find λ so that the test error

$$\mathbb{E}_{(X,Y)}(Y - \hat{w}(\lambda)^\top X)^2$$

is as small as possible. The parameter λ is called hyperparameter.

Model Selection on Validation Set

Split a labeled data into training data of size n and test data of size m

- training data: $\mathcal{S}_n$
- validation data: $\bar{\mathcal{S}}_m$

Given model hyperprameter θ, we train a prediction function

$$\hat{f}_\theta = \mathcal{A}(\theta, \mathcal{S}_n) \in \mathcal{F}$$

based on training data $\mathcal{S}_n$.

We then select $\hat{\theta}$ based on validation data $\bar{\mathcal{S}}$ so that the test error

$$\mathbb{E}_{\mathcal{D}}\phi(\hat{f}_{\hat{\theta}}, Z)$$

is small.

Model Selection Algorithm

Let $\{q(\theta) \geq 0\}$ be a sequence of non-negative numbers that satisfies the inequality

$$\sum_{\theta=1}^{\infty} q(\theta) \leq 1. \tag{1}$$

Consider the following model selection algorithm that selects $\hat{\theta}$ to approximately minimize:

$$Q(\hat{\theta}, \mathcal{A}(\hat{\theta}, \mathcal{S}_n), \bar{\mathcal{S}}_m) \leq \inf_{\theta} Q(\theta, \mathcal{A}(\theta, \mathcal{S}_n), \bar{\mathcal{S}}_m) + \tilde{\epsilon}, \tag{2}$$

where

$$Q(\theta, f, \bar{\mathcal{S}}_m) = \phi(f, \bar{\mathcal{S}}_m) + r_m(q(\theta)).$$

Discrete Model Selection Result

Theorem 2 (Model Selction on Validation Data, Thm 8.2)

Assume $\sup_{Z,Z'}[\phi(f,Z) - \phi(f,Z')] \leq M$. *Consider* (2) *with*

$$r_m(q) = M\sqrt{\frac{\ln(1/q)}{2m}}.$$

Then with probability at least $1-\delta$ *over the random selection of* $\mathcal{S}_m$*:*

$$\phi(\mathcal{A}(\hat{\theta}, \mathcal{S}_n), \mathcal{D}) \leq \inf_{\theta} Q(\theta, \mathcal{A}(\theta, \mathcal{S}_n), \bar{\mathcal{S}}_m) + \tilde{\epsilon} + M\sqrt{\frac{\ln(1/\delta)}{2m}}.$$

This implies the following oracle inequality. With probability at least $1-\delta$ *over the random sampling of* $\bar{\mathcal{S}}_m$*:*

$$\phi(\mathcal{A}(\hat{\theta}, \mathcal{S}_n), \mathcal{D}) \leq \inf_{\theta}[\phi(\mathcal{A}(\theta, \mathcal{S}_n), \mathcal{D}) + r_m(q(\theta))] + \tilde{\epsilon} + M\sqrt{\frac{2\ln(2/\delta)}{m}},$$

where $q(\theta)$ *satisfies* (1)*.*

Proof of Theorem 2

For each model θ, let $\hat{f}_\theta = \mathcal{A}(\theta, \mathcal{S}_n)$. We obtain from the additive Chernoff bound that with probability at least $1 - q(\theta)\delta$:

$$\begin{aligned}\mathbb{E}_{Z\sim\mathcal{D}}\phi(\hat{f}_\theta, Z) \leq& \frac{1}{m}\sum_{Z\in\bar{\mathcal{S}}_m}\phi(\hat{f}_\theta, Z) + M\sqrt{\frac{\ln(1/(q(\theta)\delta))}{2m}} \\ \leq& \frac{1}{m}\sum_{Z\in\bar{\mathcal{S}}_m}\phi(\hat{f}_\theta, Z) + M\sqrt{\frac{\ln(1/q(\theta))}{2m}} + M\sqrt{\frac{\ln(1/\delta)}{2m}}.\end{aligned}$$

Taking the union bound over θ, we know that the above claim holds for all $\theta \geq 1$ with probability at least $1 - \delta$. This result, combined with the definition of $\hat{\theta}$ in (2), leads to the first desired bound.
Now by applying the Chernoff bound for an arbitrary θ that does not depend on $\bar{\mathcal{S}}_m$, we obtain with probability at least $1 - \delta/2$:

$$Q(\theta, \hat{f}_\theta, \bar{\mathcal{S}}_m) \leq \mathbb{E}_{Z\sim\mathcal{D}}\phi(\hat{f}_\theta, Z) + r_m(q(\theta)) + M\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

By combining this inequality with the first bound of the theorem, we obtain the second desired inequality.

