4 October 2021

CS4912(F) - Computational Learning Theory. Chapter 4 Model Selection.

1. Motivation

(1) So far we studied results that let us use sample-based estimates, in particular, empirical error $\hat{R}_S(h)$, and express properties about what we are ultimately interested in, namely, generalisation error $R(h)$. The next natural question is how to exploit these results for designing an algorithm and analysing it.

(2) A reasonable-looking answer is to pick a hypothesis set $\mathcal{H}$ and to pick $h \in \mathcal{H}$ for a given sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ that minimises the empirical risk $\hat{R}_S(h)$. This approach is called empirical risk minimisation (ERM). At first look - you may feel that this is as good as we can get.

(3) But it turns out that we can do better than ERM, in a sense. The idea is to impose a structure on $\mathcal{H}$ by assuming a family of hypothesis sets $\{\mathcal{H}_k\}_{k \in \mathbb{N}}$ with $\mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$ and to perform search at two levels, one at the level of a hypothesis set $\mathcal{H}_k$ and the other at the level of hypothesis $h \in \mathcal{H}_k$. The search for a good hypothesis set $\mathcal{H}_k$ is called model selection, the topic of this chapter. ($\mathcal{H}_k$ is called a model). By the way, this two-level search can be done one after the other (as in n-folds cross validation) or simultaneously (as in structural risk minimisation).

(4) What should we do in order to do model selection well? We should define a good success measure of model (i.e., hypothesis set $\mathcal{H}_k$) which also permits an efficient approximate optimiser. In this chapter, we will study such measures and derive idealised learning algorithms that (for instance,) search for a good model $\mathcal{H}_k$ using those measures and and for a good hypothesis in the model $\mathcal{H}_k$.

(5) By the way, I used "idealised" above to mean that derived algorithms might not be implementable because, for instance, they require search over infinitely many entities and their search objectives cannot be computed. However, these algorithms serve as a basis for real implementable algorithms which typically perform further approximation. As you expect, idealised algorithms are easier to analyse mathematically than real implementable algorithms

2 Analysis of error.

(1) Let $D \in \Pr(\mathcal{X} \times \mathcal{Y})$ with $\mathcal{Y} = \{-1, +1\}$.

Recall that the best error we can achieve is

$$R^* = \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} [\min(\mathbb{P}[y=1|x], \mathbb{P}[y=-1|x])]$$

which is called **Bayes error**.

(2) For a hypothesis set $\mathcal{H} \subseteq [\mathcal{X} \to \mathcal{Y}]$ and $h \in \mathcal{H}$, the **excess error** of h is $R(h) - R^*$.

It can be decomposed into two parts as follows:

$$R(h) - R^* = \underbrace{R(h) - \inf_{h' \in \mathcal{H}} R(h')}_{\text{called estimation error.}} + \underbrace{\inf_{h' \in \mathcal{H}} R(h') - R^*}_{\text{called approximation error.}}$$

(3) Note that if $\mathcal{H}$ is rich, the approximation error is small but the estimation error is large. The opposite situation arises when $\mathcal{H}$ is too simple.

(4) The structural risk minimization (SRM) uses a success measure on hypo. sets $\mathcal{H}_k$ (or models) that balances est error and approx. error., as we will see shortly.

3. Empirical risk minimization

(1) Before studying algorithms with model selection, let's look at the most basic learning algorithm, empirical risk minimization.

(2) ERM algorithm for a hypo. set $\mathcal{H}$.

Input $S \in (\mathcal{X} \times \mathcal{Y})^m$

Output $h_S^{ERM} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_S(\mathcal{H})$.

→ we assume that the minimum exists.

(3) This algorithm has the following guarantee on estimation error.

Thm. $\mathbb{P}_{S \sim D^m} \left[R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right] \leq 2 \exp\left(-2m \left(\frac{\varepsilon}{2} - R_m(\mathcal{H}) \right)^2 \right)$

estimation error.

Equivalently, for all $\delta > 0$,

$$\mathbb{P}_{S \sim D^m} \left[R(h_S^{ERM}) \leq \inf_{h \in \mathcal{H}} R(h) + 2R_m(\mathcal{H}) + 2\sqrt{\frac{\log 2/\delta}{2m}} \right] \geq 1 - \delta.$$

Why? Because of Prop 4.1 and the Rademacher-complexity bound.

