Measure Pretraining Bias

PDF (sagemaker-dg.pdf#clarify-measure-data-bias)
Kindle (https://www.amazon.com/dp/B07JVSBS9J)
RSS (amazon-sagemaker-release-notes.rss)

Measuring bias in ML models is a first step to mitigating bias. Each measure of bias corresponds to a different notion of fairness. Even considering simple concepts of fairness leads to many different measures applicable in various contexts. For example, consider fairness with respect to age, and, for simplicity, that middle-aged and rest of the age groups are the two relevant demographics, referred to as *facets*. In the case of an ML model for lending, we may want small business loans to be issued to equal numbers of both demographics. Or, when processing job applicants, we may want to see equal numbers of members of each demographic hired. However, this approach may assume that equal numbers of both age groups apply to these jobs, so we may want to condition on the number that apply. Further, we may want to consider not whether equal numbers apply, but whether we have equal numbers of qualified applicants. Or, we may consider fairness to be an equal acceptance rate of qualified applicants across both age demographics, or, an equal rejection rate of applicants, or both. You might use datasets with different proportions of data on the attributes of interest. This imbalance can conflate the bias measure you choose. The models might be more accurate in classifying one facet than in the other. Thus, you need to choose bias metrics that are conceptually appropriate for the application and the situation.

We use the following notation to discuss the bias metrics. The conceptual model described here is for binary classification, where events are labeled as having only two possible outcomes in their sample space, referred to as positive (with value 1) and negative (with value 0). This framework is usually extensible to multicategory classification in a straightforward way or to cases involving continuous valued outcomes when needed. In the binary classification case, positive and negative labels are assigned to outcomes recorded in a raw dataset for a favored facet *a* and for a disfavored facet *d*. These labels y are referred to as *observed labels* to distinguish them from the *predicted labels* y' that are assigned by a machine learning model during the training or inferences stages of the ML lifecycle. These labels are used to define probability distributions $P_a(y)$ and $P_d(y)$ for their respective facet outcomes.

- labels:
  - y represents the n observed labels for event outcomes in a training dataset.
  - y' represents the predicted labels for the n observed labels in the dataset by a trained model.
- outcomes:
  - A positive outcome (with value 1) for a sample, such as an application acceptance.
    - $n^{(1)}$ is the number of observed labels for positive outcomes (acceptances).
    - $n'^{(1)}$ is the number of predicted labels for positive outcomes (acceptances).
  - A negative outcome (with value 0) for a sample, such as an application rejection.
    - $n^{(0)}$ is the number of observed labels for negative outcomes (rejections).
    - $n'^{(0)}$ is the number of predicted labels for negative outcomes (rejections).
- facet values:
  - facet *a* – The feature value that defines a demographic that bias favors.
    - n_a is the number of observed labels for the favored facet value: $n_a = n_a^{(1)} + n_a^{(0)}$ the sum of the positive and negative observed labels for the value facet *a*.

- n'_a is the number of predicted labels for the favored facet value: $n'_a = n'^{(1)}_a + n'^{(0)}_a$ the sum of the positive and negative predicted outcome labels for the facet value *a*. Note that $n'_a = n_a$.

- facet *d* – The feature value that defines a demographic that bias disfavors.
  - n_d is the number of observed labels for the disfavored facet value: $n_d = n^{(1)}_d + n^{(0)}_d$ the sum of the positive and negative observed labels for the facet value *d*.
  - n'_d is the number of predicted labels for the disfavored facet value: $n'_d = n'^{(1)}_d + n'^{(0)}_d$ the sum of the positive and negative predicted labels for the facet value *d*. Note that $n'_d = n_d$.

- probability distributions for outcomes of the labeled facet data outcomes:
  - $P_a(y)$ is the probability distribution of the observed labels for facet *a*. For binary labeled data, this distribution is given by the ratio of the number of samples in facet *a* labeled with positive outcomes to the total number, $P_a(y^1) = n_a^{(1)}/ n_a$, and the ratio of the number of samples with negative outcomes to the total number, $P_a(y^0) = n_a^{(0)}/ n_a$.
  - $P_d(y)$ is the probability distribution of the observed labels for facet *d*. For binary labeled data, this distribution is given by the number of samples in facet *d* labeled with positive outcomes to the total number, $P_d(y^1) = n_d^{(1)}/ n_d$, and the ratio of the number of samples with negative outcomes to the total number, $P_d(y^0) = n_d^{(0)}/ n_d$.

Models trained on data biased by demographic disparities might learn and even exacerbate them. To identify bias in the data before expending resources to train models on it, SageMaker Clarify provides data bias metrics that you can compute on raw datasets before training. All of the pretraining metrics are model-agnostic because they do not depend on model outputs and so are valid for any model. The first bias metric examines facet imbalance, but not outcomes. It determines the extent to which the amount of training data is representative across different facets, as desired for the application. The remaining bias metrics compare the distribution of outcome labels in various ways for facets *a* and *d* in the data. The metrics that range over negative values can detect negative bias. The following table contains a cheat sheet for quick guidance and links to the pretraining bias metrics.

