CS 7545: Machine Learning Theory **Fall 2018**

Lecture 17: Statistical Learning Theory

Lecturer: Jacob Abernethy *Scribes: Manon Huguenin, Hang Zhang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

17.1 Supervised Learning Setting

17.1.1 Key ingredients

The key ingredients for the *supervised learning setting* are the following:

1. Observation/Input space $\mathbb{X}$:
   e.g: $\mathbb{X} \in \mathbb{R}^d$, where d is the dimension of the space.

2. Label space $\mathbb{Y}$:
   e.g: - $\mathbb{Y} = \{0, 1\}$ "binary class".
   - $\mathbb{Y} = \{1, 2, 3, ..., K\}$ "multi-class".
   - $\mathbb{Y} = \mathbb{R}$ "regression".

3. Prediction space $\mathbb{Y}'$: may not be the same as $\mathbb{Y}$.
   e.g: $\mathbb{Y}' = [0, 1]$ while $\mathbb{Y} = \{0, 1\}$.

4. A distribution $\mathcal{D} \in \Delta(\mathbb{X} \times \mathbb{Y})$ associated with the instance of a learning problem:
   Samples $(\mathbf{x}, y)$ are drawn from $\mathcal{D}$ ($(\mathbf{x}, y) \sim \mathcal{D}$), where $\mathbf{x}$ is the random observation and y is its label (type).

5. A hypothesis space $\mathcal{H}$: a set of functions mapping $\mathbb{X}$ to $\mathbb{Y}$.
   e.g: a) "Decision stumps": $\mathcal{H} = \{h_{i,\alpha}(\mathbf{x}) = \mathbb{1}[x_i > \alpha] : \alpha \in \mathbb{R}, i = 1, \cdots, d\}$.
   b) "Linear thresholds": $\mathcal{H} = \{h_{\boldsymbol{\omega},b}(\mathbf{x}) = \mathbb{1}[\langle \boldsymbol{\omega}, \mathbf{x} \rangle \leq b] : \boldsymbol{\omega} \in \mathbb{R}^d, b \in \mathbb{R}\}$.
   c) "Neural networks": $\mathcal{H} = \{h_{\mathbf{M}_1,\mathbf{b}_1,\mathbf{M}_2,\mathbf{b}_2,\ldots,\mathbf{M}_K,\mathbf{b}_K}(\mathbf{x}) = \sigma(\mathbf{b}_K + \mathbf{M}_K\sigma(\mathbf{b}_{K-1} + \mathbf{M}_{K-1}\sigma(\ldots\sigma(\mathbf{b}_1 + \mathbf{M}_1\mathbf{x}))))$, for all matrices $\mathbf{M}_1, \ldots, \mathbf{M}_K$, and all offsets $\mathbf{b}_1, \ldots, \mathbf{b}_K$, and with $\sigma(\cdot)$ being the function such that $\sigma(\nu) = \frac{1}{1+\exp(\nu)}$.

6. A loss function $\ell : \mathbb{Y}' \times \mathbb{Y} \to \mathbb{R}$: the quantity $\ell(\hat{y}, y)$ is to describe "how bad is $\hat{y}$ an estimate of y".
   e.g: a) "0-1 loss": $\ell(\hat{y}, y) = \mathbb{1}[\hat{y} \neq y]$.
   b) "hinge loss": $\ell(\hat{y}, y) = \max(1 - \hat{y}y, 0)$, where $\mathbb{Y} = \{-1, 1\}$.
   c) "square loss": $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

17.1.2 Important concepts

Definition 17.1 (Risk/Generalization error) *Given a distribution $\mathcal{D} \in \Delta(\mathbb{X} \times \mathbb{Y})$, a loss function $\ell(\cdot, \cdot)$, the **risk** or **generalization error** of a hypothesis h is defined as*

$$\mathcal{R}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)].$$

Remark 17.2 *The **key idea** in statistical learning is that you want the risk (generalization error) to be small.*

