Fairness in Criminal Justice Risk Assessment

Goal: Define different kinds of (statistical) fairness and their Tradeoffs

Terminology

Confusion Table:

	Failure Predicted (positive)	Success Predicted (negative)	Conditional Procedure Error
Failure – A positive	tp True Positive	fn False Negative	$\frac{fn}{tp+fn}$ = False negative rate
Success – A negative	fp False Positive	tn True Negative	$\frac{fp}{tn+fp}$ = False positive rate
Conditional Use Error	$\frac{fp}{tp+fp}$ = Failure Predict. Error	$\frac{fn}{tn+fn}$ = Success Predict. Error	$\frac{fp+fn}{fp+fn+tn+tp}$ = Overall procedure error

Margins of the table → estimates of corresponding probabilities in a population

Sample size: total number of observations $tp+fn+fp+tn = N$

Base rate: proportion of actual failures: $\frac{tp+fn}{N}$

or " " successes: $\frac{tn+fp}{N}$

Prediction: proportion predicted to fail: $\frac{tp+fp}{N}$

Prediction: " " succeed: $\frac{tn+fn}{N}$

Overall procedure error: proportion of cases misclassified

Conditional procedure error: proportion of cases ~~wrong~~ incorrectly classified conditional on actual outcome

" use ": " " predicted " " predicted outcome

Cost ratio: ratio of false negatives to false positives: $\frac{fn}{fp}$

or " " positives to negatives: $\frac{fp}{fn}$

⇒ Example parole release

Statistical Framework

Goal: draw inference beyond data being analyzed

→ **Idea**: each observation is randomly realized from a single joint probability distribution $P(Y, L, S)$ = „target population"

- Y = outcome of interest that is supposed to be inferred, eg. arrest while on probation
- L = legitimate predictors, eg. prior convictions
- S = protected predictors, eg race, gender, ...

⇒ population = limitless number of IID observations on $P(Y, L, S)$

In the population there is some function $f(L, S)$ linking predictors L, S to the expectation of Y: $E(Y|L, S)$. If Y is categorical: cond. expectations = cond. probabilities

A fitting procedure $h(L, S)$ is applied to the data
- Structure: eg. linear regression
- Optimization algorithm: minimizing sum of least squares

⇒ algorithm fitting $h(L, S)$ produces hypothesis $\hat{f}(L, S)$ which is the source of fitted values $\hat{Y}$

NOTE: allow S to participate in fitting procedure because it is associated with Y

→ $\hat{f}(L, S)$ will almost certainly be a biased estimator of the true response surface
- important legitimate predictors not available
- measurement error
- functional form arrived at for $\hat{f}(L, S)$ might be incorrect

Joint probability distribution is essentially abstraction of a high dimensional histogram for a finite population → expanded to limitless observations

Assumption: Data is realized independently from the same joint prob. dist = IID
→ whether this assumption makes sense for real data depends

Defining algorithmic fairness

Proceed as if $\hat{f}(L, S)$ provides estimates that are the same as the corresponding population values => do not conflate discussion about accuracy & fairness

1) Overall accuracy equality: $\frac{t_p + t_n}{t_n + f_p + t_n + t_p}$ is the same across all protected groups

-> t_p and t_n are equally desirable -> not true for many cost-weighted settings

Not really used because there is no distinction between success & failure accuracy

2) Statistical parity: $\frac{t_p + f_p}{N}$ or $\frac{t_n + f_n}{N}$ (marginal dist. of predictions) is the same across all protected groups

-> different from one another, but the same for all protected groups

-> "demographic parity": can lead to undesirable results, e.g. incarceration for women that are no public safety risk to maintain proportions

3) Conditional procedure accuracy equality: $\frac{t_p}{t_p + f_n}$ or $\frac{t_n}{t_n + f_p}$ is the same across all protected groups

=> "equalized odds": the accuracy of a predicted classification, given the true outcome class

e.g. accuracy of being classified as being arrested while on parole, given the person was actually arrested on parole

4) Conditional use accuracy equality: $\frac{t_p}{t_p + f_p}$ or $\frac{t_n}{t_n + f_n}$ is the same across all protected groups

=> the accuracy of being part of a particular class, given the predicted classification

e.g. accuracy of being arrested on parole, given the algorithm predicted rearrest on parole

5) Treatment equality: $\frac{fn}{fp}$ or $\frac{fp}{fn}$ is the same across all protected groups

→ „treatment" implies: policy lever to achieve other kinds of fairness

e.g. if fn are taken to be more costly for men ⇒ conditional procedure accuracy equality

BUT: men and women are not treated equally, as incorrectly classifying a failure on parole is more costly for men

6) Total fairness: all previously defined notions of fairness are achieved simultaniously

Other notions of fairness are not discussed, because they cannot be operationalized.

