CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 19: November 9

Lecturer: Csaba Szepesvári *Scribes: Aniket Sharma*

Note: *LaTeX template courtesy of UC Berkeley EECS dept. (link to directory)*
Disclaimer: *These notes have* ***not*** *been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Lecture 19 video

19.1 Outline

- Model Selection Problem
- Model Selection using Validation Data
- Model Selection using Training Data
- Bayesian Model Selection and Averaging

19.2 Model Selection Problem

We have a set of function classes $\mathcal{G}_i \in \mathbb{R}^z, i \in \mathbb{N}$, and

$$g_n^{(i)} = argmin_{g \in \mathcal{G}_i} P_n g$$
$$Pg_n^{(i)} \leq inf_{g \in \mathcal{G}_i} Pg + penalty_i(n, \delta) \qquad \text{, wp } 1-\delta$$
$$Pg_n = min_i Pg_n^{(i)}$$

We want to find the class such that the empirical performance is the best

$$g_n \in argmin_{g \in \cup_i \mathcal{G}_i} Png$$

Note: If $VC(\mathcal{G}_i) = d_i$, then $penalty_i(n) = \sqrt{\frac{d_i ln(\frac{1}{\delta})}{n}}$.

19.3 Model Selection using Validation Data

We have $z_{1:n}, z'_{1:m} \sim P^{\otimes(n+m)}$, where $z_{1:n}$ is the training data and $z'_{1:m}$ is the validation data.

$$P'_m = \frac{1}{m} \Sigma_{i=1}^m \delta_{z'_i}$$
$$I = argmin_{i \in \mathbb{N}} P'_m g_n^{(i)} + \sqrt{ln\left(\frac{1}{q_i}\right)}$$

Here, $\sqrt{ln\left(\frac{1}{q_i}\right)}$ is the "complexity" penalty. Also, $\Sigma q_i \leq 1, q_i \geq 0$. A typical choice will be $q_i = \frac{1}{i(i+1)}$ or $q_i = \frac{1}{(i+1)^2}$.
We want to consider less complex classes first (Occam's razor) like $d_1 \leq d_2 \leq \ldots$ for VC classes.

Theorem 19.1. *Let* $sup_{z,z'} sup_{g\in\cup_i \mathcal{G}_i} g(z) - g(z') \leq M$*, then*

1. *wp* $1-\delta$,

$$Pg_n^I \leq inf_{i\in\mathbb{N}} P'_m g_n^i + \sqrt{ln\left(\frac{1}{q_i}\right)} + M\sqrt{\frac{ln\left(\frac{1}{\delta}\right)}{2m}}$$

2. *wp* $1-\delta$,

$$Pg_n^I \leq inf_{i\in\mathbb{N}} P g_n^i + \sqrt{ln\left(\frac{1}{q_i}\right)} + M\sqrt{\frac{ln\left(\frac{2}{\delta}\right)}{2m}}$$

19.4 Model Selection using Training Data

An alternative approach would be to use the training data for model selection instead of splitting.

$$(I, \mathcal{G}) := argmin\{P_n g + R_i(g, z_{1:n}) : i \in \mathbb{N}, g \in \mathcal{G}_i\}$$

Here, $R_i(g, z_{1:n})$ is the data-dependent penalty.

Theorem 19.2. $\Sigma q_i \leq 1, q_i \geq 0, \forall\delta \in (0,1)$,

$$\alpha P_g \leq P_n g + \varepsilon_i(g, z_{1:n}) + \left(\frac{ln\left(\frac{c_0}{\delta}\right)}{\lambda n}\right)^\beta$$

for some $\alpha, \beta, \lambda > 0, c_0 \geq 1$,

$$R_i(g, z_{1:n}) \geq \varepsilon_i(g, z_{1:n}) + 2^{max(0,\beta-1)}\left(\frac{ln\left(\frac{c_0}{q_i}\right)}{\lambda n}\right)^\beta$$

Part 1: $\forall\delta \in (0,1)$ *wp* $1-\delta$*:* $\forall i \in \mathbb{N}, g \in \mathcal{G}$,

$$\alpha P_g \leq P_n g + R_i(g, z_{1:n}) + 2^{max(0,\beta-1)}\left(\frac{ln\left(\frac{c_0}{q_i}\right)}{\lambda n}\right)^\beta$$

