Introduction to Machine Learning

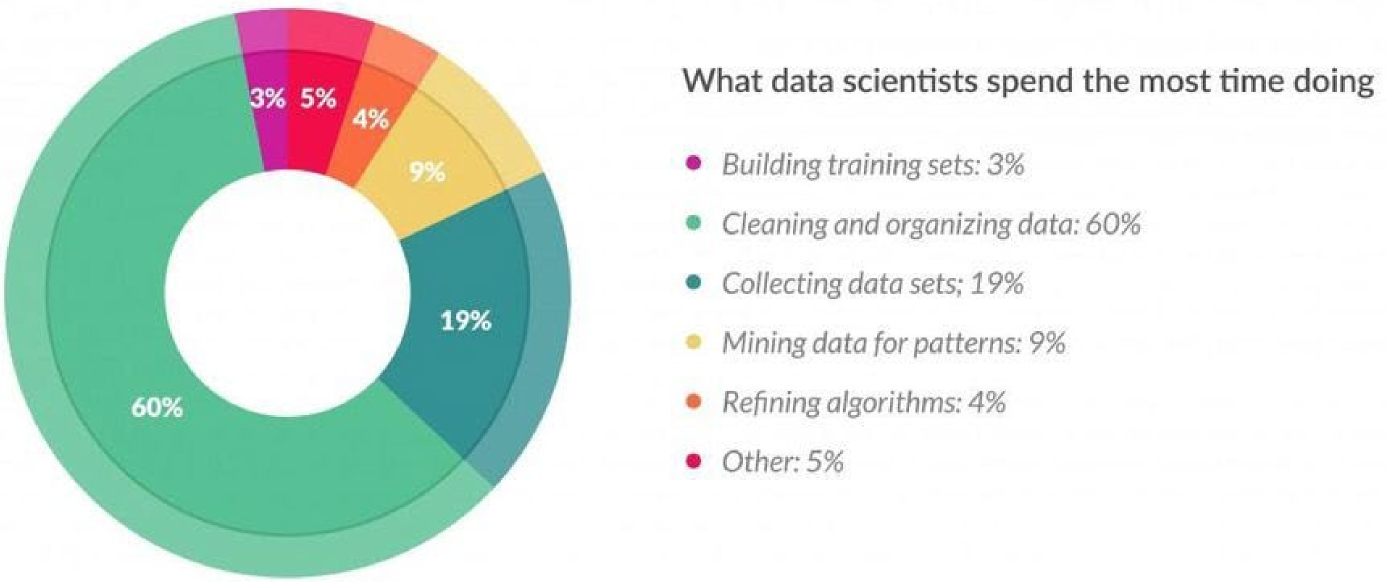
Practical Considerations:

Features
and
Handling Class Imbalance

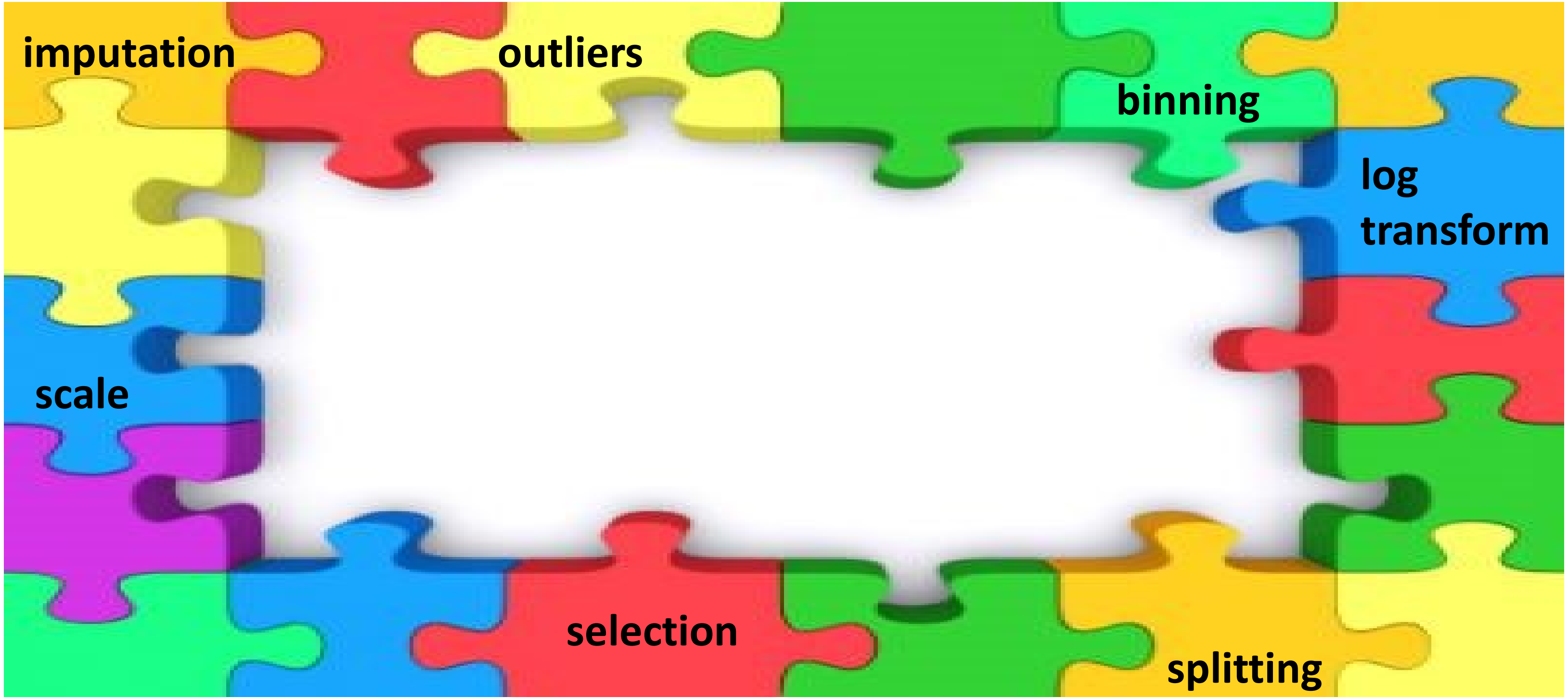
Features

“garbage in, garbage out”

Features


3%
5%
4%
9%
19%
60%
What data scientists spend the most time doing
Building training sets: 3%
Cleaning and organizing data: 60%
Collecting data sets; 19%
Mining data for patterns: 9%
Refining algorithms: 4%
Other: 5%

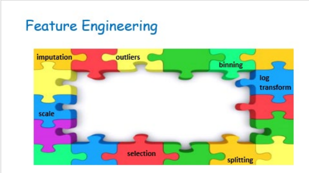
Feature Engineering



Imputation

- Why are values missing?

- What do we do?

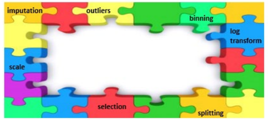


Outliers

- Standard deviation
- Percentiles

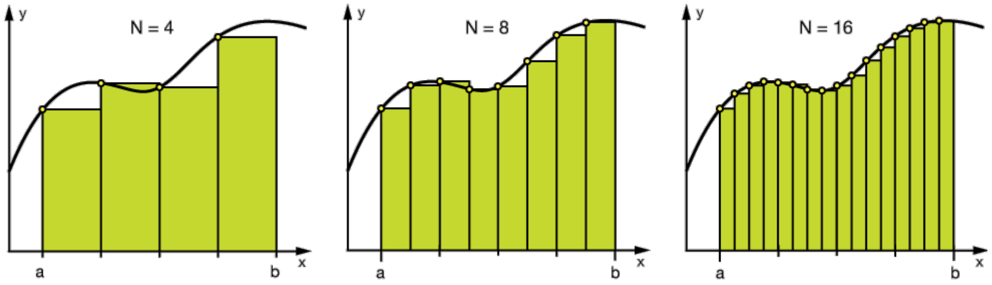
- What do we do?

