

Ethical Considerations for Classification Algorithms

Table of contents

- Introduction to Classification
- Breast Cancer Screening
- Criminal Risk Assessment Algorithms

Introduction to Classification

Classification Algorithms

- We'll consider binary classification algorithms which use data to predict one of two outcomes which we'll refer to “**positive**” and “**negative**” classes.
- Most problems have a fundamental asymmetry: not all mistakes are created equal!
 - Example: consider a medical test that returns “positive for cancer” or “negative for cancer”

- Most machine learning classifiers have tunable threshold parameters which allow us to increase (or decrease) the number of positive predictions which necessarily decreases (increases) the negative predictions

Evaluating Classifiers

Evaluating Classifiers

There are many other commonly metrics which are used to understand classification accuracy:

Sources: [9][10][11][12][13][14][15][16] view · talk · edit

		Predicted condition			
		Predicted positive	Predicted negative	Informedness, bookmaker informedness (BM) $= \text{TPR} + \text{TNR} - 1$	Prevalence threshold (PT) $= \frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Positive (P) [a]	True positive (TP), hit ^[b]	False negative (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\text{TP}}{\text{P}} = 1 - \text{FNR}$	False negative rate (FNR), miss rate, type II error ^[c] $= \frac{\text{FN}}{\text{P}} = 1 - \text{TPR}$
	Negative (N) ^[d]	False positive (FP), false alarm, overestimation	True negative (TN), correct rejection ^[e]	False positive rate (FPR), probability of false alarm, fall-out, type I error ^[f] $= \frac{\text{FP}}{\text{N}} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{\text{TN}}{\text{N}} = 1 - \text{FPR}$
		Prevalence $= \frac{\text{P}}{\text{P} + \text{N}}$	Positive predictive value (PPV), precision $= \frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR}$	Negative predictive value (NPV) $= \frac{\text{TN}}{\text{TN} + \text{FN}} = 1 - \text{FOR}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$
					Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$
		Accuracy (ACC) $= \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$	False discovery rate (FDR) $= \frac{\text{FP}}{\text{TP} + \text{FP}} = 1 - \text{PPV}$	False omission rate (FOR) $= \frac{\text{FN}}{\text{TN} + \text{FN}} = 1 - \text{NPV}$	Markedness (MK), deltaP (Δp) $= \text{PPV} + \text{NPV} - 1$
					Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$
		Balanced accuracy (BA) $= \frac{\text{TPR} + \text{TNR}}{2}$	F ₁ score $= \frac{2 \text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2 \text{TP}}{2 \text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) $= \sqrt{\text{PPV} \times \text{TPR}}$	phi or Matthews correlation coefficient (MCC) $= \frac{\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}}}{\sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}$
					Threat score (TS), critical success index (CSI), Jaccard index $= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

