CIS 7000-005: Debugging Data & Models Fall 2022

Biases — September 1

Prof. Eric Wong

Bias in ML is not always obvious. In this module, we'll go over what bias could be, how it manifests, and the ramifications.

- What is bias?
- When is bias good/bad?
- Where does bias come from?

1 Types of bias

Bias, in its most general form, can be described as "a tendency for or against something". There are a ton of different kinds of biases, from mathematical to social biases. We can loosely categorize ML-adjacent bias into three main categories:

1. Statistical bias
2. Learning bias
3. Data bias

2 Statistical bias

The notion of bias in statistics is perhaps the the cleanest mathematical type of bias, which describes whether a statistical estimator of a quantitative parameter tends to agree with the true underlying value. More formally, let θ be a parameter, and let $\hat{\theta}(X)$ be a statistic that estimates θ from a dataset X. Then, if

$$\mathbb{E}_X[\hat{\theta}(X)] = \theta \tag{1}$$

we say that $\hat{\theta}(X)$ is an *unbiased* estimator for θ. Otherwise, we say that $\hat{\theta}(X)$ is biased and we call $\mathbb{E}_X[\hat{\theta}(X)] - \theta$ the bias of the statistic $\hat{\theta}(X)$.

Review:

- What does statistical bias measure? It measures the quality or accuracy of our numerical estimators.
- Is statistical bias good or bad? Generally bad, since it means that our estimates are not quite as close as we can get to the parameter we're trying to measure.
- Can we fix this? Generally yes, usually we can mathematically adjust our estimated quantities such that the expectation works out to get an unbiased estimator.

3 Learning bias (or the bias-variance tradeoff)

In machine learning, we often refer to a particular form of bias induced by the learning algorithm. A particular learning algorithm makes certain assumptions (e.g. linearity for linear models) which may not match the true underlying model.

Consider a model $f_D(x)$ trained on a dataset D that predicts y for a given input x. We would like to minimize the expected risk, or

$$\mathbb{E}_{D,x,y}[(f_D(x) - y))^2] \tag{2}$$

where the randomness over the dataset is averaged out. The classic notion of bias comes from a tradeoff between bias and variance of this expected risk. For example, for squared error, we can add and subtract the mean prediction over datasets $\mu(x) = \mathbb{E}_D[f_D(x)]$ like so:

$$\begin{aligned}\mathbb{E}_{D,x,y}[(f_D(x) - y))^2] &= \mathbb{E}_{D,x,y}[(f_D(x) - \mu(x) + \mu(x) - y))^2] \\ &= \mathbb{E}_{D,x}[(f_D(x) - \mu(x))^2] + \mathbb{E}_{x,y}[(\mu(x) - y))^2] \\ &= \mathbb{E}_{D,x}[(f_D(x) - \mu(x))^2] + \mathbb{E}_{x,y}[(\mu(x) - y))^2]\end{aligned} \tag{3}$$

where the cross term vanishes by definition of $\mu(x)$. The first term is the *variance* of the prediction over D. In other words, it measures how much the estimated function f_D varies from the average function $\mu(x)$ over random re-draws of the dataset. The second term can decompose again as

$$\begin{aligned}\mathbb{E}_{x,y}[(\mu(x) - y))^2] &= \mathbb{E}_{x,y}[(\mu(x) - \mathbb{E}_y[y|x] + \mathbb{E}_y[y|x] - y))^2] \\ &= \mathbb{E}_{x,y}[(\mu(x) - \mathbb{E}_y[y|x])^2] + \mathbb{E}_{x,y}[(\mathbb{E}_y[y|x] - y))^2]\end{aligned} \tag{4}$$

where again the cross term vanishes. The first term here is referred to as the *bias*, which measures how far the average function $\mu(x)$ varies from the true mean $\mathbb{E}_y[y|x]$. The second term is referred to as the *noise*, which measures how far the true mean $\mathbb{E}_y[y|x]$ varies from actual samples y.

Since there is nothing we can do about noise, typically we look at the tradeoff between the first two terms, the variance and the bias. This is the statistical notion of bias.

