ENSURING MACHINES ARE WELL BEHAVED

TODAY'S CLASS

GOALS

1. Constraining Models and Providing Guarantees
2. Confidence Intervals
3. Approaches to find models that guarantee:
   - Bias and Fairness (balancing accuracy/outcomes for protected groups)
   - Performance (overall accuracy/performance/money)
   - Minimize adverse outcomes

QUIZ

THREE LAWS OF ROBOTICS

GOALS

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

THREE LAWS OF ROBOTICS

GOALS

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey ch orders would conflict with the F
3. A robot must prot not conflict with the First or Second L

What if we cannot guarantee these laws?

ENSURING INTELLIGENT MACHINES ARE WELL-BEHAVED

Overview and guide to one approach on constraining model/agent behavior with gaurantees.

https://aisafety.cs.umass.edu/

Paper in Science: https://www.science.org/doi/10.1126/science.aag3311

Science papers are readable and understandable by general audiance.

ORIGINAL OBJECTIVE

FINDING AN APPROXIMATION

Constraining the error:

$$\forall x \in \mathcal{X}, |f(x) - f_*(x)| \leq \epsilon$$

AVERAGE ERROR

FINDING AN APPROXIMATION

Objective function:

$$l(\theta) = \mathbf{E}\left[\left(f(X, \theta) - Y\right)^2\right]$$

AVERAGE ERROR

FINDING AN APPROXIMATION

Objective function:

$$l(\theta) = \mathbf{E}\left[\left(f(X,\theta) - Y\right)^2\right]$$

Can we guarantee

$$l(\theta) \leq \epsilon$$

EVALUATING $l(\theta)$

Evaluation of $l(\theta)$

$$\widehat{\theta}* \leftarrow \arg\min_{\theta} l_{D_{train}}(\theta)$$

$$l(\theta) \approx l_{D_{test}}\left(\widehat{\theta}*\right)$$

EVALUATING $l(\theta)$

Evaluation of $l(\theta)$

$$\hat{\theta}^* \leftarrow \arg\min_{\theta} l_{D_{train}}(\theta)$$

$$l(\theta) \approx l_{D_{test}}\left(\hat{\theta}^*\right)$$

Need infinite data to have an accurate evaluation

$$\lim_{|D_{test}| \to \infty} | l_{D_{test}}(\theta) - l(\theta) | \to 0$$

CONSTRAINING LOSS

If $l_{D_{test}}\left(\widehat{\theta}*\right) < \epsilon$ is $l\left(\widehat{\theta}*\right) < \epsilon$?

CONSTRAINING LOSS

If $l_{D_{test}}\left(\widehat{\theta}*\right) < \epsilon$ is $l\left(\widehat{\theta}*\right) < \epsilon$?

Not necessarily:

Due to noise $l_{D_{test}}\left(\widehat{\theta}*\right) < \epsilon$, but $l\left(\widehat{\theta}*\right) > \epsilon$ or vice versa

We also may not be able to find a $\widehat{\theta}*$ such that $l\left(\widehat{\theta}*\right) < \epsilon$

UPPER BOUNDING LOSS

Idea: Find an upper-bound estimate $l_{upper}(\theta)$ on $l(\theta)$

$$l(\theta) \leq l_{upper}(\theta)$$

If $l_{upper}(\theta) < \epsilon$ then $l(\theta) < \epsilon$

UPPER BOUNDING LOSS

n number of samples in D_{test}

Find a function $C : \mathbb{N} \to \mathbb{R}$ such that

$$\forall \theta,\ l(\theta) \leq l_{D_{test}}(\theta) + C(n)$$

UPPER BOUNDING LOSS

n number of samples in D_{test}

Find a function $C : \mathbb{N} \to \mathbb{R}$ such that

$$\forall \theta,\ l(\theta) \leq l_{D_{test}}(\theta) + C(n)$$

$C(n)$ provides a worst-case upper bound on the loss function

Worst-case:

- Any model parameters θ
- Any data D_{test}

UPPER BOUNDING LOSS

n number of samples in D_{test}

Find a function $C : \mathbb{N} \rightarrow \mathbb{R}$ such that

$$\forall \theta,\ l(\theta) \leq l_{D_{test}}(\theta) + C(n)$$

$C(n)$ provides a worst-case upper bound on the loss function

Worst-case:

- Any model parameters θ
- Any data D_{test}

Problems with this approach?

UPPER BOUNDING LOSS

Worst-case bounds are usually very conservative:

$l(\theta) \ll l_{D_{test}}(\theta) + C(n)$ – upper bound is much larger than $l(\theta)$

We will often say we cannot guarantee θ satisfies $l(\theta) < \epsilon$ even if it does

Reason:

$C(n)$ has to work for both good and bad θ

Has to work for any data distribution D_{test}

PROBABILISTIC CONSTRAINTS

Idea: Guarantee with a high probability that a model satisfies the constraint.

With 99% confidence, we know that $\widehat{\theta}*$ satisfies $l(\widehat{\theta}*) \leq \epsilon$

We trade certainty in the upper bound for a better estimate.

