Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

Bounds on Probabilities

Outline

- Recap of statistical learning problem
- Motivation for probability bounds
- Recap: CLT, Markov inequality
- Chernoff bound and Hoeffding's inequality

Risk and risk minimisation

- Risk / generalisation error of predictor h with respect to distribution $\mathcal{D}$

$$L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(h(x), y)\big]$$

  - We sample $(x, y) \sim \mathcal{D}$, and measure the expected error of h

Risk and risk minimisation

- Risk / generalisation error of predictor h with respect to distribution $\mathcal{D}$

$$L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(h(x), y)\big]$$

  - We sample $(x, y) \sim \mathcal{D}$, and measure the expected error of h

- Risk minimisation: We "ideally" want a predictor with low risk

$$\underset{h\in\mathcal{Y}^{\mathcal{X}}}{\text{minimise}}\ L_{\mathcal{D}}(h)$$

  - We cannot compute $L_{\mathcal{D}}(h)$ without knowledge of $\mathcal{D}$
  - Learner $\mathcal{A}$ only has access to training sample $S \sim \mathcal{D}^m$

Empirical risk minimisation

- Empirical risk of h on sample S

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i)$$

  - $L_S(h) =$ training error of h w.r.t. sample S
  - For $S \sim \mathcal{D}^m$, sample average is an estimate of $L_{\mathcal{D}}(h)$

Empirical risk minimisation

- Empirical risk of h on sample S

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i)$$

  - $L_S(h) =$ training error of h w.r.t. sample S
  - For $S \sim \mathcal{D}^m$, sample average is an estimate of $L_{\mathcal{D}}(h)$

- Empirical risk minimisation (ERM)

$$\underset{h \in \mathcal{Y}^{\mathcal{X}}}{\text{minimise}} \; L_S(h)$$

  - Replace $L_{\mathcal{D}}(h)$ by its estimate computed on training sample S

Is solving ERM same as solving RM?

$$\text{RM: } \underset{h \in \mathcal{Y}^{\mathcal{X}}}{\text{minimise}} \; \underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(h(x), y)\big]}_{L_{\mathcal{D}}(h)} \qquad \text{vs.} \qquad \text{ERM: } \underset{h \in \mathcal{Y}^{\mathcal{X}}}{\text{minimise}} \; \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i)}_{L_S(h)}$$

Is solving ERM same as solving RM?

$$\text{RM: } \underset{h \in \mathcal{Y}^{\mathcal{X}}}{\text{minimise}} \ \underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h(x), y)\right]}_{L_{\mathcal{D}}(h)} \qquad \text{vs.} \qquad \text{ERM: } \underset{h \in \mathcal{Y}^{\mathcal{X}}}{\text{minimise}} \ \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i)}_{L_S(h)}$$

- We first need to understand relation between $L_{\mathcal{D}}(h)$ and $L_S(h)$ for any predictor h

Is solving ERM same as solving RM?

$$\text{RM: } \underset{h\in\mathcal{Y}^{\mathcal{X}}}{\text{minimise}} \; \underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(h(x),y)\big]}_{L_{\mathcal{D}}(h)} \qquad \text{vs.} \qquad \text{ERM: } \underset{h\in\mathcal{Y}^{\mathcal{X}}}{\text{minimise}} \; \underbrace{\frac{1}{m}\sum_{i=1}^{m}\ell(h(x_i),y_i)}_{L_S(h)}$$

- We first need to understand relation between $L_{\mathcal{D}}(h)$ and $L_S(h)$ for any predictor h
- **Verify:** $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)] = \mathbb{E}_{S\sim\mathcal{D}^m}[L_S(h)] = L_{\mathcal{D}}(h)$ for any fixed h
- But how close is $L_S(h)$, which is random, close to its expectation

Infinite sample setting, $m \to \infty$

- For a fixed h, define $Z_i = \ell(h(x_i), y_i)$ and $\mu = L_{\mathcal{D}}(h) = \mathbb{E}[Z_i]$

Infinite sample setting, $m \to \infty$

- For a fixed h, define $Z_i = \ell(h(x_i), y_i)$ and $\mu = L_{\mathcal{D}}(h) = \mathbb{E}[Z_i]$

- Law of large numbers: As $m \to \infty$,

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} Z_i \longrightarrow \mu = L_{\mathcal{D}}(h) \quad \text{in probability}$$

Infinite sample setting, $m \to \infty$

- For a fixed h, define $Z_i = \ell(h(x_i), y_i)$ and $\mu = L_{\mathcal{D}}(h) = \mathbb{E}[Z_i]$

- Law of large numbers: As $m \to \infty$,

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} Z_i \longrightarrow \mu = L_{\mathcal{D}}(h) \quad \text{in probability}$$

- Does this imply that as $m \to \infty$,

$\widehat{h}_S$ that minimises $L_S(h)$ is also minimiser for $L_{\mathcal{D}}(h)$?