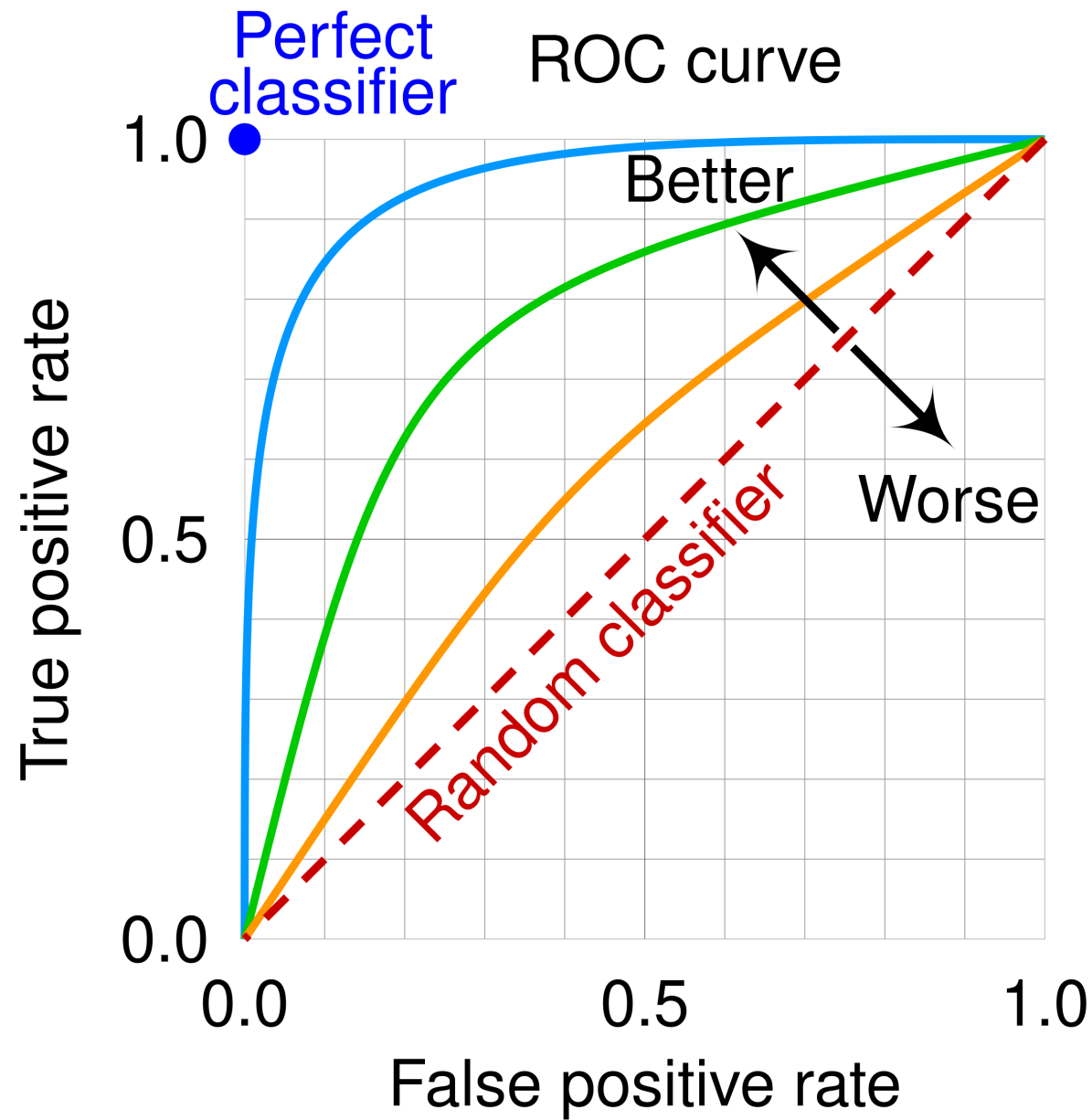
Source: Wikipedia, “Receiver operating characteristic”

In this module, we'll focus on the tradeoff between false positive rate and false negative rate

A Tradeoff: FPR vs FNR

- Tunable threshold parameter which allow us to increase or decrease the number of positive predictions
- Increasing positive predictions typically decreases the false negative rate but increases the false positive rate
- This tradeoff is often displayed using a Receiver Operator Curve (**ROC**)

The ROC Curve



Source: Wikipedia, "Receiver operating characteristic"

Breast Cancer Screening

Example: Breast Cancer Screening

Consider a breast cancer screening procedure,
e.g. mammography.

- True Positive: Cancer detected when cancer is present
- False Positive: Cancer detected when no cancer is present
- True Negative: No cancer detected when no cancer is present
- False Negative: No cancer detected when cancer is present



Instructor Note



These screening procedures may or may not actually involve AI or machine learning algorithms, but the example is still a useful one for understanding the ethical tradeoffs and considerations for classification accuracy.



Possible Discussion Questions



What are some possible consequences of a false positive result?

Some possible consequences to discuss:

- psychological harm: anxiety, depression, and reduced quality of life from false alarms
- physical harm: Unnecessary biopsies, additional radiation exposure, potential surgical complications
- economic burden: Costs of follow-up testing, lost work time, insurance implications
- healthcare resource allocation: Overuse of limited medical resources



Possible Discussion Questions



What are some possible consequences of a false negative result?

Possible consequences of false negatives:

- delayed treatment: Cancer progresses while undetected, potentially becoming more aggressive or metastatic
- survival implications: later-stage diagnosis often means worse prognosis
- trust in medical systems: patients may lose confidence in screening programs
- legal and professional liability: malpractice concerns for missed diagnoses

Example: Breast Cancer Screening

Equity

Cancer risk and screening accuracy vary depending on a number of factors like race and ethnicity, age and breast density. In addition, not all populations have equal access to state-of-the-art screening technologies.

