Concentration Inequalities

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

November 7, 2022

Recap

Last lecture: Careful mathematical reasoning to **prove** that

Recap

Last lecture: Careful mathematical reasoning to **prove** that

- Loss at deployment converged to **expected loss** as $N \to \infty$

Recap

Last lecture: Careful mathematical reasoning to **prove** that

- Loss at deployment converged to **expected loss** as $N \to \infty$
- Test set loss converged to **expected loss** as $N \to \infty$

Recap

Last lecture: Careful mathematical reasoning to **prove** that

- Loss at deployment converged to **expected loss** as $N \to \infty$
- Test set loss converged to **expected loss** as $N \to \infty$
- Variance of test set loss scaled as $\frac{c}{N}$

Recap

Last lecture: Careful mathematical reasoning to **prove** that

- Loss at deployment converged to **expected loss** as $N \to \infty$
- Test set loss converged to **expected loss** as $N \to \infty$
- Variance of test set loss scaled as $\frac{c}{N}$

Cornerstone of the argument was a **theorem**: Weak LLN:

$$\mathbb{P}(|X_n - \mu| < \epsilon) = 1 \qquad \text{for } X_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \{X_n\}\text{iid}\,, \mu = \mathbb{E}[X_n] \tag{1}$$

Recap

Last lecture: Careful mathematical reasoning to **prove** that

- Loss at deployment converged to **expected loss** as $N \to \infty$
- Test set loss converged to **expected loss** as $N \to \infty$
- Variance of test set loss scaled as $\frac{c}{N}$

Cornerstone of the argument was a **theorem**: Weak LLN:

$$\mathbb{P}(|X_n - \mu| < \epsilon) = 1 \qquad \text{for } X_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \{X_n\}\text{iid}\,, \mu = \mathbb{E}[X_n] \tag{1}$$

- Doesn't say anything about the **accuracy** for finite N!

Recap

Last lecture: Careful mathematical reasoning to **prove** that

- Loss at deployment converged to **expected loss** as $N \to \infty$
- Test set loss converged to **expected loss** as $N \to \infty$
- Variance of test set loss scaled as $\frac{c}{N}$

Cornerstone of the argument was a **theorem**: Weak LLN:

$$\mathbb{P}(|X_n - \mu| < \epsilon) = 1 \qquad \text{for } X_n = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \{X_n\}\text{iid} \,, \mu = \mathbb{E}[X_n] \tag{1}$$

- Doesn't say anything about the **accuracy** for finite N!
- Intuitively, low variance $\implies$ unlikely to be far from mean.

Recap

Last lecture: Careful mathematical reasoning to **prove** that

- Loss at deployment converged to **expected loss** as $N \to \infty$
- Test set loss converged to **expected loss** as $N \to \infty$
- Variance of test set loss scaled as $\frac{c}{N}$

Cornerstone of the argument was a **theorem**: Weak LLN:

$$\mathbb{P}(|X_n - \mu| < \epsilon) = 1 \qquad \text{for } X_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \{X_n\}\text{iid}\,, \mu = \mathbb{E}[X_n] \tag{1}$$

- Doesn't say anything about the **accuracy** for finite N!
- Intuitively, low variance $\implies$ unlikely to be far from mean.
- Can we use this? Can we make this **precise**?

Concentration Inequalities

- Theorems are useful because they are **black boxes**

Concentration Inequalities

- Theorems are useful because they are **black boxes**
- Abstract away details of a complex argument,
  to give you simple answers

Concentration Inequalities

- Theorems are useful because they are **black boxes**
- Abstract away details of a complex argument,
  to give you simple answers
- Today: We break open the black box of the LLN (i.e. the proof)

Concentration Inequalities

- Theorems are useful because they are **black boxes**
- Abstract away details of a complex argument,
  to give you simple answers
- Today: We break open the black box of the LLN (i.e. the proof)
- We find tools that will help us answer questions about finite N!

Concentration Inequalities

- Theorems are useful because they are **black boxes**
- Abstract away details of a complex argument,
  to give you simple answers
- Today: We break open the black box of the LLN (i.e. the proof)
- We find tools that will help us answer questions about finite N!