Prop 4.1

$$\mathbb{P}\left[R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right] \leq \mathbb{P}\left[\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| > \frac{\varepsilon}{2} \right]$$

Prop 4.1

Proof of thm.

$$\mathbb{P}\left[R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right]$$

$$\leq \mathbb{P}\left[\exists h \in \mathcal{H} .\ R(h) > \hat{R}_S(h) + \frac{\varepsilon}{2} \text{ or } R(h) < \hat{R}_S(h) - \frac{\varepsilon}{2} \right].$$

union-bound

$$\leq \mathbb{P}\left[\exists h \in \mathcal{H} .\ R(h) > \hat{R}_S(h) + \frac{\varepsilon}{2} \right] + \mathbb{P}\left[\exists h \in \mathcal{H} .\ R(h) < \hat{R}_S(h) - \frac{\varepsilon}{2} \right].$$

$$\leq \delta + \delta'$$

By Thm 3.5 and its dual lower-bound version. (which we didn't work out)

for all $\delta', \delta \geq \exp\left(-2m\left(\frac{\varepsilon}{2} - \mathcal{R}_m(\mathcal{H})\right)^2\right)$

□.

Proof of Prop 4.1

Pick $\varepsilon > 0$. Then, $\exists h_\varepsilon \in \mathcal{H}$ s.t. $R(h_\varepsilon) \leq \inf_{h \in \mathcal{H}} R(h) + \varepsilon$.

$$R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) = R(h_S^{ERM}) - R(h_\varepsilon) + R(h_\varepsilon) - \inf_h R(h).$$

$$\leq R(h_S^{ERM}) - R(h_\varepsilon) + \varepsilon.$$

here we use the fact that h_S^{ERM} is a minimizer.

$$\leq R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_\varepsilon) - R(h_\varepsilon) + \varepsilon.$$

$$\left(\because \hat{R}_S(h_S^{ERM}) \leq \hat{R}_S(h) \text{ for all } h \in \mathcal{H}\right)$$

$$\leq 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| + \varepsilon.$$

Since ε is arbitrary, $R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) \leq 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)|$.

Thus, $\mathbb{P}\left[R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right] \leq \mathbb{P}\left[2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| > \varepsilon \right]$ Q.

(4) ERM performs poorly and h_S^{ERM} is difficult to compute.

- Complex $\mathcal{H}$ --- large estimation error indicated by the bound in the thm that we just proved.
- Simple $\mathcal{H}$ --- large approximation error.

(5) The algorithms that we will look at next overcome this poor performance issue using model selection.

4. Structural risk minimization (SRM)

(1) We assume a countable family of hypo. sets $\{\mathcal{H}_k\}_{k \in \mathbb{N}}$ s.t.

$$\mathcal{H}_k \subseteq \mathcal{H}_{k+1} \quad \text{for all } k \in \mathbb{N}.$$

The SRM solves the following opt. pb:

$$h_S^{SRM} = \underset{k \geq 1,\, h \in \mathcal{H}_k}{\arg\min} \; F_k(h)$$

where

$$F_k(h) = \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}.$$

(2) Note the term $\mathcal{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}$. If we didn't have this term, h_S^{SRM} would be the same as h_S^{ERM} over the hypo. set $\mathcal{H} = \bigcup_k \mathcal{H}_k$.

(3) Where does this term come from? It come from the generalisation bound: for all $D \in \Pr(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}$, $\delta > 0$,

$$\mathbb{P}\left[\forall k \in \mathbb{N}.\ \forall h \in \mathcal{H}_k.\quad R(h) \leq \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}}\right] \geq 1 - \delta.$$

which we will show soon.

(4) By using $\mathcal{H}_k$ with large k, we can make $\hat{R}_S(h)$ small by choosing an appropriate $h \in \mathcal{H}_k$. But then we have to pay the price for this in the remaining term $R_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}$, which will become large. $\hat{R}_S(h)$ is related to approximation error, and $R_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}$ estimation error. Thus, $F_k(h) = \hat{R}_S(h) + R_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}$. From the generalization bound considers both types of errors and optimizing it lets us find a good balancing point between these two errors.

(5) For $h \in \mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$, let $k(h) = \min\{k \in \mathbb{N} \mid h \in \mathcal{H}_k\}$.

Lemma. $\forall D \in \mathcal{P}(X \times Y)$. $\forall m \in \mathbb{N}$. $\forall \delta > 0$.

$$\mathbb{P}_{S \sim D^m}\left[\forall k \, \forall h \in \mathcal{H}_k : R(h) \le \hat{R}_S(h) + R_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}}\right] \ge 1 - \delta.$$

Proof.

$$\mathbb{P}_{S \sim D^m}\left[\exists k \, \exists h \in \mathcal{H}_k : R(h) > \hat{R}_S(h) + R_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}}\right]$$