Note: We cannot guarantee that we will find a $\widehat{\theta}*$ that satisfies this guarantee.

PROBABILISTIC CONSTRAINTS

Upper confidence bound: $l_{upper}: \Theta \times \mathbb{D} \to \mathbb{R}$

$\alpha \in (0,1)$ – confidence level

$$\Pr\left(l(\theta) \leq l_{upper}(\theta, D_{test})\right) \geq 1 - \alpha$$

l_{upper} can adapt to θ and D_{test}

α specifies the failure rate of the limit

$\alpha = 0.05$ means we have 95% confidence

l_{upper} is a confidence interval

PROBLEM SETTING

ACCOUNTING FOR UNCERTIANITY

X Random Variable from some unknown distribution F_X

$\theta -$ parameter we care about, e.g., $\theta = \mathbf{E}[X]$

$D_n = X_1, X_2, \ldots, X_n$ sample of n draws of X

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Is $\bar{X} \geq \theta$ or $\bar{X} \leq \theta$?

CONFIDENCE INTERVALS

WHAT ARE THEY?

$l : \mathbb{R}^n \to \mathbb{R}$ – lower confidence bound function

$u : \mathbb{R}^n \to \mathbb{R}$ – upper confidence bound function

$\alpha \in (0,1)$ – confidence level

$$\Pr\left(\theta \in \left[l\left(D_n\right), u\left(D_n\right)\right]\right) \geq 1 - \alpha$$

One-sided Intervals

$$\Pr\left(\theta \leq u\left(D_n\right)\right) \geq 1 - \alpha$$

$$\Pr\left(\theta \geq l\left(D_n\right)\right) \geq 1 - \alpha$$

CONFIDENCE INTERVALS

WHAT THEY ARE NOT

$$\Pr\left(\theta \in \left[l\left(D_n\right), u\left(D_n\right)\right]\right) \geq 1 - \alpha$$

Random Variable

Not a statement that θ falls in between two values

$$\Pr\left(\theta \in \left[0.7, 0.8\right]\right) \geq 1 - \alpha$$

No random variables

CONFIDENCE INTERVALS

WHAT THEY ARE

$$\Pr\left(\theta \in \left[l\left(D_n\right), u\left(D_n\right)\right]\right) \geq 1 - \alpha$$

Random Variable

The probability that values constructed from the random sample will contain the parameter

For at least $100 \times (1 - \alpha)\,\%$ of samples of D_n, $\theta \in \left[l\left(D_n\right), u\left(D_n\right)\right]$

CONFIDENCE INTERVALS

HOW WE USE THEM

Compare the parameter to a constant, e.g., are heads more likely than tails?

$$X \in \{0,1\}, p = \Pr(X = 1)$$

$$\Pr\left(p \geq l\left(D_n\right)\right) \geq 1 - \alpha$$

If $l\left(D_n\right) > 0.5$ then with confidence $1 - \alpha$, heads are more likely than tails

CONFIDENCE INTERVALS

HOW WE USE THEM

Comparing means of X and Y

$$\Pr\left(\mathbf{E}\left[X\right] \geq l\left(D_n^X\right)\right) \geq 1 - \frac{\alpha}{2}$$

$$\Pr\left(\mathbf{E}[Y] \leq u\left(D_n^Y\right)\right) \geq 1 - \frac{\alpha}{2}$$

If $l\left(D_n^X\right) > u\left(D_n^Y\right)$, then with confidence $1 - \alpha$, $\mathbf{E}[X] > \mathbf{E}[Y]$

CONFIDENCE INTERVALS

HOW WE USE THEM

Comparing means of X and Y

$$\Pr\left(\mathbf{E}\left[X\right] \geq l\left(D_n^X\right)\right) \geq 1 - \frac{\alpha}{2}$$

$$\Pr\left(\mathbf{E}[Y] \leq u\left(D_n^Y\right)\right) \geq 1 - \frac{\alpha}{2}$$

Reduce the failure rate so that both hold with the target rate α

If $l\left(D_n^X\right) > u\left(D_n^Y\right)$, then with confidence $1 - \alpha$, $\mathbf{E}[X] > \mathbf{E}[Y]$

BOOLES INEQUALITY

CORRECTING FOR MULTIPLE COMPARISONS AND COMBINING INTERVALS

Events $A_1, A_2, A_3, \ldots$

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \Pr\left(A_i\right)$$

https://en.wikipedia.org/wiki/Boole%27s_inequality

BOOLES INEQUALITY

CORRECTING FOR MULTIPLE COMPARISONS AND COMBINING INTERVALS

Let A_i be the event that a confidence interval with confidence level α_i fails.

$$\Pr\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \Pr\left(A_i\right) = \sum_{i=1}^{k} \alpha_i$$

The probability that no confidence interval fails

$$1 - \Pr\left(\bigcup_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} \alpha_i$$

$\alpha_i = \dfrac{1}{k}$ works, but we can distribute the uncertainty any way we want