Questions:

1. How accurate is our estimate of the expected loss?
2. How big should our test set be, to get a certain accuracy?

Overview

Intro: Proof of Weak Law of Large Numbers

Generalisation Error Bounds

Assumptions and Quality of Bounds

Conclusion

Weak Law of Large Numbers

- For a sequence of iid RVs $X_1, X_2, X_3, \ldots, X_N$
- with mean $\mu = \mathbb{E}[X]$
- we can define a new RV $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$
- for which will hold:

$$\lim_{N\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| < \epsilon\right) = 1 \tag{2}$$

Weak Law of Large Numbers

- For a sequence of iid RVs $X_1, X_2, X_3, \ldots, X_N$
- with mean $\mu = \mathbb{E}[X]$
- we can define a new RV $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$
- for which will hold:

$$\lim_{N\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| < \epsilon\right) = 1 \tag{2}$$

- How to prove this?

Weak Law of Large Numbers

- For a sequence of iid RVs $X_1, X_2, X_3, \ldots, X_N$
- with mean $\mu = \mathbb{E}[X]$
- we can define a new RV $\overline{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n$
- for which will hold:

$$\lim_{N \to \infty} \mathbb{P}\left(|\overline{X}_n - \mu| < \epsilon\right) = 1 \tag{2}$$

- How to prove this?
- Let's understand how far samples lie from the mean.

Weak Law of Large Numbers

- For a sequence of iid RVs $X_1, X_2, X_3, \ldots, X_N$
- with mean $\mu = \mathbb{E}[X]$
- we can define a new RV $\overline{X}_N \; = \frac{1}{N}\sum_{n=1}^{N} X_n$
- for which will hold:

$$\lim_{N\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| < \epsilon\right) = 1 \tag{2}$$

- How to prove this?
- Let's understand how far samples lie from the mean.
- For positive RVs, since $|\overline{X}_n - \mu| \geqslant 0$!

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Proof:

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Proof:

$$\mathbb{E}[X] = \int_0^\infty x p_X(x) \mathrm{d}x \tag{4}$$

(5)

(6)

(7)

(8)

(9)

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Proof:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x p_X(x) \mathrm{d}x && (4)\\
&= \int_0^a x p_X(x) \mathrm{d}x + \int_a^\infty x p_X(x) \mathrm{d}x && (5)\\
& && (6)\\
& && (7)\\
& && (8)\\
& && (9)
\end{aligned}$$

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Proof:

$$\mathbb{E}[X] = \int_0^\infty x p_X(x) \mathrm{d}x \tag{4}$$

$$= \int_0^a x p_X(x) \mathrm{d}x + \int_a^\infty x p_X(x) \mathrm{d}x \tag{5}$$

$$\geqslant \int_a^\infty x p_X(x) \mathrm{d}x \tag{6}$$

(7)

(8)

(9)

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Proof:

$$\mathbb{E}[X] = \int_0^\infty x p_X(x) \mathrm{d}x \tag{4}$$

$$= \int_0^a x p_X(x) \mathrm{d}x + \int_a^\infty x p_X(x) \mathrm{d}x \tag{5}$$

$$\geqslant \int_a^\infty x p_X(x) \mathrm{d}x \tag{6}$$

$$\geqslant \int_a^\infty a p_X(x) \mathrm{d}x \tag{7}$$

(8)

(9)

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Proof:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x p_X(x) \mathrm{d}x && (4)\\
&= \int_0^a x p_X(x) \mathrm{d}x + \int_a^\infty x p_X(x) \mathrm{d}x && (5)\\
&\geqslant \int_a^\infty x p_X(x) \mathrm{d}x && (6)\\
&\geqslant \int_a^\infty a p_X(x) \mathrm{d}x && (7)\\
&= aP(X \geqslant a) && (8)\\
& && (9)
\end{aligned}$$