(Union bound)

$$\le \sum_{k=1}^{\infty} \mathbb{P}\left[\exists h \in \mathcal{H}_k : R(h) > \hat{R}_S(h) + R_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}}\right]$$

$$\le \sum_k \exp\left(-2m\left(\sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}}\right)^2\right)$$

(Thm 3.5 & $= \sqrt{\frac{\log 1/\delta'}{2m}}$)

$$\le \sum_k \exp\left(-2m\left(\frac{\log k}{m} + \frac{\log 2/\delta}{2m}\right)\right)$$

$$= \sum_k \frac{1}{k^2}(\delta/2) = \frac{\pi^2}{6}(\delta/2) \le \delta. \qquad \square$$

Thm 4.2 $\forall D \in P_r(\mathcal{X}\times\mathcal{Y})\ \forall m\in\mathbb{N}\ \forall \delta>0$

$$\mathbb{P}_{S\sim D^m}\left[R(h_S^{SRM}) \le \inf_{h\in\mathcal{H}\ (=\bigcup_{k\in\mathbb{N}}\mathcal{H}_k)} \left(R(h) + 2R_m(\mathcal{H}_{k(h)}) + 2\sqrt{\frac{\log k}{m}} \right) + \sqrt{\frac{2\log\frac{4}{\delta}}{m}} \right] \ge 1-\delta.$$

smaller. (often much smaller) than $R_m(\mathcal{H})$

※ ERM bound for $\mathcal{H}=\bigcup_k \mathcal{H}_k$. from Prob 4.1. & thm there

$$\mathbb{P}_{S\sim D^m}\left[R(h_S^{ERM}) \le \inf_{h\in\mathcal{H}} \left(R(h) + 2R_m(\mathcal{H}) \right) + \sqrt{\frac{2\log\frac{2}{\delta}}{m}} \right] \ge 1-\delta.$$

Proof of thm 4.2.

Lemma in the previous page. with $\sqrt{\frac{\log 2/\delta}{2m}} = \frac{\varepsilon}{2}$.

Let $\varepsilon = \sqrt{(2\log 4/\delta)/m}$. We have to show that.

$$\mathbb{P}\left[R(h_S^{SRM}) - \inf_{h\in\mathcal{H}} \left(R(h) + 2R_m(\mathcal{H}_{k(h)}) + 2\sqrt{\frac{\log k}{m}} \right) > \varepsilon \right] \le \delta.$$

This holds (by dominated convergence) if for all $h\in\mathcal{H}$,

$$\mathbb{P}\left[R(h_S^{SRM}) - \left(R(h) + 2R_m(\mathcal{H}_{k(h)}) + 2\sqrt{\frac{\log k}{m}} \right) > \varepsilon \right] \le \delta.$$

we will show this sufficient condition. Let $h_S = h_S^{SRM}$ and $k_S = k(h_S)$

$$\mathbb{P}\left[R(h_S) - \left(R(h) + 2R_m(\mathcal{H}_{k(h)}) + 2\sqrt{\frac{\log k}{m}} \right) > \varepsilon \right]$$

$$= \mathbb{P}\left[\left(R(h_S) - F_{k_S}(h_S) \right) + \left(F_{k_S}(h_S) - R(h) - 2R_m(\mathcal{H}_{k(h)}) - 2\sqrt{\frac{\log k}{m}} \right) > \varepsilon \right]$$

$$\le \mathbb{P}\left[R(h_S) - F_{k_S}(h_S) > \frac{\varepsilon}{2} \right] + \mathbb{P}\left[\text{"} > \frac{\varepsilon}{2} \right]$$

$$\le \mathbb{P}\left[R(h_S) - F_{k_S}(h_S) > \frac{\varepsilon}{2} \right] + \mathbb{P}\left[F_{k(h)}(h) - R(h) - 2R_m(\mathcal{H}_{k(h)}) - 2\sqrt{\frac{\log k}{m}} > \frac{\varepsilon}{2} \right]$$

$$\le 2\exp\left(-2m\cdot\left(\frac{\varepsilon}{2}\right)^2 \right) + \mathbb{P}\left[\hat{R}_S(h) - R(h) - R_m(\mathcal{H}_{k(h)}) - \sqrt{\frac{\log k}{m}} > \frac{\varepsilon}{2} \right]$$

$$\le 2\exp\left(-2m\left(\frac{\varepsilon}{2}\right)^2 \right) + 2\exp\left(-2m\left(\frac{\varepsilon}{2}\right)^2 \right) = 4\exp\left(-\frac{m\varepsilon^2}{2}\right) = \delta.$$

↑ Dual version of the lemma in the previous page, i.e.,

$\mathbb{P}\left[\forall k\ \forall h\in\mathcal{H}_k.\ R(h) \ge \hat{R}_S(h) - R_m(\mathcal{H}_k) - \sqrt{\frac{\log k}{m}} - \sqrt{\frac{\log 2/\delta}{2m}} \right] \ge 1-\delta.$ □

(b) Drawback of SRM: ① Difficult computationally.
- $R_m(\mathcal{H})$ is hard to compute.
- $\widehat{R}_S(h)$ is not easy to optimize.