TWO-SIDED INTERVAL

TWO ONE-SIDED INTERVALS

If $\Pr\left(\theta \leq u\left(D_n\right)\right) \geq 1 - \alpha/2$, and $\Pr\left(\theta \geq l\left(D_n\right)\right) \geq 1 - \alpha/2$, then

$$\Pr\left(\theta \in \left[l\left(D_n\right), u\left(D_n\right)\right]\right) \geq 1 - \alpha$$

CI FOR THE MEAN

T-TEST

Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample variance: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

If X is normally distributed:

$$\Pr\left(\mathbf{E}[X] \in \left[\bar{X} + t_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right]\right) = 1 - \alpha$$

$t_{v,\alpha}$ is the α quantile of Student's t-distribution with $n-1$ degrees of freedom

$T = \frac{\bar{X} - \mathbf{E}[X]}{\hat{\sigma}/\sqrt{n}}$ random variable described by Student's t-distribution

CI FOR THE MEAN

T-TEST

Central Limit Theorem:

For a large number of i.i.d. random variables, $X_1, X_2, \ldots, X_n$, with finite variance, $\bar{X}$ has approximately a normal distribution, no matter the distribution of X_i

$$\lim_{n\to\infty} \Pr\left(\mathbf{E}[X] \in \left[\bar{X} + t_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right]\right) \geq 1 - \alpha$$

CI FOR THE MEAN

HOEFFDINGS INEQUALITY

$X_1, X_2, \ldots, X_n$ be *independent* random variables such that $X_i \in [a, b]$

$$l\left(D_n\right) = \bar{X} - (b-a)\sqrt{\frac{\ln(2/\alpha)}{2n}}$$

$$u\left(D_n\right) = \bar{X} + (b-a)\sqrt{\frac{\ln(2/\alpha)}{2n}}$$

Valid for all distributions and sample sizes $n \geq 1$

Does not need i.i.d. data

Very loose intervals, probably need 1,000 samples to compare to random variables.

UPPER CONFIDENCE INTERVAL FOR LOSS

BASED ON THE T-TEST

$$D_{test} = (x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

$$l_{D_{test}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i, \theta)$$

$$\hat{\sigma}^2_{D_{test}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(l(x_i, y_i, \theta) - l_{D_{test}}(\theta) \right)^2$$

$$l_{upper}(\widehat{\theta}*, D_{test}) = l_{D_{test}}(\widehat{\theta}*) + t_{n-1,1-\alpha} \frac{\hat{\sigma}_{D_{test}}}{\sqrt{n}}$$

PROBABILISTIC CONSTRAINTS

PROCESS

Find $\widehat{\theta}*$ using D_{train}

Test for the constraint

If $l_{upper}(\widehat{\theta}*, D_{test}) \leq \epsilon$

Return $\widehat{\theta}*$

Else

?

PROBABILISTIC CONSTRAINTS

PROCESS

Find $\widehat{\theta}*$ using D_{train}

Test for the constraint

If $l_{upper}(\widehat{\theta}*, D_{test}) \leq \epsilon$

Return $\widehat{\theta}*$

Else

Return No Solution Found

PROBABILISTIC CONSTRAINTS

PROCESS

Find $\widehat{\theta}*$ using D_{train}

Test for the constraint

If $l_{upper}(\widehat{\theta}*, D_{test}) \leq \epsilon$

Return $\widehat{\theta}*$

Else

Return No Solution Found

Once we use it, D_{test} we cannot reuse it or we will not have a guarantee anymore.

MUST collect new data.

SELDONIAN MACHINE LEARNING

PROCESS

Search algorithm alg, e.g., $\widehat{\theta}* \leftarrow \text{alg}(D_{train})$

Constraint function $g : \Theta \rightarrow \mathbb{R}$, $g(\widehat{\theta}*) = l(\widehat{\theta}*) - \epsilon$

confidence level α

Find algorithm alg

$$\arg\max_{\text{alg}} f(\text{alg})$$

$$\text{s.t., } \Pr\left(g(\text{alg}(D)) \leq 0\right) \geq 1 - \alpha$$

SELDONIAN MACHINE LEARNING

PROCESS

General Process:

Split data D into D_{train}, D_{test}

Find candidate $\theta_{candidate}$ using D_{train}

Test candidate using upper confidence bound on g

If: $g(\theta_{candidate}, D_{test}) \leq 0$

Return $\theta_{candidate}$

Else:

Return No Solution Found

SELDONIAN MACHINE LEARNING

PROCESS

General Process:

Split data D into D_{train}, D_{test}

Find candidate $\theta_{candidate}$ using D_{train}

Test candidate using upper confidence bound on g

If: $g(\theta_{candidate}, D_{test}) \leq 0$

Return $\theta_{candidate}$

Else:

Return No Solution Found

Guarantees that if solutions return fail the constraint at most $100 \times \alpha$ % of the time.

SELDONIAN MACHINE LEARNING

PROCESS

See https://aisafety.cs.umass.edu/ for tutorials and code to implement these methods

NEXT CLASS

Presentations

- Everyone is required to attend.