Markov's inequality

For a RV $X > 0$, and $a > 0$, then

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \tag{3}$$

Proof:

$$\mathbb{E}[X] = \int_0^\infty x p_X(x) \mathrm{d}x \tag{4}$$

$$= \int_0^a x p_X(x) \mathrm{d}x + \int_a^\infty x p_X(x) \mathrm{d}x \tag{5}$$

$$\geqslant \int_a^\infty x p_X(x) \mathrm{d}x \tag{6}$$

$$\geqslant \int_a^\infty a p_X(x) \mathrm{d}x \tag{7}$$

$$= a P(X \geqslant a) \tag{8}$$

$$\implies P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \qquad \text{Done.} \tag{9}$$

Markov's inequality

For positive RVs (like deviations) with finite means:

Markov's inequality

For positive RVs (like deviations) with finite means:

- Large values are increasingly unlikely! ($\propto \frac{1}{a}$)

Markov's inequality

For positive RVs (like deviations) with finite means:

- Large values are increasingly unlikely! ($\propto \frac{1}{a}$)
- The expectation determines how large values can be

Markov's inequality

For positive RVs (like deviations) with finite means:

- Large values are increasingly unlikely! ($\propto \frac{1}{a}$)
- The expectation determines how large values can be

Such bounds are powerful because they abstract away details of the distribution, which we may not know!

Markov's inequality

For positive RVs (like deviations) with finite means:

- Large values are increasingly unlikely! ($\propto \frac{1}{a}$)
- The expectation determines how large values can be

Such bounds are powerful because they abstract away details of the distribution, which we may not know!

Chebyshev's Inequality

For a RV X, with finite exp $X = \mu$, and finite $\mathbb{V}[X] = \sigma^2$, then for $k > 0$

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2} \tag{10}$$

Chebyshev's Inequality

For a RV X, with finite exp $X = \mu$, and finite $\mathbb{V}[X] = \sigma^2$, then for $k > 0$

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2} \tag{10}$$

Proof: Apply Markov's inequality to the RV of the squared deviation:

Chebyshev's Inequality

For a RV X, with finite exp $X = \mu$, and finite $\mathbb{V}[X] = \sigma^2$, then for $k > 0$

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2} \tag{10}$$

Proof: Apply Markov's inequality to the RV of the squared deviation:

$$P\big((X - \mu)^2 \geqslant a\big) \leqslant \frac{\mathbb{E}\big[(X - \mu)^2\big]}{a} \tag{11}$$

$$= \frac{\sigma^2}{a} \tag{12}$$

(13)

(14)

Chebyshev's Inequality

For a RV X, with finite exp $X = \mu$, and finite $\mathbb{V}[X] = \sigma^2$, then for $k > 0$

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2} \tag{10}$$

Proof: Apply Markov's inequality to the RV of the squared deviation:

$$P\big((X-\mu)^2 \geqslant a\big) \leqslant \frac{\mathbb{E}\big[(X-\mu)^2\big]}{a} \tag{11}$$

$$= \frac{\sigma^2}{a} \tag{12}$$

$$\implies P\big((X-\mu)^2 \geqslant k^2\sigma^2\big) \leqslant \frac{1}{k^2} \qquad \text{sub } a = k^2\sigma^2 \tag{13}$$

$$\tag{14}$$

Chebyshev's Inequality

For a RV X, with finite exp $X = \mu$, and finite $\mathbb{V}[X] = \sigma^2$, then for $k > 0$

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2} \tag{10}$$

Proof: Apply Markov's inequality to the RV of the squared deviation:

$$P\big((X-\mu)^2 \geqslant a\big) \leqslant \frac{\mathbb{E}\big[(X-\mu)^2\big]}{a} \tag{11}$$

$$= \frac{\sigma^2}{a} \tag{12}$$

$$\implies P\big((X-\mu)^2 \geqslant k^2\sigma^2\big) \leqslant \frac{1}{k^2} \qquad \text{sub } a = k^2\sigma^2 \tag{13}$$

$$\implies P(|X-\mu| \geqslant k\sigma) \leqslant \frac{1}{k^2} \qquad \text{Done.} \tag{14}$$

Chebyshev's Inequality

For **any** RV with finite mean and variance, we **limit** the probability of being k standard deviations from the mean.

