Using Permutations Tests to Evaluate Fairness

Cyrus DiCiccio

LinkedIn Corporation

Sriram Vasudevan

LinkedIn Corporation

Kinjal Basu

LinkedIn Corporation

Krishnaram Kenthapadi

Amazon Web Services

Deepak Agarwal

Enter their title

Overview

1 Some Definitions of Fairness

2 Why Evaluating Fairness WIth Respect to these Definitions is Challenging

3 Permutation Tests as a Generally Applicable Solution

Some ways to evaluate fairness

Traditional ML Evaluation Metrics

- AUC
- Classification Error Rates
- Prediction Error

Fairness Definitions

- Equality of Opportunity
- Equalized Odds
- Disparate Impact
- Unawareness
- Demographic parity

Equalized Odds as an Example

A model score used for a classifier should be independent of protected attributes and the true label of an observation

Formally, if members belong to one of two groups, say according to an attribute that takes values 1 or 2, and training labels are denoted by y taking values +1 or -1,

$F(s|y = +1, a = 1) = F(s|y = +1, a = 2)$ and $F(s|y = -1, a = 1) = F(s|y = -1, a = 2)$

where $F(s|y, a)$ is the conditional distribution of the model scores

Challenges With Evaluating Equalized Odds

Equalized odds is a definition rather than a single number summary of fairness

How can we translate this into a test statistic to see if our model is fair? There are many choices, but one is

$$\sup_s \ \sup_y \ |F(s \mid y\,,\, a = 1) \ - \ F(s \mid y\,,\, a \ = \ 2)|$$

Non-Parametric Testing

The test statistic for equalized odds can be difficult to interpret and it can be challenging to derive the distribution of the statistic

Is the value sufficiently large that there is evidence of model bias?

Non-parametric testing (i.e. bootstrap or permutation tests) provide a flexible framework for assessing the strength of evidence that a model is unfair

Our paper discusses permutation tests: common misapplications, and modifications making this methodology appropriate for fairness applications

Brief Review of Permutation Tests

Observe data from two populations:

$$X_1, \ldots, X_{n_x} \sim P_X \qquad \text{and} \qquad Y_1, \ldots, Y_{n_y} \sim P_Y$$

Are the populations the same?

$$H : P_X = P_Y$$

A reasonable test statistic might be

$$T = \bar{X} - \bar{Y}$$

Brief Review of Permutation Tests (Continued)

A p-value is the chance of observing a test statistic at least as "extreme" as the value we actually observed

Permutation test approach:

- Randomly shuffle the population designations of the observations
- Recompute the test statistic T
- Repeat many times

Permutation p-value: the proportion of permuted datasets resulting in a larger test statistic than the original value

This test is exact!

A Fairness Example

Consider testing whether the true positive rate of a classifier is equal between two groups

Test Statistic: difference in proportion of negative labeled observations that are classified as positive between the two groups

Permutation test: Randomly reshuffle group labels, recompute test statistic

Problems With The Permutation Test

The asymptotic distribution (under the null hypothesis) of the test statistic has variance (proportional to)

$$\frac{p_{TP}(1-p_{TP})}{p_1 p_{1,+}} + \frac{p_{TP}(1-p_{TP})}{p_2 p_{2,+}}$$

and of the permutation distribution is

$$\frac{p_{TP}(1-p_{TP})}{p_1 p_{+}} + \frac{p_{TP}(1-p_{TP})}{p_2 p_{+}}$$

where p_{TP} is the common TPR, p_i is the proportion of observations belonging to group i, $p_{i,+}$ is the proportion of observations belonging to group i that have positive labels and

$$p_{+} = p_1 p_{1,+} + p_1 p_{2,+}$$

Why This Happened

The permutation test assumes that the distributions are equal, which is much stronger than just equality of the false positive rates

The permutation test pools data, and approximates the sampling distribution as if each sample was taken from the pooled distribution

This is similar to the difference between a two-sample z-test and a pooled two-sample z-test assuming equal variances that are taught in introductory statistics courses

How To Fix This Problem?

General fix: Use an asymptotically pivotal test statistic, i.e. one that does not depend

Practically, how is this done?

Normalize the test statistic by an estimate of the standard deviation (studentize the test statistic)

This generally works for statistics which have an asymptotic normal distribution

Why This Works

If the asymptotic distribution does not depend on the underlying distributions of the test statistics, the following are equivalent in large samples

- The distribution of the test statistic where the samples are attained under the original population distributions (sampling distribution)

- The distribution of the test statistic where the samples from each population are generated from a pooled distribution of the original populations (permutation distribution)

Thank you!

Linkedin