Infinite sample setting, $m \to \infty$

- For a fixed h, define $Z_i = \ell(h(x_i), y_i)$ and $\mu = L_\mathcal{D}(h) = \mathbb{E}[Z_i]$

- Law of large numbers: As $m \to \infty$,

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} Z_i \longrightarrow \mu = L_\mathcal{D}(h) \quad \text{in probability}$$

- Does this imply that as $m \to \infty$,

  $\widehat{h}_S$ that minimises $L_S(h)$ is also minimiser for $L_\mathcal{D}(h)$?

  - Not necessarily since we need converge in probability simultaneously for all h (will discuss later)

Finite m (the practical case)

- For a fixed h, how far is $L_S(h)$ from $L_{\mathcal{D}}(h)$?

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq$$

Finite m (the practical case)

- For a fixed h, how far is $L_S(h)$ from $L_{\mathcal{D}}(h)$?

- Assume 0-1 loss: $Z_i = \mathbf{1}\left\{h(x_i) \neq y_i\right\}$

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq$$

Finite m (the practical case)

- For a fixed h, how far is $L_S(h)$ from $L_{\mathcal{D}}(h)$?
- Assume 0-1 loss: $Z_i = \mathbf{1}\{h(x_i) \neq y_i\}$
  - $Z_1, \ldots, Z_m$ independent and identically distributed Bernoullis

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq$$

Finite m (the practical case)

- For a fixed h, how far is $L_S(h)$ from $L_{\mathcal{D}}(h)$?

- Assume 0-1 loss: $Z_i = \mathbf{1}\{h(x_i) \neq y_i\}$
  - $Z_1, \ldots, Z_m$ independent and identically distributed Bernoullis
  - $\mathbb{E}[Z_i] = \mu$ and Variance$(Z_i) = \mu(1-\mu)$

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(\left| \frac{1}{m} \sum_{i=1}^{m} Z_i - \mu \right| > \epsilon \right) \leq$$

Finite m (the practical case)

- For a fixed h, how far is $L_S(h)$ from $L_{\mathcal{D}}(h)$?

- Assume 0-1 loss: $Z_i = \mathbf{1}\{h(x_i) \neq y_i\}$
  - $Z_1, \ldots, Z_m$ independent and identically distributed Bernoullis
  - $\mathbb{E}[Z_i] = \mu$ and Variance$(Z_i) = \mu(1-\mu)$

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\text{Variance}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i\right) \quad \ldots \text{ Chebyshev inequality}$$
$$= \frac{\mu(1-\mu)}{m\epsilon^2}$$

Hoeffding's inequality

- Chebyshev inequality gives poor dependence on ϵ, $\mathbb{P}(|\cdot| > \epsilon) \leq O(\frac{1}{\epsilon^2})$

- We will later need a sharper bound, $\mathbb{P}(|\cdot| > \epsilon) \leq e^{-O(\epsilon^2)}$

Hoeffding's inequality

- Chebyshev inequality gives poor dependence on ϵ, $\mathbb{P}(|\cdot| > \epsilon) \leq O(\frac{1}{\epsilon^2})$
- We will later need a sharper bound, $\mathbb{P}(|\cdot| > \epsilon) \leq e^{-O(\epsilon^2)}$

Theorem Appendix.1 (Hoeffding's inequality)

Let $Z_1, \ldots, Z_m$ be m independent random variables such that $\mathbb{P}(Z_i \in [a_i, b_i]) = 1$ for all i.

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} (Z_i - \mathbb{E}[Z_i])\right| > t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

Proof of Hoeffding's inequality

Denote $S_m = \sum_{i=1}^m Z_i$. For the proof, we proceed in three steps.

- Chernoff bounding: For all $s > 0$ we have

$$\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > t\right) \leq e^{-st} \prod_{i=1}^{n} \mathbb{E}\left[\exp(s\left(Z_i - \mathbb{E}\left[Z_i\right]\right))\right]$$

- Hoeffding's Lemma: For all $s > 0$ and all $i \in [m]$ we have

$$\mathbb{E}\left[\exp\left(s\left(Z_i - \mathbb{E}\left[Z_i\right]\right)\right)\right] \leq \exp\left(\frac{s^2\left(b_i - a_i\right)^2}{8}\right)$$

- Plug in and find $s > 0$ to obtain the sharpest upper bound.