Weak Law of Large Numbers

Proof of WLLN:

- Remember: $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$

Weak Law of Large Numbers

Proof of WLLN:

- Remember: $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$
- Note that: $\mathbb{V}\left[\overline{X}_n\right] = \frac{\mathbb{V}[X]}{N} = \frac{c}{N}$ (we assume finite variance)

Weak Law of Large Numbers

Proof of WLLN:

- Remember: $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$
- Note that: $\mathbb{V}\left[\overline{X}_n\right] = \frac{\mathbb{V}[X]}{N} = \frac{c}{N}$ (we assume finite variance)
- By Chebyshev:

$$P(|\overline{X}_n - \mathbb{E}[X]| > \epsilon) \leqslant \frac{\sigma^2}{\epsilon^2} \tag{15}$$

$$= \frac{c}{N\epsilon^2} \tag{16}$$

Weak Law of Large Numbers

Proof of WLLN:

- Remember: $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$
- Note that: $\mathbb{V}\left[\overline{X}_n\right] = \frac{\mathbb{V}[X]}{N} = \frac{c}{N}$ (we assume finite variance)
- By Chebyshev:

$$P(|\overline{X}_n - \mathbb{E}[X]| > \epsilon) \leqslant \frac{\sigma^2}{\epsilon^2} \quad (15)$$

$$= \frac{c}{N\epsilon^2} \quad (16)$$

- For **any** fixed ϵ, $\lim_{N\to\infty} \frac{c}{N\epsilon^2} = 0$

Weak Law of Large Numbers

Proof of WLLN:

- Remember: $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$
- Note that: $\mathbb{V}\left[\overline{X}_n\right] = \frac{\mathbb{V}[X]}{N} = \frac{c}{N}$ (we assume finite variance)
- By Chebyshev:

$$
\begin{aligned}
P(|\overline{X}_n - \mathbb{E}[X]| > \epsilon) &\leqslant \frac{\sigma^2}{\epsilon^2} &\quad (15)\\
&= \frac{c}{N\epsilon^2} &\quad (16)
\end{aligned}
$$

- For **any** fixed ϵ, $\lim_{N\to\infty} \frac{c}{N\epsilon^2} = 0$
- $\implies \lim_{N\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| < \epsilon\right) = 1$ Done.

Overview

LLN is a Detour

- LLN ignores the size of the variance

LLN is a Detour

- LLN ignores the size of the variance
- To prove LLN, we used a bound that **did** depend on the size of the variance!

LLN is a Detour

- LLN ignores the size of the variance
- To prove LLN, we used a bound that **did** depend on the size of the variance!

Can we use knowledge of the size of the variance
to say something more about generalisation error?

Generalisation Error Bound

A **Generalisation Error/Loss Bound** is a procedure for computing a number ϵ from data that you sample form the world, such that

Generalisation Error Bound

A **Generalisation Error/Loss Bound** is a procedure for computing a number ϵ from data that you sample form the world, such that

- with high probability,

Generalisation Error Bound

A **Generalisation Error/Loss Bound** is a procedure for computing a number ϵ from data that you sample form the world, such that

- with high probability,
- the expected loss is below ϵ.

Generalisation Error Bound

A **Generalisation Error/Loss Bound** is a procedure for computing a number ϵ from data that you sample form the world, such that

- with high probability,
- the expected loss is below ϵ.

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \delta \tag{17}$$

$$\text{ER} = \mathbb{E}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] \tag{18}$$

Classification GEB

- Consider Classification where $f : \mathcal{X} \to [0, 1]$.

Classification GEB

- Consider Classification where $f : \mathcal{X} \to [0, 1]$.
- For **testing**, we use 0-1 loss function (classification accuracy)

$$\ell(f(x; \boldsymbol{\theta}^*), y) = \begin{cases} 0 & \text{if } \texttt{int}(f(x; \boldsymbol{\theta}^*)) = y \\ 1 & \text{otherwise} \end{cases} \tag{19}$$

Classification GEB

- Consider Classification where $f : \mathcal{X} \to [0, 1]$.
- For **testing**, we use 0-1 loss function (classification accuracy)

$$\ell(f(x; \boldsymbol{\theta}^*), y) = \begin{cases} 0 & \text{if } \texttt{int}(f(x; \boldsymbol{\theta}^*)) = y \\ 1 & \text{otherwise} \end{cases} \tag{19}$$

- Remember $L_{\text{test}} = \frac{1}{N} \sum_{n=1}^{N} \ell(f(x; \boldsymbol{\theta}^*), y)$
- Remember $\mathbb{E}_{\pi(x,y)}[L_{test}] = \text{ER}$
  ($x = [x_1, x_2, \ldots]$, and $y = [y_1, y_2, \ldots]$).