- Why do these inequities exist?
- How might we ensure more equitable access and more fair screenings?



Possible Discussion Questions



- Access to advanced screening: advanced technologies are more expensive and less available in underserved communities.
- Geographic disparities: rural and low-income urban areas often lack specialized breast imaging centers, forcing women to travel long distances for screening or rely on older, less accurate equipment. Mobile screening units may use different (often older) technology with different accuracy profiles.
- Insurance and economic barriers: Even when screening is covered by insurance, costs of follow-up testing after false positives can lead women with limited resources may skip recommended follow-up, turning false positives into dangerous delays.



Possible Discussion Questions



Is it ethical to have different screening protocols for different populations if it improves overall accuracy? Could this be considered discriminatory?

Criminal Risk Assessment Algorithms

Criminal Risk Assessment Algorithms

- Machine learning is increasingly used to to inform decisions about individuals in the criminal justice system
 - Setting bond amounts
 - Length of sentence
- One major component of these decisions is trying to evaluate a defendant's risk of future crime
- *Recidivism*: the tendency of a convicted criminal to re-offend.

The COMPAS recidivism algorithm

- In 2016, [ProPublica published an a story](#) about one of the most used algorithms called ‘COMPAS’
- COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions, and was mostly used to assess the risk of a pre-trial release
- In their article, they conducted a rigorous analysis of possible bias by analyzing data and COMPAS predictions from more than 10,000 criminal defendants in Broward County, Florida

Fairness & Transparency

- Fairness: ideally, COMPAS would have the same impact on all demographic subgroups. Probably not possible!
- Transparency: COMPAS is a *closed-source* algorithm. The public is not allowed to see how the algorithm works.

Question

What information about an individual do you think is “fair” to include in an algorithm which predicts recidivism? What information would be “unfair”?

COMPAS data

The risk-scores are derived come from answers to a **137 question survey** and the defendant's criminal record.

Predictors used by COMPAS include:

- Prior arrests and convictions (and if any friends had priors)
- Address, GPA, wealth
- If the defendant's parents separated

Important

Purportedly, **race** is not used as a predictor.

Recidivism Data

There are 6172 observations in the data we'll analyze. 45% of individuals received a high risk score.

Age Group	Proportion	Race	Proportion
25 - 45	0.57	African-American	0.51
Greater than 45	0.21	Asian	0.01
Less than 25	0.22	Caucasian	0.34
		Hispanic	0.08
		Native American	0.00
		Other	0.06

Predictions

- The COMPAS algorithm is closed-source but we can learn about the algorithm by fitting **our own model** to the COMPAS predictions.
- Predict COMPAS score (“high” vs “low” risk) using data about the defendants collected by ProPublica. This yields an explainable **proxy** for the internal workings of COMPAS.
- Our data includes age, race, and prior counts
- Compare the predictions for different inputs to the model to impute how COMPAS seems to react to such changes

Predictions

Call:

```
glm(formula = score_factor ~ gender_factor + age_factor + race_factor +  
     priors_count, family = "binomial", data = compas_scores)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.85391	0.10844	-26.317	< 2e-16	***
gender_factorFemale	0.14358	0.07802	1.840	0.065727	.
age_factor25 - 45	1.45043	0.09875	14.688	< 2e-16	***
age_factorLess than 25	2.87679	0.11411	25.211	< 2e-16	***
race_factorAfrican-American	0.48841	0.06840	7.140	9.3e-13	***
race_factorAsian	-0.33565	0.47979	-0.700	0.484193	
race_factorHispanic	-0.44845	0.12638	-3.549	0.000387	***

Note

The classifier we use here is called a “logistic regression”. You’ll learn much more about this in later statistics courses.

Exploring Predictions

- Use our classifier to predict the *probability* that COMPAS would give an individual a high score
- Looks at inputs which are identical except in one characteristic
- e.g., compare predictions for different ages and different races

The importance of age

Compare predictions for two hypothetical individuals identical in all characteristics except age.

Sex	Age Group	Race	Priors	Prob. High
Male	25 - 45	African-American	0	0.29
Male	Greater than 45	African-American	0	0.09

The importance of race

Compare predictions for two hypothetical individuals identical in all characteristics except race.

Sex	Age Group	Race	Priors	Prob. High
Male	25 - 45	African-American	0	0.29
Male	25 - 45	Caucasian	0	0.20

 Important

But “race” is purportedly not an input to COMPAS!

Proxy Variables

Just because **race** isn't an input to the algorithm does not mean the algorithm makes the same predictions for all race groups!

Question

How might the COMPAS algorithm implicitly factor racial information into its predictions even though **race** is not an input?

Examining FPR and FNR

- ProPublica provides a binary variable named `two_year_recid` which indicates whether an individual committed a new crime within two years of the screening or not.
- We can compare the collected COMPAS risk predictions to the `two_year_recid` variable which we'll assume is our “ground truth” to compute FPR and FNR
- There is a tradeoff between FPR and FNR!

Examining FPR and FNR