Part 2: $\forall\delta \in (0,1), \forall i \in \mathbb{N}, g \in \mathcal{G}$,

$$P_n g + R_i(g, z_{1:n}) \leq \mathbb{E}[\alpha' P_n g + \alpha'' R_i(g, z_{1:n})] + \varepsilon'_i(g, \delta)$$

then wp $1-\delta$,

$$\alpha PG \leq inf_{i\in\mathbb{N}, g\in\mathcal{G}_i}\left[\alpha' Pg + \alpha''\mathbb{E}[R_i(g, z_{1:n})] + \varepsilon'\left(g, \frac{\delta}{2}\right)\right] + 2^{max(0,\beta-1)}\left(\frac{ln\left(\frac{c_0}{q_i}\right)}{\lambda n}\right)^\beta$$

19.4.1 Concentration of Empirical Rademacher Complexity

Theorem 19.3.

$$R_i(g, z_{1:n}) \geq 2R_n(\mathcal{G}_i, P) + M_i\sqrt{\frac{ln\left(\frac{1}{q_i}\right)}{2n}}$$

where, $M_i = sup_{g\in\mathcal{G}_i} sup_{z,z'\in\mathcal{Z}} g(z) - g(z')$. *Then,*

1. *wp* $1-\delta$*:* $i \in \mathbb{N}, g \in \mathcal{G}_i$,

$$Pg \leq P_n g + R_i(g, z_{1:n}) + M_i\sqrt{\frac{ln\left(\frac{1}{\delta}\right)}{2n}}$$

2. *wp* $1-\delta$,

$$PG \leq inf_{i\in\mathbb{N}, g\in\mathcal{G}_i} Pg + R_i(g, z_{1:n}) + 2M_i\sqrt{\frac{ln\left(\frac{2}{\delta}\right)}{2n}}$$

Theorem 19.4. $M \geq sup_g sup_{z,z'} g(z) - g(z')$*, then wp* $1-\delta$,

$$R_n(\mathcal{G}, P) \leq R(\mathcal{G}, z_{1:n}) + M\sqrt{\frac{ln\left(\frac{1}{\delta}\right)}{2n}}$$

Also wp $1-\delta$,

$$R_n(\mathcal{G}, P) \geq R(\mathcal{G}, z_{1:n}) - M\sqrt{\frac{ln\left(\frac{1}{\delta}\right)}{2n}}$$

Here, $R(\mathcal{G}, z_{1:n})$ is the empirical Rademacher complexity.

Corollary 19.5. *wp* $1-\delta$*:* $\forall g \in \mathcal{G}$,

$$Pg \leq P_n g + 2R(\mathcal{G}, z_{1:n}) + 3M\sqrt{\frac{ln\left(\frac{2}{\delta}\right)}{2n}}$$

Theorem 19.6.

$$R_i(z_{1:n}) \geq R(\mathcal{G}_i, z_{1:n} + 3M_i\sqrt{\frac{ln\left(\frac{2}{q_i}\right)}{2n}}$$

Then,

1. *wp* $1-\delta$*:* $\forall i \in \mathbb{N}, g \in \mathcal{G}_i$,

$$Pg \leq P_n g + R_i(z_{1:n}) + 3M_i\sqrt{\frac{ln\left(\frac{1}{\delta}\right)}{2n}}$$

2. *wp* $1-\delta$,

$$PG \leq inf_{i\in\mathbb{N}, g\in\mathcal{G}_i} Pg + \mathbb{E}[R_i(z_{1:n})] + 4M_i\sqrt{\frac{ln\left(\frac{2}{\delta}\right)}{2n}}$$

19.5 Bayesian Model Selection and Averaging

Consider the Gibb's algorithm,

$$g \sim exp(-\beta n P_n g)\pi_0(dg)$$

Here, $g \in \mathcal{G}$ and $\pi_0(dg)$ is the prior.
Take $\Sigma q_i = 1$,

$$(I, \mathcal{G}) \sim P_i \pi_i(dg) exp(-\beta n P_n g)$$

Here, $\pi_i(dg)$ is the prior for class $\mathcal{G}_i$.
Now, we can use the Bayesian formula for Gibbs model selection and select a model randomly but in practice model averaging often leads to superior performance.
For $f \in \mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$,

$$\tilde{P_n}(df, i) = P_i \pi_i(dg) exp(-\beta n P_n l(f))$$

Here, $\tilde{P_n}(df, i)$ is the posterior.
Then we can make the predictions using,

$$\Sigma_i \int f(x) \tilde{P_n}(df, i)$$

Claim: Averaging $>>>$ Any Model Selection