Estimate Accuracy

Previously: applying $h(L,S)$ to data ⇒ $\hat{f}(L,S)$ that estimates true response surface

But: no credible claim, that response surface is being estimated in unbiased manner

Still: With more samples → smaller estimation error

bias + variance = no conventional confidence bound

⇒ bias cannot be removed w/o comparison to truth, which is unknown

Alternative: $\hat{f}(L,S)$ estimates a response surface in population that is an approximation of the true in the population $\hat{f}(L,S)$ has the same form as the approximation → probability estimates of cross-tabulation on data = prob est. of a Y-by-$\hat{Y}$ cross-tabulation of applying $h(L,S)$

⇒ since data is assumed to be IID ⇒ estimates are asymptotically unbiased

Estimate accuracy is then measured by out-of-sample performance

Tradeoffs

Fundamentally: $h(L,S)$ capitalizes on non-redundant associations that L and S have with Y

=> excluding or discarding S will reduce accuracy

But fairness is important as well => examinations of trade offs

Proven Impossibility Theorems:

„When the base rates differ by protected groups and when there is not separation, one cannot have both conditional use accuracy and equality in false negative/positive rate." (reverse cond. procedure acc.)

Definitions:

Separation: observations are seperable if perfectly accurate classification is possible

=> for each possible configuration of predictor values there is some $h(L,S)$ for which probability of membership in a given outcome class is always either 1.0 or 0.0

Calibration: proportion of people experiencing an outcome (base rate)
=
proportion of people predicted to experience an outcome

- indicator for algorithm performance
- matter of fairness if different quality of calibration across groups

=> If there is variation in base rates and no separation
→ goal of complete race and gender neutrality is unachievable

=> example parole release

Potential Solutions

1) Preprocessing: eliminate sources of unfairness in data before $h(L,S)$

- Remove later dependencies between L and S
  - regress in turn each predictor in L on predictors in S
    → work with residuals => „residualized" transformations of predictors
    Problem: interaction effects are not removed unless specifically included
  - transform predictors, so that fair predictors can be obtained while „preserving as much information as possible" => Euclidian distances between original & transformed predictors
    → residualization in turn using results from previous residualizations and interactions from protected classes
- Rebalancing base rates:
  - direct rebalancing: apply weights for each group seperately → base rates same across categories
  - randomly relabel some response values

Problems: 1) accuracy loss
2) differing false positive/negative rates across groups

- Direct vs. Indirect discrimination: features of protected class vs. related feature of prot. class
  - perturbe class membership:
  - perturbing outcome (label ⟷ also changes base rate!)
- randomly transform all predictors except for indicators of protected class membership
  => joint distribution less correlated on prot. class membership
  Constraints: 1) joint distr. of transformed variables is very close to joint distr. of original pred.
  2) no individual case is substantially disturbed

Problems: • effect on different kinds of fairness unclear
• accuracy price

2) In-Processing: make fairness adjustments as part of the process of constructing $h(L,S)$

- alter results of ~~prediction~~ inference to more fair results, if ~~prediction~~ result itself is very uncertain
  → reduction of out-of-sample accuracy might be small
  Problem: might have unacceptable consequence for fpr/fnr rates

- add a penalty term to penalize unfairness
  - for example: penalty for violations of conditional procedure acc. equality
  Problems: - loss function typically not convex
  - undesirable implications for other kinds of fairness

3) Post-processing: adjust results after applying $h(L,S)$ to make it more fair

- „equalized odds" (cond. proc. acc. eq.): probabilistic switching of assigned labels
  - binary case: $\hat{f}(L,S)$ assigns 0 or 1
  - Construct linear program model: assign each label a prob $[0,1]$ to switch label
  => minimize the classification error s.t. one of the two fairness controls by changing the values of the probabilities

Problems: - implications for other kinds of fairness unclear
- conditional use acc eg. can suffer
- lower accuracy (=> target: overall class error as error of worst group)
- reassignment prob. larger when base rates are more disparate

Making fairness operational: consider question of benchmark:
of equality is achieved, with regard to what?
e.g. fairness in terms of prison sentence length between black and white
offenders does not make statement about length of sentences except for the two
equally long

Future Work:
- Combination of technical & policy challenges
- Benchmark: current practice → even small changes can lead to meaningful improvement

But: used data reflects past practices → current decisions affect future training data
↳ algorithmic solutions/results need to be regularly updated

Technical questions: - data long?
- discarding of historic data? How much?
- more weight for recent training data?

⇒ Example of In-Process Fairness

Conclusion:
fairness is a subtle problem ⇒ tradeoffs inevitable
⇒ Core Problem: Different base rates across protected groups

Paths forward:
1) Statistical procedures can improve transparency, accuracy of fairness aspect to judgment
2) Tradeoffs need to be made explicit & available as hyperparameters
Measures of fairness need to be formalized
3) Decisions about tradeoffs need to be made by stakeholders
→ matters of law, values of the political process
4) There will be no quick solutions