Approximate ERM Learner

Consider a countable family of approximate ERM algorithms

$$\{\mathcal{A}(\theta, \cdot) : \theta = 1, 2, \ldots\},$$

each characterized by its model space $\mathcal{F}(\theta)$.

The approximate ERM algorithm $\mathcal{A}(\theta, \cdot)$ returns a function $\hat{f}_\theta \in \mathcal{F}(\theta)$ such that

$$\phi(\hat{f}, \mathcal{S}_n) \leq \inf_{f \in \mathcal{F}(\theta)} \phi(f, \mathcal{S}_n) + \epsilon', \tag{3}$$

where we use the notation of Definition 1.

Oracle Inequality for Approximate ERM Learner

Corollary 3 (Cor 8.3)

Consider approximate ERM Learner (3)*. Assume further that* $\sup_{Z,Z'}[\phi(f,Z) - \phi(f,Z')] \leq M$ *for all* f*, and we use* (2) *to select* $\hat{\theta}$*:*

$$r_m(q) = M\sqrt{\frac{\ln(1/q)}{2m}}.$$

Then the following result holds with probability at least $1-\delta$ *over random selection of* $\mathcal{S}_n$ *and* $\bar{\mathcal{S}}_m$*:*

$$\phi(\mathcal{A}(\hat{\theta},\mathcal{S}_n),\mathcal{D}) \leq \inf_{\theta}\left[\inf_{f\in\mathcal{F}(\theta)}\phi(f,\mathcal{D}) + 2R_n(\mathcal{G}(\theta),\mathcal{D}) + r_m(q(\theta))\right] + \tilde{\epsilon} + \epsilon' + M\sqrt{\frac{2\ln(4/\delta)}{n}} + M\sqrt{\frac{2\ln(4/\delta)}{m}},$$

where $R_n(\mathcal{G}(\theta),\mathcal{D})$ *is the Rademacher complexity of* $\mathcal{G}(\theta) = \{\phi(f,\cdot) : f \in \mathcal{F}(\theta)\}$ *and* $q(\theta)$ *satisfies* (1)*.*

Result used in the Proof of Corollary 3

Corollary 4 (Cor 6.21)

Assume that for some $M \geq 0$:

$$\sup_{w \in \Omega} \sup_{z, z'} \left[\phi(w, z) - \phi(w, z')\right] \leq M.$$

Then the approximate ERM method

$$\phi(\hat{w}, \mathcal{S}_n) \leq \min_{w \in \Omega} \phi(w, \mathcal{S}_n) + \epsilon'$$

satisfies the following oracle inequality. With probability at least $1 - \delta$:

$$\phi(\hat{w}, \mathcal{D}) \leq \inf_{w \in \Omega} \phi(w, \mathcal{D}) + \epsilon' + 2R_n(\mathcal{G}, \mathcal{D}) + 2M\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof of Corollary 3

Consider any model θ. We have from Theorem 2 that with probability $1 - \delta/2$,

$$\phi(\mathcal{A}(\hat{\theta}, \mathcal{S}_n), \mathcal{D}) \leq [\phi(\mathcal{A}(\theta, \mathcal{S}_n), \mathcal{D}) + r_m(q(\theta))] + \tilde{\epsilon} + M\sqrt{\frac{2\ln(4/\delta)}{m}}.$$

Moreover, from Corollary 4, we know that with probability at least $1 - \delta/2$:

$$\phi(\mathcal{A}(\theta, \mathcal{S}_n), \mathcal{D}) \leq \inf_{f \in \mathcal{F}(\theta)} \phi(f, \mathcal{D}) + \epsilon' + 2R_n(\mathcal{G}(\theta), \mathcal{D}) + 2M\sqrt{\frac{\ln(4/\delta)}{2n}}.$$

Taking the union bound, both inequalities hold with probability at least $1 - \delta$, which leads to the desired bound.

Example

Example 5 (Expl 8.4)

Consider a $\{0,1\}$ valued binary classification problem, with binary classifiers $\mathcal{F}(\theta) = \{f_\theta(w,x) \in \{0,1\} : w \in \Omega(\theta)\}$ of VC-dimension $d(\theta)$. The Rademacher complexity of $\mathcal{G}(\theta)$ is no larger than $(16\sqrt{d(\theta)})/\sqrt{n}$ (See Example 6.26). Take $q(\theta) = 1/(\theta+1)^2$. Then we have from Corollary 3 that

$$\mathbb{E}_{\mathcal{D}} \mathbb{1}(f_{\hat{\theta}}(\hat{w}, X) \neq Y) \leq \inf_{\theta, w \in \Omega(\theta)} \left[\mathbb{E}_{\mathcal{D}} \mathbb{1}(f_\theta(w, X) \neq Y) + \frac{32\sqrt{d(\theta)}}{\sqrt{n}} + \sqrt{\frac{\ln(\theta+1)}{m}} \right] + \tilde{\epsilon} + \epsilon' + \sqrt{\frac{2\ln(4/\delta)}{n}} + \sqrt{\frac{2\ln(4/\delta)}{m}}.$$

This result shows that the model selection algorithm of (2) can automatically balance the model accuracy $\mathbb{E}_{\mathcal{D}} \mathbb{1}(f_\theta(w, X) \neq Y)$ and model dimension $d(\theta)$. it can adaptively choose the optimal model θ, up to a penalty of $O(\sqrt{\ln(\theta+1)/n})$.

