Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 1: Calibration

Sep 11, 2026

Predictions

Batch/ Offline

labeled examples

(x, y) $x \in \mathcal{X}$

features   outcome $y \in \{0,1\}$ ← binary outcomes

distribution over examples

$\mathcal{D}, \ (X, Y)$

predictor $\quad f: \mathcal{X} \to [0,1]$

Useful running example is weather forecasting

Sequential/ Online

predictor   nature

for $t = 1, \dots, T$:

$\xleftarrow{x_t}$

chooses p_t

$\xleftarrow{y_t}$

Question: When can we treat p_t or $f(x) = p$ as if it were a real probability?

Why is this not just "accuracy"?

Definition: $\mathrm{MSE}_D(f) := \underset{D}{\mathbb{E}}\left[(f(x)-y)^2\right]$

$f^*(x) = \underset{D}{\mathbb{P}}[Y=1 \mid X=x]$

f^* is the unique minimizer of $\mathrm{MSE}_D(f)$

Las Vegas: Rains 10% of the time

Predictor $f_{LV}(x) = 0$

$\mathrm{MSE}_{D_{LV}}(f_{LV}) = 0.1$

Portland: Rains 50% of the time

$\mathbb{P}[Y=1 \mid X=\text{cloudy}] = .75$

$\mathbb{P}[Y=1 \mid X=\text{sunny}] = .25$

Predictor $f_{PO}(\text{cloudy}) = .75$

$f_{PO}(\text{sunny}) = .25$

$\mathrm{MSE}_{D_{PO}}(f_{PO}) = 0.1875$

(Average) Bias

Definition: For distribution D and predictor f the average bias is

$$\beta_D(f) = \mathbb{E}_D[f(x) - y]$$

f_{PO} was unbiased $(\beta_D(f) = 0)$

f_{LV} was biased $(\beta_D(f) = -0.1)$

$$f_{unbiased}(x) = f(x) - \beta_D(f)$$

Fact: $MSE_D(f_{unbiased}) = MSE_D(f) - \beta^2$

Why isn't bias enough?

<u>Example:</u> $f'_{PO}(\text{cloudy}) = .25$ $\quad f_{PO}(\text{cloudy}) = .75$

$f'_{PO}(\text{sunny}) = .75$ $\quad f_{PO}(\text{sunny}) = .25$

f, f' are both unbiased, but only one is <u>calibrated</u>

<u>Definition:</u> A predictor f is <u>calibrated</u> for distribution D if for every p in the range of f, $\mathbb{P}_D[f(X)=p] > 0$

$$\mathbb{E}_D[Y \mid f(X)=p] = p$$