Proof of Hoeffding's inequality (Chernoff bound)

Let Z be a random variable and let $s > 0$. Then,

$$\begin{aligned}
\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > t\right) &= \mathbb{P}\left(\exp\left(s(S_m - \mathbb{E}\left[S_m\right])\right) > e^{st}\right) \\
&\leq e^{-st} \cdot \mathbb{E}\left[\exp\left(s(S_m - \mathbb{E}\left[S_m\right])\right)\right] \\
&= e^{-st} \prod_{i=1}^{n} \mathbb{E}\left[\exp(s\left(Z_i - \mathbb{E}\left[Z_i\right]\right))\right]
\end{aligned}$$

First, we use the Markov inequality. Then, we use independence of $Z_1, \ldots, Z_m$.

Proof of Hoeffding's inequality (Hoeffding's Lemma)

Let Z have zero mean and assume $Z \in [a, b]$ almost surely. Then,

$$\begin{aligned}\mathbb{E}\left[\exp(sZ)\right] &\leq \mathbb{E}\left[\frac{b-Z}{b-a}e^{sa} + \frac{Z-a}{b-a}e^{sb}\right] \\ &= e^{sa}\left(\frac{b}{b-a} + \frac{-a}{b-a}e^{s(b-a)}\right)\end{aligned}$$

First, we use the fact that $z \mapsto e^{sz}$ is a convex function. Then, plug in $\mathbb{E}\left[Z\right] = 0$. Defining $\theta = \frac{-a}{b-a}$ and $u = s(b-a)$ this can be written as $e^{\varphi(u)}$ for a function

$$\varphi(u) = -\theta u + \log\left(1 - \theta + \theta e^{u}\right)$$

The next step will be to bound $\varphi(u)$.

Proof of Hoeffding's inequality (Hoeffding's Lemma)

By Taylor expansion there exists some $v \in [0, u]$ such that $\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{1}{2}u^2\varphi''(v)$. Taking derivatives, we compute

$$\varphi(0) = \varphi'(0) = 0$$
$$\varphi''(v) = \left(\frac{\theta e^v}{1-\theta+\theta e^v}\right) \cdot \left(1 - \frac{\theta e^v}{1-\theta+\theta e^v}\right) \leq \frac{1}{4}$$

where we observe $\left(\frac{\theta e^v}{1-\theta+\theta e^v}\right) \leq 1$ and note that $x \mapsto x(1-x)$ is bounded by $\frac{1}{4}$ on $[0, 1]$. This yields

$$\varphi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$$

and implies Hoeffding's Lemma.

Hoeffding's inequality (finding the best $s > 0$)

What we have so far:

$$\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > t\right) \leq e^{-st} \prod_{i=1}^{m} \exp\left(\frac{s^2\left(b_i - a_i\right)^2}{8}\right)$$

Up to now, $s > 0$ was arbitrary. We now minimise the RHS with respect to s to find the best upper bound. Simple calculus yields $s = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$ and thus

$$\min_{s>0}\left(e^{-st} \prod_{i=1}^{m} \exp\left(\frac{s^2\left(b_i - a_i\right)^2}{8}\right)\right) = \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right)$$

This proves Hoeffding's inequality.

In-class exercise: Chebyshev vs. Hoeffding

- Let $Z_1, \ldots, Z_m \sim_{iid}$ Bernoulli(μ), $\mu = 0.5$
- Chebyshev's inequality: $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq \frac{0.25}{m\epsilon^2}$
- Hoeffding's inequality: $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq 2e^{-2m\epsilon^2}$

In-class exercise: Chebyshev vs. Hoeffding

- Let $Z_1, \ldots, Z_m \sim_{iid}$ Bernoulli(μ), $\mu = 0.5$
- Chebyshev's inequality: $\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} Z_i - \mu \right| > \epsilon \right) \leq \frac{0.25}{m\epsilon^2}$
- Hoeffding's inequality: $\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} Z_i - \mu \right| > \epsilon \right) \leq 2e^{-2m\epsilon^2}$
- How do the bounds compare for $m = 10^3$ and $\epsilon = 10^{-2}$?

In-class exercise: Chebyshev vs. Hoeffding

- Let $Z_1, \ldots, Z_m \sim_{iid}$ Bernoulli(μ), $\mu = 0.5$

- Chebyshev's inequality: $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq \frac{0.25}{m\epsilon^2}$

- Hoeffding's inequality: $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right) \leq 2e^{-2m\epsilon^2}$

- How do the bounds compare for $m = 10^3$ and $\epsilon = 10^{-2}$?

- Suppose we wish to fix the bound $\mathbb{P}(|\cdot| > \epsilon) \leq 10^{-3}$.
  How large should ϵ be, assuming $m = 10^3$?