Connection to learning theory. Learning bias is analogous to the distance between the hypothesis outputed by a learning algorithm and the true hypothesis in the unrealizable setting for statistical learning. For a given hypothesis class $\mathcal{H}$, let $h_{ERM} = \min_{h \in \mathcal{H}} R(h)$ be the hypothesis outputted by ERM. If $h^* \notin \mathcal{H}$ is the true underlying hypothesis, then the "learning bias" is exactly the square of the expected difference of the risks of h^* and the average ERM classifier $E[h_{erm}]$:

$$\mathbb{E}[(R(h^*) - R(h_{ERM}))^2] \tag{5}$$

Connection to statistical bias. Learning bias can be viewed as a special case of statistical bias, where $\hat{\theta}(X)$ is our model of the world estimated from a dataset X, and θ is the true underlying process. Then, the learning bias is simply the statistical bias of our estimated model from the true process.

As an example, consider the following simple 1D regression data. A classifier with low variance and high bias is a linear classifier, whereas a classifier with high variance and low bias is a neural network, as shown in the following figures:

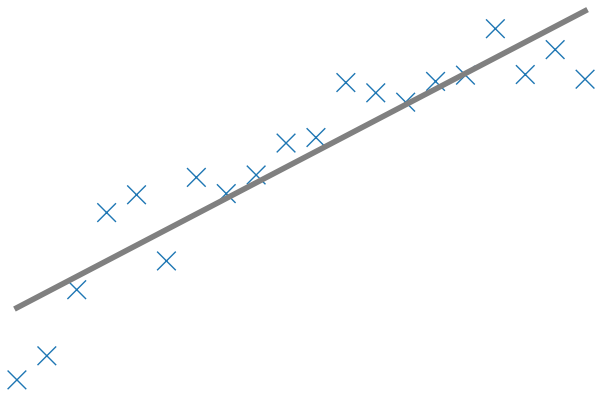

Figure 1: Fitting a linear classifier to data, which has high bias from the hypothesis class and low variance.

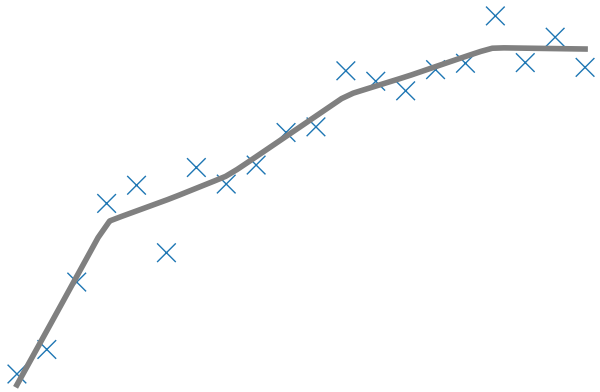

Figure 2: Fitting a neural network to data, which has high bias from the hypothesis class and low variance.

We can estimate all the bias, variance, and noise quantities in synthetic settings and visualize it in the following graphic.

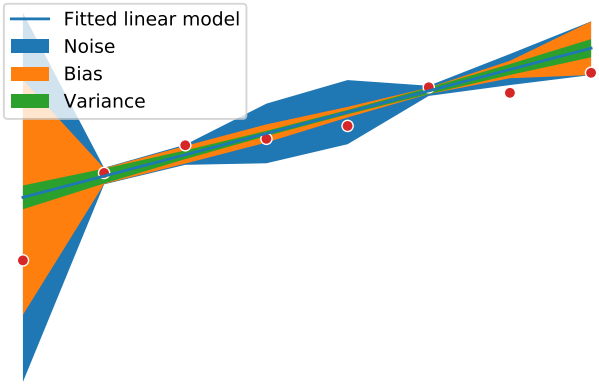

Figure 3: Plotting the bias, variance, and noise terms for an estimate of a linear classifier (quantities are amplified for visibility).