② Works only when the family $\{\mathcal{H}_k\}_k$ is countable and increasing.

5. Cross Validation (CV)

Here we mean

(1) Cross validation used for model selection. It avoids the computation of $R_m(\mathcal{H})$, which is hard or impossible.

(2) Idea: Use two independent samples S_1 and S_2, the former for finding an optimal solution for each hyp. set $\mathcal{H}_k$, and the latter for estimating the generalisation errors of those solutions. Then, pick one based on the estimated errors.

(3) Algorithm description.

- We assume a family $\{\mathcal{H}_k\}_{k\in\mathbb{N}}$ s.t $\mathcal{H}_k \subseteq \mathcal{H}_{k+1}$ for all k.
- Input : $S \in (\mathcal{X}\times\mathcal{Y})^m$.

Divide S into a sample S_1 of size $(1-\alpha)m$ and the sample S_2 of the remaining αm pairs.

(α: fixed number (hyper-parameter).)

(S_2: called validation set.)

$$k^* = \operatorname{argmin}_{k\geq 1} \widehat{R}_{S_2}\left(h^{ERM}_{S_1,k}\right),$$

($\widehat{R}_{S_2}$: note that we use S_2, not S_1.)

$$h^{CV}_S = h^{ERM}_{S_1,k^*}$$

where $h^{ERM}_{S_1,k}$ is an ERM solution wrt. $\mathcal{H}_k$ and S_1.

Return h^{CV}_S.

(4) Note the simplicity of CV. For instance, we don't need to compute $R_m(\mathcal{H}_k)$.

(5) How well does CV work? The next result answers this question.

Thm 4.4 $\forall D \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})\ \forall m \in \mathbb{N}\ \forall \delta > 0$

$$\mathbb{P}^{\otimes m}\left[R(\hat{h}^{CV}_S) \le \inf_K \left(R(\hat{h}^{ERM}_{S_1,K}) + 2\sqrt{\frac{\log \mathrm{card}(\{\hat{h}^{\alpha}_S : K\})}{\alpha m}} \right) + 2\sqrt{\frac{\log \frac{4}{\delta}}{\alpha m}}\right] \ge 1-\delta.$$

$|S_2| = \alpha m$; note αm, not $(1-\alpha)m$ here.

where S_1 is a subsequence of S of size $(1-\alpha)m$ that is computed by CV.

How to prove thm 4.4? We use Prop. 4.3 below:

Prop 4.3 $\forall D \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})\ \forall m \in \mathbb{N}\ \forall \alpha > 0\ \forall \varepsilon > 0$

$$\mathbb{P}^{\otimes m}\left[\sup_{k\ge 1}\left(\left| R(\hat{h}^{ERM}_{S_1,k}) - \hat{R}_{S_2}(\hat{h}^{ERM}_{S_1,k}) \right| - \sqrt{\frac{\log k}{\alpha m}} \right) > \varepsilon \right] \le 4\exp(-2\alpha m \varepsilon^2).$$

where does it come from?

$4 = 2 \times 2$ ~ union-bound over all k and the term $\sqrt{\frac{\log k}{\alpha m}}$

Hoeffding for $|-|$.

Proof. union bound.

$$\mathbb{P}_S\Big[\sup_k \Big|R(h^{ERM}_{S_1,k}) - \hat{R}_{S_2}(h^{ERM}_{S_1,k})\Big| - \sqrt{\frac{\log k}{dm}} > \varepsilon\Big]$$

$$\le \sum_{k=1}^{\infty} \mathbb{P}_S\Big[\Big|R(h^{ERM}_{S_1,k}) - \hat{R}_{S_2}(h^{ERM}_{S_1,k})\Big| > \varepsilon + \sqrt{\frac{\log k}{dm}}\Big]$$

$$= \sum_{k=1}^{\infty} \mathbb{E}_{S_1}\Big[\mathbb{P}_{S_2}\Big[\quad \text{“} \quad \Big| S_1\Big]\Big]$$

$$\le \sum_{k=1}^{\infty} \mathbb{E}_{S_1}\Big[2\exp\Big(-\frac{2\big(dm\big(\varepsilon + \sqrt{\frac{\log k}{dm}}\big)\big)^2}{dm}\Big)\Big]$$

Hoeffding.

$$\le \sum_{k=1}^{\infty} 2\exp\big(-2dm\varepsilon^2 - 2\log k\big) = \sum_{k=1}^{\infty} \frac{2}{k^2}\exp(-2dm\varepsilon^2)$$

$$= \frac{2\times\pi^2}{6}\exp(-2dm\varepsilon^2) \le 4\exp(-2dm\varepsilon^2)$$

□.