Chebyshev GEB

Apply Chebyshev:

Chebyshev GEB

Apply Chebyshev:

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \frac{\sigma^2}{\epsilon^2} \tag{20}$$

$$\tag{22}$$

Chebyshev GEB

Apply Chebyshev:

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \frac{\sigma^2}{\epsilon^2} \tag{20}$$

$$\sigma^2 = \mathbb{V}_{\pi(x,y)}[L_{\text{test}}] \tag{21}$$

$$= \frac{1}{N}\mathbb{V}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] \tag{22}$$

Chebyshev GEB

Apply Chebyshev:

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \frac{\sigma^2}{\epsilon^2} \tag{20}$$

$$\sigma^2 = \mathbb{V}_{\pi(x,y)}[L_{\text{test}}] \tag{21}$$

$$= \frac{1}{N}\mathbb{V}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] \tag{22}$$

Notice: $\mathbb{V}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] < 0.25$!

Chebyshev GEB

Apply Chebyshev:

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \frac{\sigma^2}{\epsilon^2} \tag{20}$$

$$\sigma^2 = \mathbb{V}_{\pi(x,y)}[L_{\text{test}}] \tag{21}$$

$$= \frac{1}{N}\mathbb{V}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] \tag{22}$$

Notice: $\mathbb{V}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] < 0.25$!

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \frac{0.25}{N\epsilon^2} \tag{23}$$

Chebyshev GEB

Apply Chebyshev:

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \frac{\sigma^2}{\epsilon^2} \tag{20}$$

$$\sigma^2 = \mathbb{V}_{\pi(x,y)}[L_{\text{test}}] \tag{21}$$

$$= \frac{1}{N}\mathbb{V}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] \tag{22}$$

Notice: $\mathbb{V}_{\pi(x,y)}[\ell(f(x;\boldsymbol{\theta}^*), y)] < 0.25$!

$$\mathbb{P}(|L_{\text{test}} - \text{ER}| > \epsilon) < \frac{0.25}{N\epsilon^2} \tag{23}$$

$$\implies \mathbb{P}(\text{ER} > L_{\text{test}} + \epsilon) < \frac{0.25}{N\epsilon^2} \tag{24}$$

(Draw double-sided plot on board. L_{test} is RV, and we only care about under-estimation of ER.)

Example Chebyshev GEB

Q1: How accurate is our estimate of the expected loss?

- You train a NN on MNIST

Example Chebyshev GEB

Q1: How accurate is our estimate of the expected loss?

- You train a NN on MNIST
- Test error with $N = 10000$ gives $L_{\text{test}} = 0.01$

Example Chebyshev GEB

Q1: How accurate is our estimate of the expected loss?

- You train a NN on MNIST
- Test error with $N = 10000$ gives $L_{\text{test}} = 0.01$
- Then Chebyshev gives us the guarantee that

Example Chebyshev GEB

Q1: How accurate is our estimate of the expected loss?

- You train a NN on MNIST
- Test error with $N = 10000$ gives $L_{\text{test}} = 0.01$
- Then Chebyshev gives us the guarantee that

$$\mathbb{P}(\text{ER} > L_{\text{test}} + 0.03) < \frac{0.25}{N \cdot 0.03^2} = 0.0278 \quad \text{Pretty confident} \tag{25}$$

$$\mathbb{P}(\text{ER} > L_{\text{test}} + 0.01) < \frac{0.25}{N \cdot 0.01^2} = 0.25 \quad \text{Not confident} \tag{26}$$

$$\mathbb{P}(\text{ER} > L_{\text{test}} + 0.001) < \frac{0.25}{N \cdot 0.001^2} = 25 \quad \textbf{Vacuous} \tag{27}$$

How good is this?

- We can guarantee with high probability that the classifier isn't an order of magnitude worse than L_{test} indicates

How good is this?