Remark 17.3 *Because the distribution $\mathcal{D}$ is inaccessible, risk $\mathcal{R}(h)$ cannot be calculated directly. Instead, we adopt its empirical value as the substitution.*

Definition 17.4 (Empirical risk) *Given samples $\mathcal{S} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)\} \subseteq \mathbb{X} \times \mathbb{Y}$, the **empirical risk** of a hypothesis h is defined as*

$$\hat{\mathcal{R}}_m(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i).$$

17.2 Empirical Risk Minimization

17.2.1 Introduction

Definition 17.5 (*Empirical Risk Minimization* (ERM)) *Given a data set $\mathcal{S}$, the* empirical risk minimization *proposes to minimize the* empirical risk*, i.e.,*

$$\hat{h}_m^{\mathrm{ERM}} = \mathrm{argmin}_{h\in\mathcal{H}}\ \hat{\mathcal{R}}_m(h), \tag{17.1}$$

where $\hat{\mathcal{R}}_m(\cdot)$ denotes the empirical risk.

Example 17.6 (Least square regression) *We define here the hypothesis set as $\mathcal{H} = \{h_{\boldsymbol{\omega}}(\mathbf{x}) : \langle \boldsymbol{\omega},\ \mathbf{x}\rangle,\ \ \boldsymbol{\omega} \in \mathbb{R}^d\}$ and the loss function as $l(\hat{y}, y) = \|\hat{y} - y\|_2^2$. Then we can write the estimate $\boldsymbol{\omega}^*$ with ERM as*

$$\boldsymbol{\omega}^* = \mathrm{argmin}_{\boldsymbol{\omega}}\ \frac{1}{m}\left(\sum_{i=1}^{m} \|y_i - \langle \boldsymbol{\omega},\ \mathbf{x}_i\rangle\|_2^2\right) \Longrightarrow \boldsymbol{\omega}^* = \underbrace{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}}_{\boldsymbol{X}^{\dagger}}\boldsymbol{Y},$$

where $\boldsymbol{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_m]$, $\boldsymbol{Y} = [y_1\ y_2\ \cdots\ y_m]^{\mathrm{T}}$, and $(\cdot)^{\dagger}$ denotes the Moore-Penrose inverse*.*

17.2.2 Evaluation Quantities

To evaluate the algorithm, there are two important evaluation quantities: **estimation error** and **approximation error**.

Definition 17.7 (Estimation Error) *For ERM hypothesis $\hat{h}_m^{\mathrm{ERM}}$, its* estimation error *is defined as*

$$\mathcal{R}(\hat{h}_m^{\mathrm{ERM}}) - \min_{h^*\in\mathcal{H}} \mathcal{R}(h^*).$$

Remark 17.8 *The first term $\mathcal{R}(\hat{h}_m^{\mathrm{ERM}})$ corresponds to the risk of $\hat{h}_m^{\mathrm{ERM}}$.*

The second term $\min_{h^*\in\mathcal{H}} \mathcal{R}(h^*)$ corresponds to the risk of the best hypothesis h^* that can be chosen from the hypothesis set $\mathcal{H}$.

Definition 17.9 (Approximation Error) ***Approximation error** is defined as*

$$\min_{h^*\in\mathcal{H}} \mathcal{R}(h^*) - \min_{\substack{h^{**}\in \text{ all possible} \\ \text{functions}}} \mathcal{R}(h^{**}),$$

*where the second term $\min_{\substack{h^{**}\in \text{ all possible} \\ \text{functions}}} \mathcal{R}(h^{**})$ is also known as the **Bayes risk**.*

Fact 17.10 (Bias Variance Trade-off) [1] *The above two quantities are affected by the complexity of the hypothesis set $\mathcal{H}$: as the hypothesis set $\mathcal{H}$ becomes more complex,* [2] *the estimation error goes up ($\uparrow$) while the approximation error goes down ($\downarrow$).*

Here the "variance" *corresponds to the* estimation error*, since they measure the sensitivity to the data set $\mathcal{S}$, while* "bias" *corresponds to the* approximation error*.*

[1] This terminology is suitable only when the loss function is $\|\cdot\|_2^2$ while inaccurate when other loss function is adopted.

