ECE595 / STAT598: Machine Learning I

Lecture 23.3: Probability Inequality - Advance Inequalities II

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

- Lecture 22 Is Learning Feasible?
- Lecture 23 Probability Inequality
- Lecture 24 Probably Approximate Correct

Today's Lecture:

- Basic Inequalities
  - Markov and Chebyshev
  - Interpreting the results
- Advance Inequalities
  - Chernoff inequality
  - Hoeffding inequality

Hoeffding Inequality

Let us revisit the Bad event:

$$
\begin{aligned}
\mathbb{P}[|\nu - \mu| \geq \epsilon] &= \mathbb{P}[\nu - \mu \geq \epsilon \quad \textbf{or} \quad \nu - \mu \leq -\epsilon] \\
&\leq \underbrace{\mathbb{P}[\nu - \mu \geq \epsilon]}_{\leq A} + \underbrace{\mathbb{P}[\nu - \mu \leq -\epsilon]}_{\leq A}, \qquad \text{Union bound} \\
&\leq 2A, \qquad \text{(What is } A\text{? To be discussed.)}
\end{aligned}
$$

Theorem (Hoeffding Inequality)

Let $X_1, \ldots, X_N$ be random variables with $0 \leq X_n \leq 1$, then

$$
\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq 2\underbrace{e^{-2\epsilon^2 N}}_{=A}
$$

The e-trick + Markov Inequality

Let us check one side:

$$\mathbb{P}[\nu - \mu \geq \epsilon] = \mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} X_n - \mu \geq \epsilon\right] = \mathbb{P}\left[\sum_{n=1}^{N}(X_n - \mu) \geq \epsilon N\right]$$

$$= \mathbb{P}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)} \geq e^{s\epsilon N}\right], \qquad \forall s > 0$$

$$\leq \frac{\mathbb{E}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)}\right]}{e^{s\epsilon N}}, \qquad \text{Markov Inequality}$$

$$= \left(\frac{\mathbb{E}\left[e^{s(X_n-\mu)}\right]}{e^{s\epsilon}}\right)^N, \qquad \text{Independence}$$

If we let $Z_n = X_n - \mu$, then

$$\mathbb{E}[e^{s(X_n-\mu)}] = M_{Z_n}(s) = \text{MGF of } Z_n.$$

Hoeffding Lemma

So now we have

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq \left(\frac{\mathbb{E}\left[e^{s(X_n - \mu)}\right]}{e^{s\epsilon}} \right)^N$$

Lemma (Hoeffding Lemma)

If $a \leq X_n \leq b$, then

$$\mathbb{E}\left[e^{s(X_n-\mu)}\right] \leq e^{\frac{s^2(b-a)^2}{8}}$$

This leads to

$$\begin{aligned} \mathbb{P}[\nu - \mu \geq \epsilon] &= \left(\frac{\mathbb{E}\left[e^{s(X_n - \mu)}\right]}{e^{s\epsilon}} \right)^N \\ &\leq \left(\frac{e^{\frac{s^2}{8}}}{e^{s\epsilon}} \right)^N = e^{\frac{s^2 N}{8} - s\epsilon N}, \qquad \forall s > 0. \end{aligned}$$

Minimization

Finally, we arrive at:

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{s^2 N}{8} - s\epsilon N}.$$

Since holds for all $s > 0$, in particular it holds for the minimizer:

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{s_{\min}{}^2 N}{8} - s_{\min}\epsilon N} = \min_{s>0} \left\{ e^{\frac{s^2 N}{8} - s\epsilon N} \right\}$$

Minimizing the exponent gives: $\frac{d}{ds}\left\{\frac{s^2 N}{8} - s\epsilon N\right\} = \frac{sN}{4} - \epsilon N = 0$. So $s = 4\epsilon$.

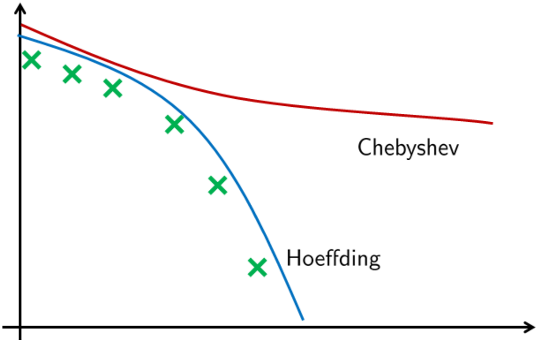
$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{(4\epsilon)^2 N}{8} - (4\epsilon)\epsilon N} = e^{-2\epsilon^2 N}.$$

Hoeffding Inequality

Theorem (Hoeffding Inequality)

Let $X_1, \ldots, X_N$ be random variables with $0 \leq X_n \leq 1$, then

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$



Compare Hoeffding and Chebyshev

Chebyshev:

$$\mathbb{P}\left[|\nu - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{N\epsilon^2}.$$

Hoeffding:

$$\mathbb{P}\left[|\nu - \mu| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 N}.$$

Both are in the form of

$$\mathbb{P}\left[|\nu - \mu| \geq \epsilon\right] \leq \delta.$$

Equivalent to: **For probability at least** $1 - \delta$, we have

$$\mu - \epsilon \leq \nu \leq \mu + \epsilon.$$

Error bar / **Confidence interval** of ν.

$$\delta = \frac{\sigma^2}{N\epsilon^2} \;\Rightarrow\; \epsilon = \frac{\sigma}{\sqrt{\delta N}}$$

$$\delta = 2e^{-2\epsilon^2 N} \;\Rightarrow\; \epsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$$

Example

Chebyshev: For probability at least $1-\delta$, we have

$$\mu - \frac{\sigma}{\sqrt{\delta N}} \le \nu \le \mu + \frac{\sigma}{\sqrt{\delta N}}.$$

Hoeffding: For probability at least $1-\delta$, we have

$$\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \le \nu \le \mu + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}.$$

Example:

- Alex: I have data $X_1, \ldots, X_N$. I want to estimate μ. How many data points N do I need?
- Bob: How much δ can you tolerate?
- Alex: Alright. I only have limited number of data points. How good my estimate is? (ϵ)
- Bob: How many data points N do you have?

Example

Chebyshev: For probability at least $1-\delta$, we have

$$\mu - \frac{\sigma}{\sqrt{\delta N}} \leq \nu \leq \mu + \frac{\sigma}{\sqrt{\delta N}}.$$

Hoeffding: For probability at least $1-\delta$, we have

$$\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \leq \nu \leq \mu + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}.$$

Let $\delta = 0.01$, $N = 10000$, $\sigma = 1$.

$$\epsilon = \frac{\sigma}{\sqrt{\delta N}} = 0.1$$

$$\epsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} = 0.016$$

Let $\delta = 0.01$, $\epsilon = 0.01$, $\sigma = 1$.

$$N \geq \frac{\sigma^2}{\epsilon^2\delta} = 1,000,000.$$

$$N \geq \frac{\log\frac{2}{\delta}}{2\epsilon^2} \approx 26,500.$$

Reading List

- Abu-Mustafa, Learning from Data, Chapter 2.
- Martin Wainwright, High Dimensional Statistics, Cambridge University Press 2019. (Chapter 2)
- Cornell Note,
  https://www.cs.cornell.edu/~sridharan/concentration.pdf
- CMU Note,
  http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf
- Stanford Note,
  http://cs229.stanford.edu/extra-notes/hoeffding.pdf