Computational learning theory

CE-477: Machine Learning - CS-828: Theory of Machine Learning
Sharif University of Technology
Fall 2024

Fatemeh Seyyedsalehi

References of the lecture are mentioned in the last slide

Feasibility of learning

- Does the training set $\mathcal{D}$ tell us anything out of $\mathcal{D}$?
  - $\mathcal{D}$ does <u>not tells</u> us something <u>certain</u> about f outside of $\mathcal{D}$
  - However, it can <u>tell</u> us something <u>likely</u> about f outside of $\mathcal{D}$

- Probability helps us to find learning theory

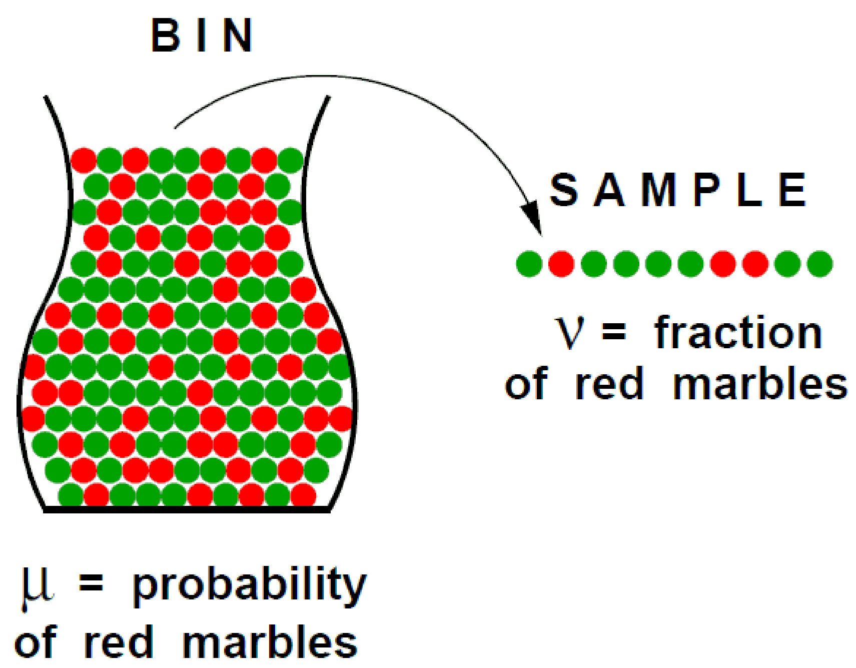
Generalizability of Learning

- Generalization error is important to us
- Why should doing well on the training set tell us anything about generalization error?
  - Can we relate error on training set to generalization error?
- Which are conditions under which we can actually prove that learning algorithms will work well?

Two lemma from the probability

- Considering i.i.d. random variables $Z_1, Z_2, \dots, Z_n$ from a Bernoulli distribution with parameter μ.
  - We can estimate μ as follows

$$\hat{\mu} = \frac{1}{n} \Sigma_{i=1}^{n} Z_i$$



Two lemma from the probability

- Considering i.i.d. random variables $Z_1, Z_2, \dots, Z_n$ from a Bernoulli distribution with parameter μ.
  - We can estimate μ as follows

$$\hat{\mu} = \frac{1}{n} \Sigma_{i=1}^{n} Z_i$$

- Hoeffding's Inequality
  - In a big sample (large N), $\hat{\mu}$ is probably close to μ (within ϵ):

$$\Pr[|\hat{\mu} - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

Two lemma from the probability

- The union bound
  - An axiom in probability theory

$$P(A_1 \cup A_2 \cup \cdots A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$$

- Using just these two lemmas, we will be able to prove some of the deepest and most important results in learning theory.

PAC framework

- **P**robably **a**pproximately **c**orrect

- A framework and set of assumptions under which numerous results on learning theory were proved.
  - Training and testing data are on the same distribution
  - independently drawn training examples

- To simplify our exposition, let's restrict our attention to binary classification. Everything we'll say here generalizes to other problems, including regression and multi-class classification.