[2] Example of more complex hypothesis sets include: (*i*) polynomials of higher degree, (*ii*) neuron networks with more layers, and (*iii*) decision trees with greater depth.

17.2.3 Main Results

Here we concentrate on analyzing the *estimation error*.

Bound of estimation error: First we have

$$\begin{aligned}
&\mathcal{R}(\hat{h}_m^{\mathrm{ERM}}) - \min_{h^*\in\mathcal{H}} \mathcal{R}(h^*) \\
=\;& \underbrace{\mathcal{R}(\hat{h}_m^{\mathrm{ERM}}) - \hat{\mathcal{R}}_m(\hat{h}_m^{\mathrm{ERM}})}_{\mathcal{T}_1} + \underbrace{\hat{\mathcal{R}}_m(\hat{h}_m^{\mathrm{ERM}}) - \hat{\mathcal{R}}_m(h^*)}_{\mathcal{T}_2} + \underbrace{\hat{\mathcal{R}}_m(h^*) - \mathcal{R}(h^*)}_{\mathcal{T}_3} \\
\overset{(i)}{\leq}\;& \mathcal{T}_1 + \mathcal{T}_3 \overset{(ii)}{\leq} 2\sup_{h\in\mathcal{H}} |\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)|,
\end{aligned} \tag{17.2}$$

where (i) is because $\mathcal{T}_2 \leq 0$ according to the definition of $\hat{h}_m^{\mathrm{ERM}}$ in Eqn (17.1), and in (ii) we use $\mathcal{T}_1 \leq \sup_{h\in\mathcal{H}} |\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)|$ and $\mathcal{T}_3 \leq \sup_{h\in\mathcal{H}} |\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)|$.

Remark 17.11 *If we can control the quantity* $\sup_{h\in\mathcal{H}} |\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)|$ *in Eqn (17.2), this bound is called the* uniform deviation bound*: (i) the first term $\hat{\mathcal{R}}_m(h)$ corresponds to the* training error*; (ii) the second term $\mathcal{R}(h)$ corresponds to the* test error*.*

Hence, bounding the approximation error transforms to the problem of bounding the *uniform deviation bound*.

Non-asymptotic bound: We now want to find a non-asymptotic bound to the quantity $|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)|$.

Theorem 17.12 *If $\mathcal{H}$ is finite and $\ell(\cdot,\cdot)$ is within* $[0,\ 1]$*, then the following inequality,*

$$\sup_{h\in\mathcal{H}} |\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)| \leq \sqrt{\frac{\log \frac{2|\mathcal{H}|}{\delta}}{2m}},$$

holds with probability at least $1-\delta$, where $|\cdot|$ denotes the cardinality.

Remark 17.13 *Although the increasing m leads to a smaller bound, a complex hypothesis set $\mathcal{H}$, which has large $|\mathcal{H}|$, compensates for the decreasing smaller bound.*

Remark 17.14 *Also, the bound is $\mathcal{O}(\sqrt{\log|\mathcal{H}|})$ and is consistent with the bound of the* Halving Algorithm*.*

Remark 17.15 *One wrong way of bounding is attached in the Appendix. 17.3.*

Sketch of proof: Instead of lower bounding $\Pr\left\{\sup_{h\in\mathcal{H}} \left|\hat{\mathcal{R}}_m(h) - \ \mathcal{R}(h)\right| \leq t\right\}$, we upper bound its complementary event $\Pr\left\{\sup_{h\in\mathcal{H}}\left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| \geq t\right\}$ as