Let's rewrite the conclusion of Prop 4.3 in terms of δ, instead of ε. We set $\delta = 4\exp(-2dm\varepsilon^2)$.

Then, $\varepsilon = \sqrt{\frac{\log 4/\delta}{2dm}}$. Thus, Prop 4.3 implies:

$$\mathbb{P}_S\Big[\sup_{k\ge 1}\Big|R(h^{ERM}_{S_1,k}) - \hat{R}_{S_2}(h^{ERM}_{S_1,k})\Big| - \sqrt{\frac{\log k}{dm}} \le \sqrt{\frac{\log 4/\delta}{2dm}}\Big] \ge 1-\delta.$$

We use this bound in our proof of Thm 4.4.

Proof of Thm 4.4 If $\sup_{k\ge 1}\Big|R(h^{ERM}_{S_1,k}) - \hat{R}_{S_2}(h^{ERM}_{S_1,k})\Big| - \sqrt{\frac{\log k}{dm}} \le \sqrt{\frac{\log 4/\delta}{2dm}}$,

then for all $k\ge 1$,

$$R(h^{CV}_S) \le \hat{R}_{S_2}(h^{CV}_S) + \sqrt{\frac{\log k(h^{CV}_S)}{dm}} + \sqrt{\frac{\log 4/\delta}{2dm}} \le \hat{R}_{S_2}(h^{ERM}_{S_1,k}) + \sqrt{\frac{\log k(h^{CV}_S)}{dm}} + \sqrt{\frac{\log 4/\delta}{2dm}}$$

$$\le R(h^{ERM}_{S_1,k}) + \sqrt{\frac{\log k}{dm}} + \sqrt{\frac{\log k(h^{CV}_S)}{dm}} + 2\sqrt{\frac{\log 4/\delta}{2dm}}$$

$$\le R(h^{ERM}_{S_1,k}) + 2\sqrt{\frac{\log(\max(k, k(h^{CV}_S)))}{dm}} + 2\sqrt{\frac{\log 4/\delta}{2dm}}$$

Thus, the thm follows from the above & version of Prop 4.3

□.

(b) Any problem with CV? We waste examples in S. If α is big, S_1 has a small # of examples. Thus, our ERM solutions might not be that good. On the other hand, if α is small, $|S_2|$ is small, and the generalisation bound that we proved is loose.

Is there a way to use all examples during ERM while doing some form of cross validation? N-fold cross validation is such a method.

6. N-fold cross validation.

(1) Idea: ① Use S twice, first to find a good model or a hypothesis set $\mathcal{H}_k$, and then to find a good hypothesis h in $\mathcal{H}_k$. ② For the first, estimate the quality of generalisation of each $\mathcal{H}_k$ using multiple splits of S based on n folds.

(2) Algorithm.

① Split S into n folds. $S_1, S_2, \cdots, S_n$.
($|S| = \sum_{i=1}^{n} m_i$ where $m_i = |S_i|$).

② $\hat{R}_{CV}(\mathcal{H}_k) = \frac{1}{n} \sum_{i=1}^{n} \hat{R}_{S_i}\left(h^{ERM}_{S_{\setminus i}, k}\right)$.

where $S_{\setminus i}$ means S without the i-th fold S_i

③ Pick k that maximises $\hat{R}_{CV}(\mathcal{H}_k)$.

④ Compute $h^{ERM}_{S,k}$, and return it.

model selection based on generalisation

hypo. selection in a selected model.

(3) Note that all the examples in S are used to compute $h^{ERM}_{S,k}$.

(4) When $m_1 = \cdots = m_n$,

$\hat{R}_{CV}(\mathcal{H})$ is an unbiased estimate of

$$\mathbb{E}_{S \sim \mathcal{D}^{m-m_1}}\left[R(h^{ERM}_{S,\mathcal{H}}) \right].$$

which measures the quality of generalization of the ERM algorithm over $\mathcal{H}$.

7 Regularization and Convex Surrogate Loss.

(1) Besides the use of $R_m(\mathcal{H}_k)$ in the objective, there are two other issues in SRM. First, it assumes a countable family of hypo. sets $\{\mathcal{H}_k\}_{k \in \mathbb{N}}$, which often limits the expressiveness of $\mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$. If possible, we want the family to be uncountable, as in $\{\mathcal{H}_\gamma\}_{\gamma \in \mathbb{R}}$ so that $\mathcal{H} = \bigcup_{\gamma \in \mathbb{R}} \mathcal{H}_\gamma$ becomes highly expressive.