- We can guarantee with high probability that the classifier isn't an order of magnitude worse than L_{test} indicates
- However bound is not tight enough to distinguish different methods, which often differ in accuracy by ± 0.001

How good is this?

- We can guarantee with high probability that the classifier isn't an order of magnitude worse than L_{test} indicates
- However bound is not tight enough to distinguish different methods, which often differ in accuracy by ± 0.001
- Probably **very** pessimistic

How good is this?

- We can guarantee with high probability that the classifier isn't an order of magnitude worse than L_{test} indicates
- However bound is not tight enough to distinguish different methods, which often differ in accuracy by ± 0.001
- Probably **very** pessimistic
- Bound holds for **any** distribution with a maximum variance!

Flipping bound round

Q2: How big should our test set be, to get a certain accuracy?

$$\mathbb{P}(\mathrm{ER} > L_{\text{test}} + \epsilon) < \delta \tag{28}$$

$$\implies N > \frac{0.25}{\delta\epsilon^2} \tag{29}$$

Flipping bound round

Q2: How big should our test set be, to get a certain accuracy?

$$\mathbb{P}(\mathrm{ER} > L_{\text{test}} + \epsilon) < \delta \tag{28}$$

$$\implies N > \frac{0.25}{\delta\epsilon^2} \tag{29}$$

- For $\epsilon = 0.001$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.05$, we need $N > 5 \cdot 10^6$!
- For $\epsilon = 0.001$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.01$, we need $N > 25 \cdot 10^6$!
- For $\epsilon = 0.01$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.05$, we need $N > 50 \cdot 10^3$!
- For $\epsilon = 0.01$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.01$, we need $N > 250 \cdot 10^3$!

Overview

Intro: Proof of Weak Law of Large Numbers

Generalisation Error Bounds

Assumptions and Quality of Bounds

Conclusion

Hoeffding's inequality

For iid RVs $X_1, X_2, \ldots,$ such that $a < X_n < b$, $S_N = \frac{1}{N}\sum_n X_n$, and $t > 0$, we have

$$\mathbb{P}(|S_N - \mathbb{E}_\pi[S_N]| \geqslant t) \leqslant 2\exp\left(-\frac{2t^2N}{(b-a)^2}\right) \tag{30}$$

Proof not covered in course :)

Hoeffding GEB

Again, for classification

$$\mathbb{P}(\mathrm{ER} > L_{\text{test}} + \epsilon) \leqslant \delta \tag{31}$$

$$\implies N \geqslant \frac{\log(2\delta^{-1})}{2\epsilon^2} \tag{32}$$

Hoeffding GEB

Again, for classification

$$\mathbb{P}(\mathrm{ER} > L_{\text{test}} + \epsilon) \leqslant \delta \tag{31}$$

$$\implies N \geqslant \frac{\log(2\delta^{-1})}{2\epsilon^2} \tag{32}$$

- For $\epsilon = 0.001$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.05$, we need $N > 1.85 \cdot 10^6$!
- For $\epsilon = 0.001$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.01$, we need $N > 2.65 \cdot 10^6$!
- For $\epsilon = 0.01$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.05$, we need $N > 18.5 \cdot 10^3$!
- For $\epsilon = 0.01$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.01$, we need $N > 26.5 \cdot 10^3$!

Hoeffding GEB

Again, for classification

$$\mathbb{P}(\mathrm{ER} > L_{\text{test}} + \epsilon) \leqslant \delta \tag{31}$$

$$\implies N \geqslant \frac{\log(2\delta^{-1})}{2\epsilon^2} \tag{32}$$

- For $\epsilon = 0.001$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.05$, we need $N > 1.85 \cdot 10^6$!
- For $\epsilon = 0.001$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.01$, we need $N > 2.65 \cdot 10^6$!
- For $\epsilon = 0.01$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.05$, we need $N > 18.5 \cdot 10^3$!
- For $\epsilon = 0.01$, and $\delta = \frac{0.25}{N\epsilon^2} < 0.01$, we need $N > 26.5 \cdot 10^3$!

Significant reduction compared to Chebyshev!

Overview

Conclusion