Pretraining Bias Metrics

Bias metric	Description	Example question	Interpreting metric values

Pretraining Bias Metrics

Bias metric	Description	Example question	Interpreting metric values
Class Imbalance (CI) (./clarify-bias-metric-class-imbalance.html)	Measures the imbalance in the number of members between different facet values.	Could there be age-based biases due to not having enough data for the demographic outside a middle-aged facet?	Normalized range: [-1,+1] Interpretation: • Positive values indicate the facet *a* has more training samples in the dataset. • Values near zero indicate the facets are balanced in the number of training samples in the dataset. • Negative values indicate the facet *d* has more training samples in the dataset. Range for normalized binary & multicategory facet labels: [-1,+1] Range for continuous labels: (-∞, +∞) Interpretation: • Positive values indicate facet *a* has a higher proportion of positive outcomes. • Values near zero indicate a more equal proportion of positive outcomes between facets. • Negative values indicate facet *d* has a higher proportion of positive outcomes.
Difference in Proportions of Labels (DPL) (./clarify-data-bias-metric-true-label-imbalance.html)	Measures the imbalance of positive outcomes between different facet values.	Could there be age-based biases in ML predictions due to biased labeling of facet values in the data?	

Pretraining Bias Metrics

Bias metric	Description	Example question	Interpreting metric values
Kullback-Leibler Divergence (KL) (./clarify-data-bias-metric-kl-divergence.html)	Measures how much the outcome distributions of different facets diverge from each other entropically.	How different are the distributions for loan application outcomes for different demographic groups?	Range for binary, multicategory, continuous: [0, +∞) Interpretation: • Values near zero indicate the labels are similarly distributed. • Positive values indicate the label distributions diverge, the more positive the larger the divergence.
Jensen-Shannon Divergence (JS) (./clarify-data-bias-metric-jensen-shannon-divergence.html)	Measures how much the outcome distributions of different facets diverge from each other entropically.	How different are the distributions for loan application outcomes for different demographic groups?	Range for binary, multicategory, continuous: [0, +∞) Interpretation: • Values near zero indicate the labels are similarly distributed. • Positive values indicate the label distributions diverge, the more positive the larger the divergence.

Pretraining Bias Metrics

Bias metric	Description	Example question	Interpreting metric values
Lp-norm (LP) (./clarify-data-bias-metric-lp-norm.html)	Measures a p-norm difference between distinct demographic distributions of the outcomes associated with different facets in a dataset.	How different are the distributions for loan application outcomes for different demographics?	Range for binary, multicategory, continuous: [0, +∞) Interpretation: • Values near zero indicate the labels are similarly distributed. • Positive values indicate the label distributions diverge, the more positive the larger the divergence.
Total Variation Distance (TVD) (./clarify-data-bias-metric-total-variation-distance.html)	Measures half of the L_1-norm difference between distinct demographic distributions of the outcomes associated with different facets in a dataset.	How different are the distributions for loan application outcomes for different demographics?	Range for binary, multicategory, and continuous outcomes: [0, +∞) • Values near zero indicates the labels are similarly distributed. • Positive values indicates the label distributions diverge, the more positive the larger the divergence.

Pretraining Bias Metrics

Bias metric	Description	Example question	Interpreting metric values
Kolmogorov-Smirnov (KS) (./clarify-data-bias-metric-kolmogorov-smirnov.html)	Measures maximum divergence between outcomes in distributions for different facets in a dataset.	Which college application outcomes manifest the greatest disparities by demographic group?	Range of KS values for binary, multicategory, and continuous outcomes: [0,+1] • Values near zero indicate the labels were evenly distributed between facets in all outcome categories. • Values near one indicate the labels for one category were all in one facet, so very imbalanced. • Intermittent values indicate relative degrees of maximum label imbalance.
Conditional Demographic Disparity (CDD) (./clarify-data-bias-metric-cddl.html)	Measures the disparity of outcomes between different facets as a whole, but also by subgroups.	Do some groups have a larger proportion of rejections for college admission outcomes than their proportion of acceptances?	Range of CDD: [-1, +1] • Positive values indicate a outcomes where facet *d* is rejected more than accepted. • Near zero indicates no demographic disparity on average. • Negative values indicate a outcomes where facet *a* is rejected more than accepted.

For additional information about bias metrics, see Fairness Measures for Machine Leaning in Finance (https://pages.awscloud.com/rs/112-TZM-766/images/Fairness.Measures.for.Machine.Learning.in.Finance.pdf) .

Topics

- Class Imbalance (CI) (./clarify-bias-metric-class-imbalance.html)
- Difference in Proportions of Labels (DPL) (./clarify-data-bias-metric-true-label-imbalance.html)
- Kullback-Leibler Divergence (KL) (./clarify-data-bias-metric-kl-divergence.html)
- Jensen-Shannon Divergence (JS) (./clarify-data-bias-metric-jensen-shannon-divergence.html)

- Lp-norm (LP) (./clarify-data-bias-metric-lp-norm.html)
- Total Variation Distance (TVD) (./clarify-data-bias-metric-total-variation-distance.html)
- Kolmogorov-Smirnov (KS) (./clarify-data-bias-metric-kolmogorov-smirnov.html)
- Conditional Demographic Disparity (CDD) (./clarify-data-bias-metric-cddl.html)