```
1 compas_scores %>% xtabs(data=., ~ score_factor + two_year_recid)
```

	two_year_recid	
score_factor	0	1
LowScore	2345	1076
HighScore	1018	1733

- **FPR:** $1018/(2345 + 1018) = 0.30$
- **FNR:** $1076/(1733 + 1076) = 0.38$
- **Accuracy:** $(2345 + 1733)/n = 0.66$



Possible Discussion Questions



- What is the meaning of FPR and FNR in this context? Is one worse than the other?
- In this example who decided what false positive rate and false negative rate is acceptable? Who *should* get to decide this?

FPR and FNR for African-Americans

```
1 compas_scores %>%  
2   filter(race == "African-American") %>%  
3   xtabs(data=., ~ score_factor + two_year_recid)
```

	two_year_recid	
score_factor	0	1
LowScore	873	473
HighScore	641	1188

$$\text{FPR} = 641 / (873 + 641) = 0.57$$

$$\text{FNR} = 473 / (473 + 1188) = 0.28$$

$$\text{ACC} = (999 + 414) / (999 + 408 + 414 + 282) = 0.65$$

FPR and FNR for Caucasians

```
1 compas_scores %>%  
2   filter(race == "Caucasian") %>%  
3   xtabs(data=., ~ score_factor + two_year_recid)
```

	two_year_recid	
score_factor	0	1
LowScore	999	408
HighScore	282	414

$$\text{FPR} = 282 / (282 + 999) = 0.22$$

$$\text{FNR} = 408 / (408 + 414) = 0.50$$

$$\text{ACC} = (999 + 414) / (999 + 408 + 414 + 282) = 0.67$$

FPR and FNR

	FPR	FNR	ACC
African-American	0.57	0.28	0.65
Caucasian	0.22	0.50	0.67
All	0.30	0.38	0.66

Similar accuracy across the two race groups are attained in very different ways!

High FPR for African-Americans and high FNR for Caucasians.

The Impossibility of Fairness

- If race is not included as a predictor in the classifier it is not possible in general to match FPR and FNR across both groups!
- If we use a “race-aware” classifier, we might be able to match FPR and FNR, but only if we use different classification criteria for each race (is that fair?)
- In fairness literature these constraints are known as “the impossibility of fairness” ([Kleinberg, Mullainathan, and Raghavan 2016](#))