Model Selection on Training Data

If we have a training data dependent generalization bound, then we can obtain a model selection algorithm that minimize the generalization bound on the training data without training/validation split.

Consider the following model selection algorithm, which simultaneously finds the model hyperparameter $\hat{\theta}$ and model function $\hat{f} \in \mathcal{F}(\hat{\theta})$ on the training data $\mathcal{S}_n$:

$$Q(\hat{\theta}, \hat{f}, \mathcal{S}_n) \leq \inf_{\theta, f \in \mathcal{F}(\theta)} Q(\theta, f, \mathcal{S}_n) + \tilde{\epsilon}, \tag{4}$$

where for $f \in \mathcal{F}(\theta)$,

$$Q(\theta, f, \mathcal{S}_n) = \phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n),$$

where $\tilde{R}$ is an appropriately chosen sample dependent upper bound of the complexity for family $\mathcal{F}(\theta)$.

Theorem 6 (Uniform Convergence, Simplified from Thm 8.5)

Let $\{q(\theta) \geq 0\}$ *be a sequence of numbers that satisfy* (1)*. Assume that for each model* θ*, we have uniform convergence result as follows. With probability at least* $1 - \delta$*, for all* $f \in \mathcal{F}(\theta)$*,*

$$\phi(f, \mathcal{D}) \leq \phi(f, \mathcal{S}_n) + \hat{\epsilon}(\theta, f, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(c_0/\delta)}{n}},$$

for some constants $M(\theta) > 0$ *and* $c_0 \geq 1$*. If we choose*

$$\tilde{R}(\theta, f, \mathcal{S}_n) \geq \hat{\epsilon}(\theta, f, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(c_0/q(\theta))}{n}},$$

then with probability at least $1 - \delta$*, for all* θ *and* $f \in \mathcal{F}(\theta)$*:*

$$\phi(f, \mathcal{D}) \leq \phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(1/\delta)}{n}}.$$

Theorem 7 (Oracle Inequality, Simplified from Thm 8.5)

Under the assumptions of Theorem 6. If moreover, we have for all θ *and* $f \in \mathcal{F}(\theta)$*, the following concentration bound hold, with probability* $1-\delta$*:*

$$\phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n) \leq \mathbb{E}_{\mathcal{S}_n}\left[\phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n)\right] + \epsilon'(\theta, f, \delta).$$

Then we have the following oracle inequality for (4)*. With probability at least* $1-\delta$*:*

$$\phi(\hat{f}, \mathcal{D}) \leq \inf_{\theta, f \in \mathcal{F}(\theta)} \left[\phi(f, \mathcal{D}) + \mathbb{E}_{\mathcal{S}_n}\tilde{R}(\theta, f, \mathcal{S}_n) + \epsilon'(\theta, f, \delta/2)\right]$$
$$+ \tilde{\epsilon} + M(\theta)\sqrt{\frac{\ln(2/\delta)}{n}}.$$

Proof of Theorem 6

Taking union bound over θ, each with probability $1 - 0.5q(\theta)\delta$, we obtain that with probability at least $1 - \delta/2$, for all θ and $f \in \mathcal{F}(\theta)$,

$$\begin{aligned}
\phi(f, \mathcal{D}) \leq & \phi(f, \mathcal{S}_n) + \hat{\epsilon}(\theta, f, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(c_0/q(\theta))}{n} + \frac{\ln(2/\delta)}{n}} \\
\leq & \phi(f, \mathcal{S}_n) + \hat{\epsilon}(\theta, f, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(c_0/q(\theta))}{n}} + M(\theta)\sqrt{\frac{\ln(2/\delta)}{n}} \\
\leq & \phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(2/\delta)}{n}}.
\end{aligned}$$

The first inequality used the union bound over all $\mathcal{F}(\theta)$. The second inequality used Jensen's inequality. The third inequality used the assumption of $\tilde{R}$. This proves the desired uniform convergence result.

