SOURCE OF BIAS IN MODEL INTERPRETATION AND USAGE

LECTURE 8
SECTION 3
JUNE 15TH

INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

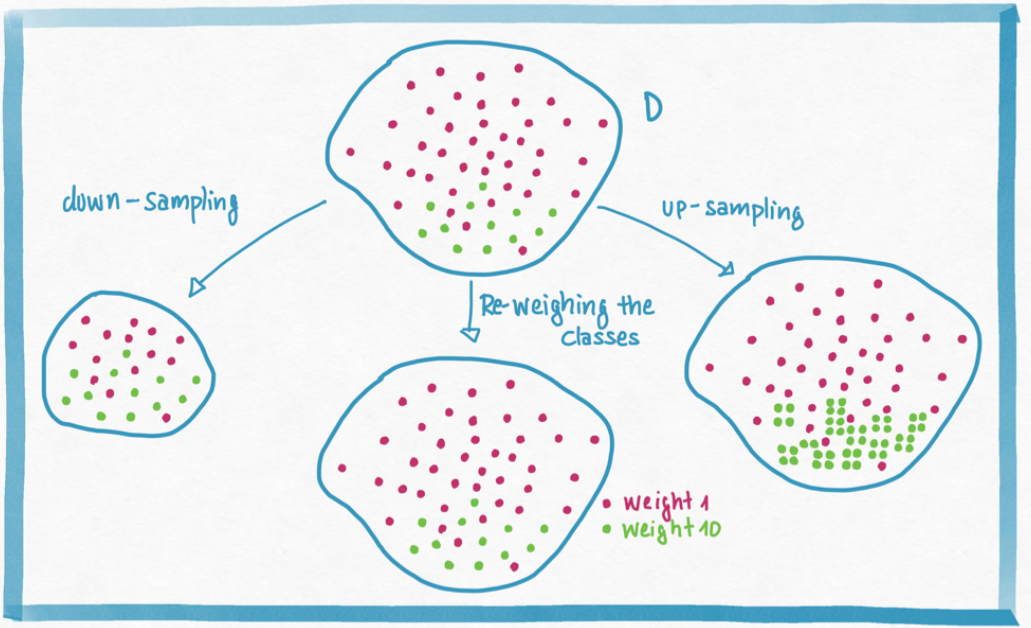
UNIVERSITY of RWANDA

SAMPLING BIAS

CLASS IMBALANCE:

We know that classifiers can be biased towards the majority class when the classes are extremely unbalanced. We can detect this problem using the confusion matrix. Changing the classification threshold can help. So can balancing the classes.

Down-Sampling is creating a new training set with less instances of the majority class by sampling a small number from the majority class in the dataset.



Up-Sampling is creating a new training set with more instances of the rare class, by sampling from the rare class in the existing dataset with replacement.

Class re-weighing is assigning more weight to loss function terms corresponding to the minority class. So the model is incentivized to do better on the minority class.

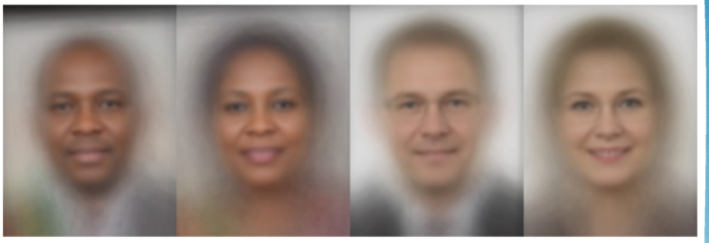
THE EFFECT OF UNBALANCED DATASETS:

It's important to balance the dataset even for regression models: if your dataset contain very few instances from Group A, your model will not be incentivized to do well on Group A. This will be a problem whenever your data contains an under-represented group, e.g. patients with rare disease, minority ethnic groups etc.

Racial & Gender Difference in Classifier Performance

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Classification Accuracy



Facial recognition softwares made by large tech companies and are commonly used in security & law enforcement consistent show lower accuracy for darker skin-tones and females.

One problem is that most of these models were trained on data with mostly white male faces.

BIAS FROM UNMEASURED COVARIATES

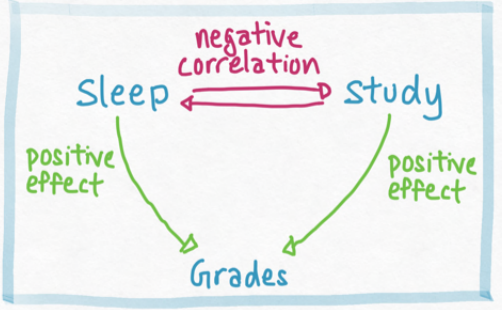
BIAS FROM UNMEASURED CONFOUNDING:

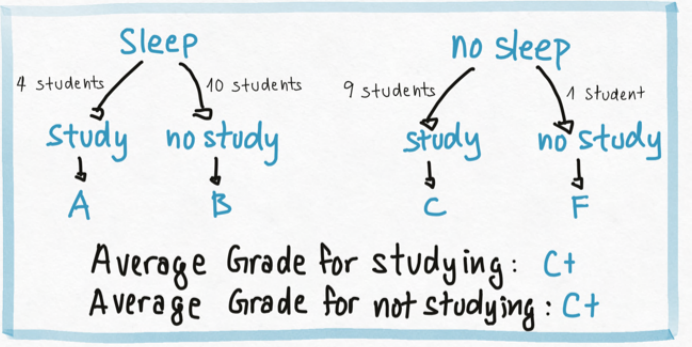
Our models can be biased by data we did not collect. If there are unmeasured variables that affect our target and or other measured covariates, we call them unobserved confounders.

Unobserved confounders can skew our model interpretation.

Example: We collected study habits of students to build a linear model to predict their grade. We find that the coefficient of "# of study hours" is zero. What does this suggest about the relationship between studying and grades?

If we know that "# of sleep hours" is an unobserved confounder, does this change our interpretation?





CORRELATION IS NOT CAUSATION:

Just because our parameters estimates are trustworthy it doesn't mean that their interpretations are obvious.

Example: Patient data has been collected in an hospital to model the probability of surviving COVID-19. Suppose our model is:

$$P(y=1 \mid x, w) = 1.6\,X_1 + 0.5\,X_2 - 0.7\,X_3 - 4$$

y: survived or not

X_1: patient has asthma or not
X_2: patient oxygen saturation
X_3: patient blood pressure

X_1 has the largest coefficient, suggesting a strong correlation between having asthma and surviving the ICU.

Is there some biological mechanism that causes asthma patients to be more resiliant against COVID-19?

What if you found that at this hospital, patients who have asthma were hospitalized earlier and monitored more intensely than non-asthma patients?

Then it is more likely that having asthma led to better care, and better care caused the increase in survival rate. Without the better care asthma would likely lower the survival rate.