Instructor Note



This content is likely for more advanced students.

- Can create two classifiers, one for each race group. If the ROCs intersect anywhere, then the FPR and FNR match at those intersection points (seemingly more fair). This is known as *equalized odds*.
- However, two individuals that match in all characteristics except race can receive a different prediction (seemingly unfair). This violates *calibration*.



Possible Discussion Questions



- If we can balance FPR and FNR, is that desirable even if we use different criteria for each race group to classify patient risk?
- Alternatively, if we use a single “race-blind” classifier, we could match either FPR or FNR across race groups but not both. Which would be preferable?

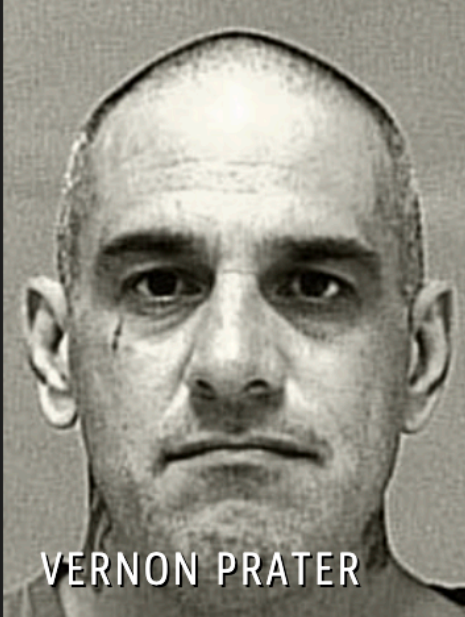
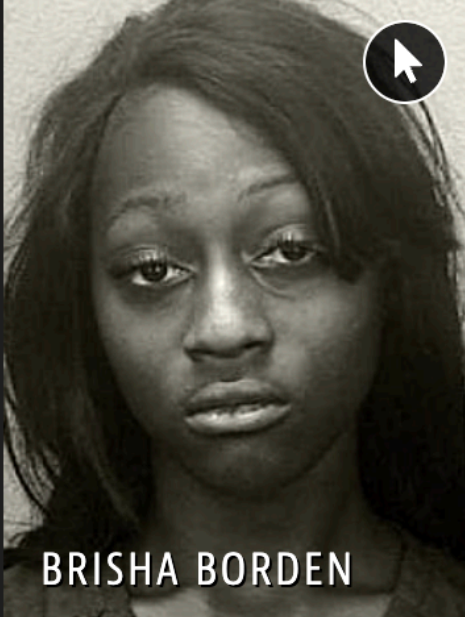
Putting a human face on it

ProPublica discusses two different incidents in Broward County:

1. In 2014, an 18-year-old girl, Brisha Borden, and her friend grabbed an unlocked bicycle and scooter and rode them down the street, then abandoning it. Police arrived and arrested the girls for burglary and theft of \$80 worth of goods.

Putting a human face on it

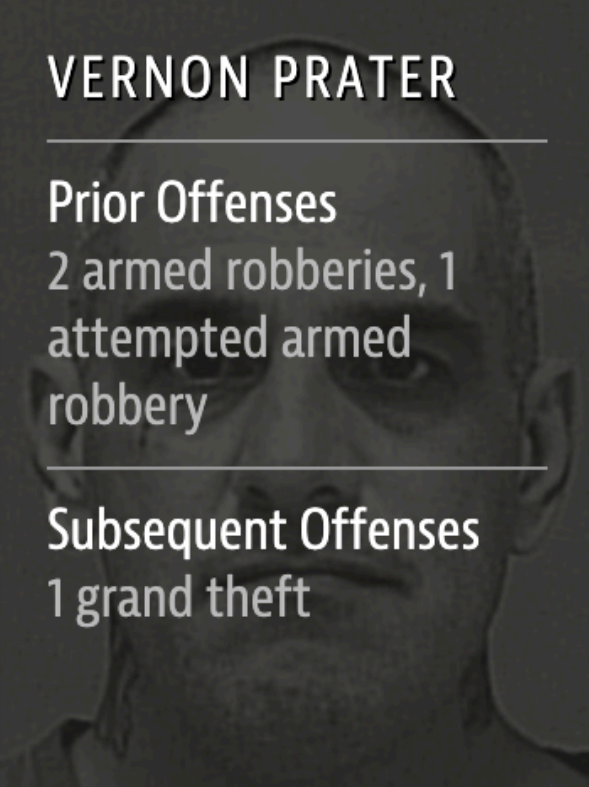
Two Petty Theft Arrests

	
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

source

Putting a human face on it

 <p>VERNON PRATER</p> <hr/> <p>Prior Offenses 2 armed robberies, 1 attempted armed robbery</p> <hr/> <p>Subsequent Offenses 1 grand theft</p> <p>LOW RISK 3</p>	 <p>BRISHA BORDEN</p> <hr/> <p>Prior Offenses 4 juvenile misdemeanors</p> <hr/> <p>Subsequent Offenses None</p> <p>HIGH RISK 8</p>
---	---

source

Read more

- [The ProPublica article, “Machine Bias”](#),
- [Methodology for the analyses used in “Machine Bias”](#)
- [Fairness and Algorithmic Decision Making - COMPAS Chapter](#)

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *arXiv Preprint arXiv:1609.05807*.