$$
\begin{aligned}
\Pr\left\{\sup_{h\in\mathcal{H}}\left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| \geq t\right\} &= \Pr\left\{\bigcup_{h\in\mathcal{H}}\left\{\left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| \geq t\right\}\right\} \\
\overset{(a)}{\leq} \sum_{h\in\mathcal{H}}\Pr\left\{\left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| \geq t\right\} &= \sum_{h\in\mathcal{H}}\Pr\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\ell(h(\mathbf{x}_i), y_i) - \mathbb{E}\ell(h(\mathbf{x}),\ y)\right| \geq t\right\} \\
\overset{(b)}{=} \sum_{h\in\mathcal{H}}\Pr\left\{\left|\frac{1}{m}\left(\sum_{i=1}^{m}\ell(h(\mathbf{x}_i), y_i) - \sum_{i=1}^{m}\mathbb{E}\ell(h(\mathbf{x}_i), y_i)\right)\right| \geq t\right\} \\
= \sum_{h\in\mathcal{H}}\Pr\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\left(\ell(h(\mathbf{x}_i), y_i) - \mathbb{E}\ell(h(\mathbf{x}_i), y_i)\right)\right| \geq t\right\} \\
\overset{(c)}{\leq} \sum_{h\in\mathcal{H}} 2\exp\left(-2mt^2\right) = 2|\mathcal{H}|\exp\left(-2mt^2\right),
\end{aligned}
$$

where (a) is because of the union bound, (b) is because $(\mathbf{x}_i,\ y_i) \overset{\text{i.i.d}}{\sim} \mathcal{D}$ and hence $\mathbb{E}\ell(h(\mathbf{x}_i), y_i) = \mathbb{E}l(h(\mathbf{x}), y)$, and in (c) we have $\ell(h(\mathbf{x}), y) \in [0,\ 1]$ and use Hoeffding's inequality.

Then we have

$$\Pr\left\{\sup_{h\in\mathcal{H}}\left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| \leq t\right\} \geq 1 - 2|\mathcal{H}|\exp\left(-2mt^2\right),$$

which completes the proof when we set t^* to satisfy $2|\mathcal{H}|\exp\left(-2mt^{*2}\right) = \delta$.

17.3 Appendix A: wrong way of bounding uniform deviation

Wrong bounding method: To bound Eqn (17.2), the goal is to bound $\sup_{h\in\mathcal{H}}|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)|$ as the following:

$$
\begin{aligned}
\hat{\mathcal{R}}_m(h) - \mathcal{R}(h) &= \frac{1}{m}\sum_{i=1}^{m}\ell(h(\mathbf{x}_i, y_i)) - \mathbb{E}\ell(h(\mathbf{x}, y)) \\
\overset{(a)}{=} \frac{1}{m}\sum_{i=1}^{m}\ell(h(\mathbf{x}_i, y_i)) - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\ell(h(\mathbf{x}_i, y_i)) &= \frac{1}{m}\sum_{i=1}^{m}\underbrace{\left(\ell(h(\mathbf{x}_i, y_i)) - \mathbb{E}\ell(h(\mathbf{x}_i, y_i))\right)}_{X_i},
\end{aligned}
$$

where in (a) we use the fact that $(\mathbf{x}_i,\ y_i) \sim \mathcal{D}$. Since $\ell(\cdot,\cdot)$ is within $[0, 1]$, we have X_i to be a RV within $[0, 1]$. With Hoeffding's inequality, we have

$$\left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| = \left|\frac{1}{m}\sum_{i=1}^{m}X_i\right| \leq \sqrt{\frac{\log\frac{2}{\delta}}{2m}},$$

with probability at least $1 - \delta$.

Problem: If we perform back-substitution, we find that large sample size, namely m, will always lead to smaller error, which is **WRONG**!! So where do we make the mistake?

Answer: When we bound $|\hat{\mathcal{R}}(h) - \mathcal{R}(h)|$ with Hoeffding's inequality, we require X_i to be independent. However, the ERM hypothesis $\hat{h}_m^{\text{ERM}}$ makes samples correlated and violates this assumption. Hence, the above derivation is incorrect.