Second, $\hat{R}_S(h)$ is often hard to optimize. Of course, we are not optimizing $\hat{R}_S(h)$ directly in SRM, but the difficulty of optimizing it may cause the opt. of SRM to be challenging. Note that this is not an issue specific to SRM. All of ERM, CV, n-folds CV use $\hat{R}_S(h)$, so that they all get affected by the difficulty of optimizing $\hat{R}_S(h)$.

(2). What should we do? A common approach for the first issue is to use regularization. In a simple incarnation of the approach,

we consider an uncountable family of hypo. sets.

$$\{\mathcal{H}_\gamma\}_{\gamma \in \mathbb{R}, \gamma > 0}, \qquad \mathcal{H} = \bigcup_{\gamma > 0} \mathcal{H}_\gamma.$$

s.t. $\mathcal{H}_\gamma = \{x \mapsto \text{sgn}(\langle w, \Phi(x) \rangle) \mid \|w\|_p \leq \gamma \text{ and } w \in \mathbb{R}^k\}$.

$\text{sgn}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0. \end{cases}$

where k and p are fixed natural numbers, Φ is a function from $\mathcal{X}$ to $\mathbb{R}^k$, and $\|w\|_p = \left(\sum_{i=1}^{k} |w_i|^p\right)^{1/p}$ is the p-norm of w. Then, we solve the following optimisation problem:

w parameter of h.

$$\underset{h \in \mathcal{H}}{\text{argmin}} \; \hat{R}_S(h) + \lambda \|w\|_p$$

called regularisation.

where λ is a hyper-parameter, (At a very high level, that is, it is a fixed constant). Just like SRM, this regularisation-based algorithm can be viewed as searching for a good hypothesis set and a good hypothesis at the same time.

(3) A popular approach for addressing the second issue is to use a convex surrogate loss. It applies to a setting where hypotheses h has the following form:

$$h: \mathcal{X} \to \mathcal{Y} = \{-1, 1\}.$$

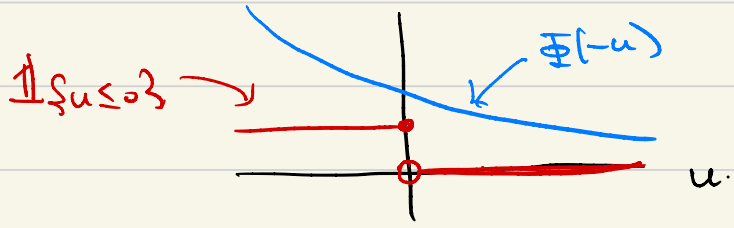
$$h(x) = \begin{cases} +1 & \text{if } f(x) \geq 0 \\ -1 & \text{otherwise.} \end{cases} = \text{sgn}(f(x)).$$

for some $f: \mathcal{X} \to \mathbb{R}$. We write $h_f(x)$ to mean $\text{sgn}(f(x))$.

To use the approach, we first have to pick a convex and non-increasing function $\Phi: [-\infty, +\infty] \to \mathbb{R}$ s.t.

$$\mathbb{1}\{u \le 0\} \le \Phi(-u) \quad \text{for all } u \in \mathbb{R},$$

which, in picture, means:



Next, we define:

$$L_\Phi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}.$$

$$L_\Phi(r, y) = \Phi(-yr)$$

$$\hat{R}_S^\Phi(h_f) = \frac{1}{m}\sum_{i=1}^m \Phi(-f(x_i)y_i). \qquad \left(S = ((x_1,y_1),\dots,(x_m,y_m))\right)$$

Finally, we substitute $\hat{R}_S^\Phi(h_f)$ for $\hat{R}_S(h_f)$, in an optimisation objective, which we saw or will see, and solve the resulting new optimisation problem.

Lemma. $\hat{R}_S(h_f) \le \hat{R}_S^\Phi(h_f)$.