Proof of Theorem 7

Now since $\hat{f}$ is the solution of (4), it follows that for all θ and $f \in \mathcal{F}(\theta)$, with probability at least $1 - \delta/2$:

$$\begin{aligned}\phi(\hat{f}, \mathcal{D}) \leq & \phi(\hat{f}, \mathcal{S}_n) + \tilde{R}(\hat{\theta}, \hat{f}, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(2/\delta)}{n}} \\ \leq & \phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n) + M(\theta)\sqrt{\frac{\ln(2/\delta)}{n}} + \tilde{\epsilon}.\end{aligned}$$

In addition, with probability at least $1 - \delta/2$:

$$\phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n) \leq \mathbb{E}_{\mathcal{S}_n}\left[\phi(f, \mathcal{S}_n) + \tilde{R}(\theta, f, \mathcal{S}_n)\right] + \epsilon'(\theta, f, \delta/2).$$

Taking the union bound, and sum of the two inequalities, we obtain the desired oracle inequality.

Model Selection Using Rademacher Complexity

Theorem 8 (Thm 8.7)

Consider the model selection algorithm in (4)*, with*

$$\tilde{R}(\theta, f, \mathcal{S}_n) = \tilde{R}(\theta) \geq 2R_n(\mathcal{F}(\theta), \mathcal{D}) + M(\theta)\sqrt{\frac{\ln(1/q(\theta))}{2n}},$$

where $M(\theta) = \sup_{f,z,z'} \left|\phi(f,z) - \phi(f,z')\right|$*, and* $q(\theta)$ *satisfies* (1)*. Then with probability at least* $1-\delta$*, for all* θ *and* $f \in \mathcal{F}(\theta)$*:*

$$\phi(f, \mathcal{D}) \leq \phi(f, \mathcal{S}_n) + \tilde{R}(\theta) + M(\theta)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Moreover, we have oracle inequality: with probability of at least $1-\delta$*,*

$$\phi(\hat{f}, \mathcal{D}) \leq \inf_{\theta, f \in \mathcal{F}(\theta)} \left[\phi(f, \mathcal{D}) + \tilde{R}(\theta) + 2M(\theta)\sqrt{\frac{\ln(2/\delta)}{2n}}\right] + \tilde{\epsilon}.$$

Proof of Theorem 8 (I/II)

Using Rademacher complexity, we know for any θ, with probability $1-\delta$, the following uniform convergence result holds for all $f \in \mathcal{F}(\theta)$:

$$\phi(f, \mathcal{D}) \leq \phi(f, \mathcal{S}_n) + 2R_n(\mathcal{F}(\theta), \mathcal{D}) + M(\theta)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

The choice of $\tilde{R}$ satisfies the condition of Theorem 6. It implies the desired uniform convergence result.

Proof of Theorem 8 (II/II)

Given fixed θ and $f \in \mathcal{F}(\theta)$, we know that

$$\left|[\phi(f, \mathcal{S}_n) + \tilde{R}(\theta)] - [\phi(f, \mathcal{S}_n') + \tilde{R}(\theta)]\right| \leq M(\theta)$$

when $\mathcal{S}_n$ and $\mathcal{S}_n'$ differ by one element. From McDiarmid's inequality, we know that with probability at least $1 - \delta$,

$$\phi(f, \mathcal{S}_n) + \tilde{R}(\theta) \leq \phi(f, \mathcal{D}) + \tilde{R}(\theta) + M(\theta)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

It follows that we can take

$$\epsilon'(\theta, f, \delta) = M(\theta)\sqrt{\frac{\ln(1/\delta)}{2n}}$$

in Theorem 7, and obtain the desired oracle inequality.

Example

Example 9

Consider the same problem considered in Example 5. We can take $M(\theta) = 1$ and $h = 0$ in Theorem 8. It implies that the model selection method (4) with

$$\tilde{R}(\theta, f, \mathcal{S}_n) = \frac{32\sqrt{d(\theta)}}{\sqrt{n}} + \sqrt{\frac{\ln(\theta + 1)}{n}}$$

satisfies the following oracle inequality. With probability $1 - \delta$:

$$\mathbb{E}_{\mathcal{D}} \mathbb{1}(f_{\hat{\theta}}(\hat{\theta}, X) \neq Y) \leq \inf_{\theta, w \in \Omega_\theta} \left[\mathbb{E}_{\mathcal{D}} \mathbb{1}(f_\theta(w, X) \neq Y) + \frac{32\sqrt{d(\theta)}}{\sqrt{n}} + \sqrt{\frac{\ln(\theta + 1)}{n}} \right] + \sqrt{\frac{2\ln(2/\delta)}{n}}.$$

The result is comparable to that of Example 5.

Summary (Chapter 8)

- Model Selection Problem
- Model Selection on Validation Data
- Model Selection on Training Data using Sample Dependent Bound