Generalization error

- Given a training set as follows,

$$S = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{i=1}^{n}$$

- The training error

$$\bar{\varepsilon}\,(h) = \frac{1}{|S|}\sum_{i=1}^{n} I\big(h(\boldsymbol{x}^{\boldsymbol{i}}) \neq y^{i}\big)$$

- The true (generalization, test) error

$$\varepsilon(h) = E_{\boldsymbol{x}\sim P(X)}[I(h(\boldsymbol{x}) \neq y)]$$

Generalization error

- Our learning algorithm is based on the minimization of the empirical risk (training error)

$$\hat{h} = \underset{h \in H}{\operatorname{argmin}}\, \bar{\varepsilon}\,(h)$$

- We would like give guarantees on the generalization error of $\hat{h}$.
  - First we show that $\bar{\varepsilon}\,(h)$ is a good approximate of $\varepsilon(h)$
  - Second we find an upper bound for the generalization error of $\hat{h}$

Generalization error

- Consider a specific hypothesis h_i and training set S, we define the following Bernoulli random variable,

$$Z = 1\{h_i(x) \neq y\}$$

- Real mean of this Bernoulli distribution $\varepsilon(h_i)$

- Now consider following random samples,

$$Z_j = 1\{h_i(x^{(j)}) \neq y^{(j)}\}$$

Generalization error

- The training error for this specific hypothesis

$$\hat{\varepsilon}(h_i) = \frac{1}{n}\sum_{j=1}^{n} Z_j.$$

- Now, we can apply the Hoeffding inequality

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2\exp(-2\gamma^2 n).$$

- This shows that, for our particular h_i, training error will be close to generalization error with high probability, assuming n is large.

Generalization error

- We show that the generalization error is close to the training error for a particular hypothesis h_i,
- Considering A_i as,

$$|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$$

- We have

$$P(A_i) \leq 2\exp(-2\gamma^2 n)$$

Generalization error

- Using the union bound,

$$
\begin{aligned}
P(\exists\, h \in \mathcal{H}. |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \cdots \cup A_k) \\
&\leq \sum_{i=1}^{k} P(A_i) \\
&\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 n) \\
&= 2k\exp(-2\gamma^2 n)
\end{aligned}
$$

Generalization error

- Subtract both sides from 1,

$$
\begin{aligned}
P(\neg\exists\, h \in \mathcal{H}. |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(\forall h \in \mathcal{H}. |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \\
&\geq 1 - 2k \exp(-2\gamma^2 n)
\end{aligned}
$$

- This is called a **uniform convergence result**, because this is a bound that holds simultaneously for all h

Generalization error

- In the discussion above, what we did was, for particular values of n and γ give a bound on the probability of

$$h \in \mathcal{H}, \ |\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma.$$

- Three quantities of interest: n, γ and the probability of error

Generalization error

For instance, we can ask the following question: Given γ and some $\delta > 0$, how large must n be before we can guarantee that with probability at least $1 - \delta$, training error will be within γ of generalization error? By setting $\delta = 2k\exp(-2\gamma^2 n)$ and solving for n, [you should convince yourself this is the right thing to do!], we find that if

$$n \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta},$$

Generalization error

Similarly, we can also hold n and δ fixed and solve for γ in the previous equation, and show [again, convince yourself that this is right!] that with probability $1-\delta$, we have that for all $h \in \mathcal{H}$,

$$|\hat{\varepsilon}(h) - \varepsilon(h)| \leq \sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

Generalization error

Define $h^* = \arg\min_{h\in\mathcal{H}} \varepsilon(h)$ to be the best possible hypothesis in $\mathcal{H}$. Note that h^* is the best that we could possibly do given that we are using $\mathcal{H}$, so it makes sense to compare our performance to that of h^*. We have:

$$\begin{aligned} \varepsilon(\hat{h}) &\le \hat{\varepsilon}(\hat{h}) + \gamma \\ &\le \hat{\varepsilon}(h^*) + \gamma \\ &\le \varepsilon(h^*) + 2\gamma \end{aligned}$$

Generalization error

Theorem. Let $|\mathcal{H}| = k$, and let any n, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

The sample complexity bound

Corollary. Let $|\mathcal{H}| = k$, and let any δ, γ be fixed. Then for $\varepsilon(\hat{h}) \leq \min_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$ to hold with probability at least $1 - \delta$, it suffices that

$$\begin{aligned} n &\geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta} \\ &= O\left(\frac{1}{\gamma^2} \log \frac{k}{\delta}\right), \end{aligned}$$

Optional reading

- When the hypothesis space is infinite
  - We can not use the size of hypothesis space in our bounds directly.
  - Instead, we can use the Vapnik-Chervonenkis VC dimension of a hypothesis space that measures the size (capacity, complexity, expressive power, …) of it.
  - New bounds can be constructed based on the VC dimension.

References

- Andrew NG., Stafnord CS229 main_notes.pdf