Proof.

$$\hat{R}_S(h_f) \le \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{h_f(x_i) \neq y_i\} \qquad \left(S = ((x_1,y_1),\dots,(x_m,y_m))\right)$$

$$= \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{\operatorname{sgn}(f(x_i)) \neq y_i\}$$

$$\le \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{y_i f(x_i) \le 0\} \le \frac{1}{m}\sum_{i=1}^m \Phi(-y_i f(x_i)) = \frac{1}{m}\sum_{i=1}^m L_\Phi(f(x_i), y_i) = \hat{R}_S^\Phi(h_f).$$

□

Lemma. $\hat{R}_S^\Phi$ is convex. That is, $\forall \alpha \in (0,1)$ $\forall f, g : \mathcal{X} \to \mathbb{R}$,

$$\hat{R}_S^\Phi(h_{\alpha f + (1-\alpha)g}) \le \alpha \hat{R}_S^\Phi(h_f) + (1-\alpha)\hat{R}_S^\Phi(h_g).$$

Proof.

$$\hat{R}_S^\Phi(h_{\alpha f+(1-\alpha)g}) = \frac{1}{m}\sum_{i=1}^m \Phi\left(-y_i\left(\alpha f(x_i) + (1-\alpha)g(x_i)\right)\right)$$

$$\le \frac{1}{m}\sum_{i=1}^m \Phi\left(\alpha(-y_i f(x_i)) + (1-\alpha)(-y_i g(x_i))\right) \le \frac{1}{m}\sum_{i=1}^m \alpha\Phi(-y_i f(x_i)) + (1-\alpha)\Phi(-y_i g(x_i))$$

(convexity of Φ)

$$= \alpha\hat{R}_S^\Phi(h_f) + (1-\alpha)\hat{R}_S^\Phi(h_g)$$

□

(4) Examples.

① Hinge loss : $\Phi(u) = \max(0, 1+u)$

② Exponential loss : $\Phi(u) = \exp(u)$

③ Logistic loss : $\Phi(u) = \log_2(1+e^u)$

(5) Theoretical analysis.

① Bayes scoring fn ... $f^*(x) = \eta(x) - \frac{1}{2}$

where $\eta(x) = \mathbb{P}[y=1 \mid x]$.

Note that h_{f^*} is h_{Bayes}, the Bayes hypothesis that achieves the minimum risk called Bayes risk R^*.

② Let $R(f) = R(h_f) = \mathbb{E}_{x,y\sim D}[\mathbb{1}_{\{\operatorname{sgn}(f(x)) \neq y\}}]$.

Note R^* (Bayes risk) $= R(f^*)$.

Lemma 4.5 $R(f) - R^* \leq 2\mathbb{E}_{x\sim D_x}[|f^*(x)|\,\mathbb{1}_{\{f(x)f^*(x)\leq 0\}}]$

$= R(h_f) - R^*$

(i.e. excess error of h_f).

Proof: Let $\eta(x) = \mathbb{P}[y=1 \mid x]$. for any $f' : \mathcal{X} \to \mathbb{R}$,

$$R(f') = \mathbb{E}_{x\sim D_x}[\eta(x)\mathbb{1}_{\{f'(x)<0\}} + (1-\eta(x))\mathbb{1}_{\{f'(x)\geq 0\}}]$$

$$= \mathbb{E}_{x\sim D_x}[\eta(x)\mathbb{1}_{\{f'(x)<0\}} + (1-\eta(x))(1-\mathbb{1}_{\{f'(x)<0\}})]$$

$$= \mathbb{E}_x[[2\eta(x)-1]\mathbb{1}_{\{f'(x)<0\}} + (1-\eta(x))]$$

$$= \mathbb{E}_x[2f^*(x)\mathbb{1}_{\{f'(x)<0\}} + (1-\eta(x))]$$

Thus, $R(f) - R(f^*) = \mathbb{E}_x[2f^*(x)[\mathbb{1}_{\{f(x)<0\}} - \mathbb{1}_{\{f^*(x)<0\}}]]$

$$\leq \mathbb{E}_x[2f^*(x)\operatorname{sgn}(f^*(x))\,\mathbb{1}_{\{f(x)f^*(x)\leq 0 \,\wedge\, (f(x),f^*(x)) \neq (0,0)\}}]$$

case analysis on $f(x)<0$ and $f^*(x)<0$.

$$= 2\mathbb{E}_x[|f^*(x)|\,\mathbb{1}_{\{f(x)f^*(x)\leq 0\}}] \quad \square$$

(3) Bayes scoring function wrt. Φ:

$$f^*_\Phi(x) = \arg\min_{u \in [-\infty, \infty]} \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u)$$

This minimizes the new risk induced by Φ (which we explained in (3)) as shown below.

↳ new risk:

$$R^\Phi(f) = \mathbb{E}_{(x,y)\sim D}[\Phi(-yf(x))].$$

[Note that: $R^\Phi(f) = \mathbb{E}_S[\hat{R}^\Phi_S(f)]$]

$R^{\Phi,*} =$

[Lemma: $R^\Phi(f^*_\Phi) = \min_{f \in [\mathcal{X} \to \mathbb{R}]} R^\Phi(f)$.]