We see that the relative quantities here make sense: in the middle, the noise term dominates because a linear classifier has low variance here and the linear model is relatively accurate in this regime. The bias of the linear model increases greatly towards the left side, where it deviates the most from the underlying log model. The low variance reflects the stability of fitting linear classifiers, with the most variance being on the edges. See the notebook for how these quantities can be calculated.

Review:

- What does learning bias measure? It measures the average error (over datasets) that results in choosing a particular hypothesis class for f_D. In other words, it measures the error that can be attributed to attributed to the learning algorithm being forced to use a particular model and learning algorithm.
- Is learning bias good or bad? It depends. Occam's razor is a learning bias towards simple answers, which implicitly assumes that simpllicity bias is good. If we have underfitting, then learning bias is usually seen as bad, whereas when we have overfitting, learning bias is seen as better.
- Can we fix it? Kind of. If we consider deep learning to be a universal function approximators we can remove learning bias by design, at least the bias induced by the hypothesis class. But deep learning itself has a different set of biases beyond hypothesis complexity that may still lead to a biased model.

4 Bias in data

Biases that occur within datasets are perhaps the most diverse and complex of these biases, many of which remain resolved to this day. What is bias in data?

Judgement based on preconcieved notions or prejudices, as opposed to say the impartial evaluation of facts—Kate Crawford, NeurIPS 2017 Keynote "The Trouble with Bias".

In other words, we can view data bias as a skew in the data. Note that not all skews are bad—indeed, machine learning is all about learning the skews that aid the most in prediction. Thus, typically we are most concerned with biases that cause harms. These harms often occur when the data doesn't match the distribution we are trying to model.

So what are these harms exactly? We can categorize many harmful biases into the following two categories:

- Harms of allocation: these harms affect downstream outcomes for subgroups, such as hiring or credit approval. These tend to be quantifiable to some degree since they are based on predictions with target outcomes (i.e. via fairness metrics).
- Harms of representation: these harms affect the representation for subgroups. Because these are not necessarily explicitly associated with some measurable downstream task, it is more difficult to formalize what exactly this harm is.

Often, harms of allocation can arise from harms of representation. The difference is mainly in whether we can quantify the harm, or if we can just recognize a difference in treatment with unknown harms.

Connection to learning bias. Some data biases can be viewed in the framework of learning bias. Recall that the (squared) learning bias from earlier was

$$\mathbb{E}_{x,y}[(\mu(x) - \mathbb{E}_y[y|x])^2] \text{ where } \mu(x) = \mathbb{E}_D[f_D(x)]. \tag{6}$$

A bias in the data can be framed as a skew between $p(D)$ and $p(x, y)$. In other words, the dataset distribution that we train our models that average out to $\mu(x)$ is different from the target distribution $p(x, y)$. In other words, our datasets are biased and we truly want to estimate $p(x, y)$. Methods for solving or removing these biases can be viewed as trying to correct for this mismatch.

Is it possible to formalize some of these harms? We can certainly try. For example, recognition bias can be viewed as a data bias that affects different subgroups differently, and hence could be framed in the language of fairness or distribution shift. With a model for identifying disparaging cultural content, one could hope to regularized or equalize denigration bias. Stereotyping and under representation bias is not as clear, since there is no "right" answer—how would you go about it?

Review:

- What does data bias measure? It measures skews in the data, typically with a focus on those that produce harms. Skews can vary greatly.
- Is data bias good or bad? It depends. Some skews in the data are good, and are what allow the models to generalize and be useful. However, not all skews are helpful, the data may not match our desired distribution, and we as a society may deem certain skews to be unacceptable regardless of whether they occur naturally.
- Can we fix it? For certain data biases that have quantifiable effects (i.e. fairness), these can be solved for to some degree. However, many data biases lack a well-defined metric for the induced harm, and it becomes less clear how to fix these.

5 References

Parts of these notes are pulled from Cosma Shalizi's course on Modern Regression at

`https://www.stat.cmu.edu/~cshalizi/mreg/15/`

and Kate Crawford's NeurIPS 2017 Keynote, "The Trouble with Bias" at

`https://www.youtube.com/watch?v=fMym_BKWQzk`