Proof: $R^\Phi(f) = \mathbb{E}_{(x,y)\sim D}[\Phi(-yf(x))]$

$= \mathbb{E}_{x \sim D_x}[\eta(x)\Phi(-f(x)) + (1-\eta(x))\Phi(f(x))]$

$\geq \mathbb{E}_{x \sim D_x}[\min_{u \in [\infty,\infty]} (\eta(x)\Phi(-u) + (1-\eta(x))\Phi(u))]$

$= R^\Phi(f^*_\Phi)$

[Lemma: $R(f) \leq R^\Phi(f)$ for all measurable $f: \mathcal{X} \to \mathbb{R}$.]

Proof:

$$R(f) = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}\{\text{sgn}(f(x)) \neq y\}]$$

$$\leq \mathbb{E}_{(x,y)\sim D}[\mathbb{1}\{yf(x) \leq 0\}] \leq \mathbb{E}_{(x,y)\sim D}[\Phi(-yf(x))]$$

$$= R^\Phi(f) \quad \square$$

Thm 4.7 Φ : convex, non-decreasing.

$\exists\ s \geq 1$ and $c > 0$ s.t. $\forall x \in \mathcal{X}$:

$$|f^*(x)|^s = \left|\eta(x) - \frac{1}{2}\right|^s \leq c^s \left[\Phi(0) - \mathbb{E}_{P[y|x]}\left[\Phi(-y f^*_\Phi(x))\right]\right]$$

$\Rightarrow$ for all $f : \mathcal{X} \to \mathbb{R}$,

$$\left(R(f) - R^*\right)^s \leq (2c)^s \left[R^\Phi(f) - R^{\Phi,*}\right]$$

Proof By the convexity of Φ,

$$\Phi(-f^*(x) f(x)) = \Phi\left((f(x)\eta(x) - 1)\right)$$

$$= \Phi\left(-\eta(x) f(x) + (1-\eta(x)) f(x)\right)$$

$$\leq \eta(x)\Phi(-f(x)) + (1-\eta(x))\Phi(f(x)) = \mathbb{E}_{P[y|x]}\left[\Phi(-yf(x))\right]$$

Jensen ↓ Lemma 4.5

$$R(f) - R(f^*) \leq \mathbb{E}_{x \sim D_x}\left[2|f^*(x)|\,\mathbb{1}_{\{\xi^*(x) f^*(x) f(x) \leq 0\}}\right]$$

$$= \mathbb{E}_x\left[\left(|2\eta(x) - 1|^s \mathbb{1}_{\{\xi^*(x) f^*(x) f(x) \leq 0\}}\right)^{\frac{1}{s}}\right]$$

$$\leq \mathbb{E}_x\left[|2\eta(x) - 1|^s \mathbb{1}_{\{\xi^*(x) f^*(x) f(x) \leq 0\}}\right]^{\frac{1}{s}}$$

$$\leq \mathbb{E}_x\left[c^s \left[\Phi(0) - \mathbb{E}_{P[y|x]}\left[\Phi(-y f^*_\Phi(x))\right]\right] \mathbb{1}_{\{\xi^*(x) f(x) \leq 0\}}\right]^{\frac{1}{s}}$$

Φ non-decreasing ↓

$$\leq 2c\, \mathbb{E}_x\left[\left[\Phi(-f^*(x) f(x)) - \mathbb{E}_{P[y|x]}\left[\Phi(-y f^*_\Phi(x))\right]\right] \mathbb{1}_{\{\xi^*(x) f(x) \leq 0\}}\right]^{\frac{1}{s}}$$

$$\leq 2c\, \mathbb{E}_x\left[\left[\mathbb{E}_{P[y|x]}\left[\Phi(-y f(x))\right] - \mathbb{E}_{P[y|x]}\left[\Phi(-y f^*_\Phi(x))\right]\right] \mathbb{1}_{\{\xi^*(x) f(x) \leq 0\}}\right]^{\frac{1}{s}}$$

$$\leq 2c\, \mathbb{E}_x\left[\left[\mathbb{E}_{P[y|x]}\left[\Phi(-y f(x))\right] - \mathbb{E}_{P[y|x]}\left[\Phi(-y f^*_\Phi(x))\right]\right]\right]^{\frac{1}{s}}$$

$$\leq 2c \left(R^\Phi(f) - R^{\Phi,*}\right)^{\frac{1}{s}}. \qquad \square$$

※ Hinge … $s = 1$, $c = \frac{1}{2}$, Exponential … $s = 2$, $c = \frac{1}{\sqrt{2}}$,

Logistic … $s = 2$, $c = \frac{1}{\sqrt{2}}$.