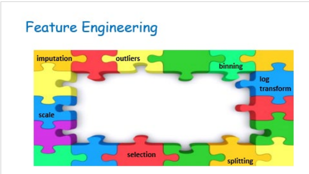
Feature Engineering



Binning

- Equal width
- Equal frequency
- Bin value

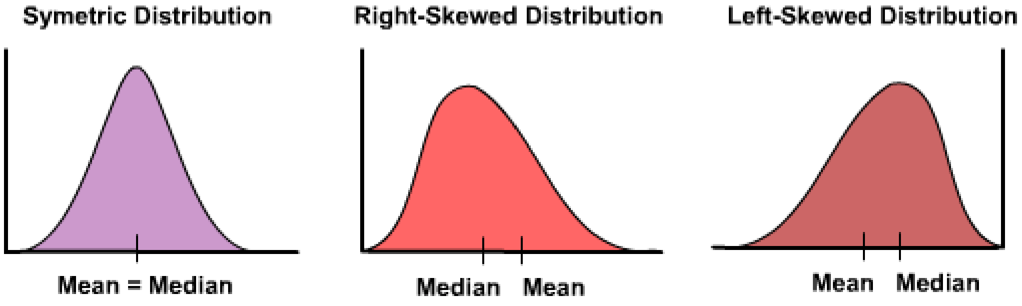


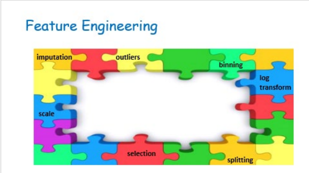


Log Transform

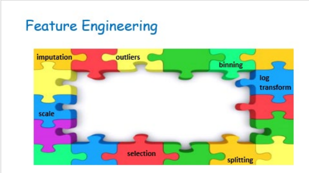
- Another way to normalize

- Log(x+1)



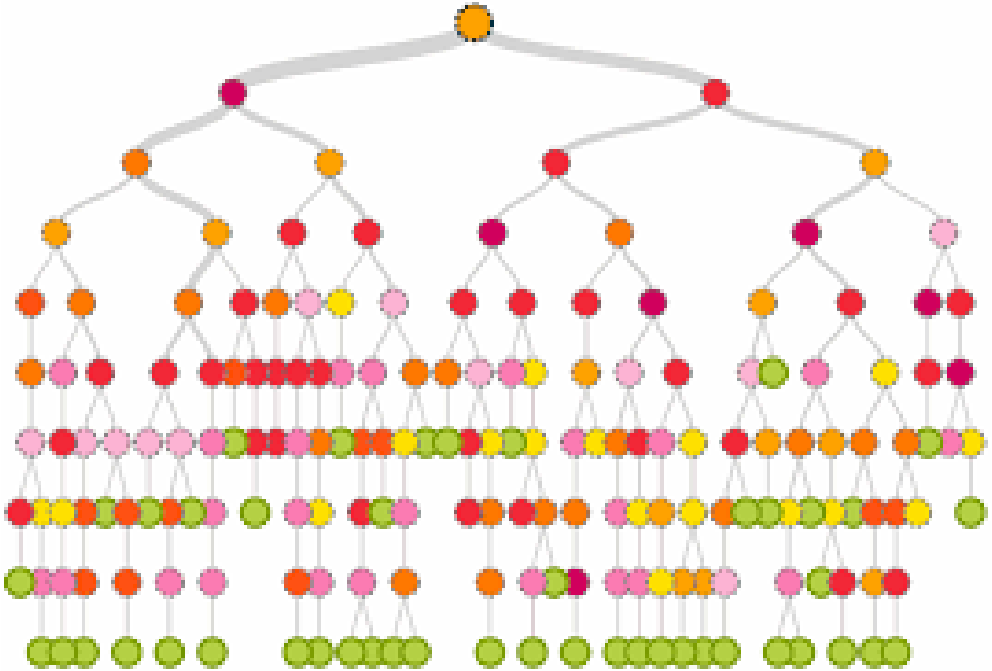


Splitting


Feature Engineering
imputation
outliers
binning
log
transform
scale
selection
splitting

Selection

- Prune features with low variance
- Prune redundant features
  - Use decision tree
  - Wrapper
  - Correlation



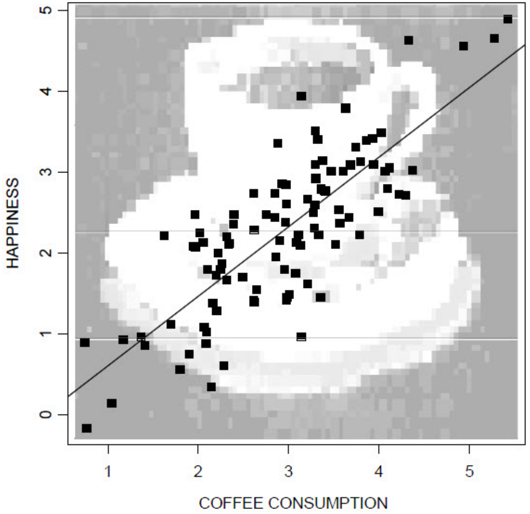
Correlation

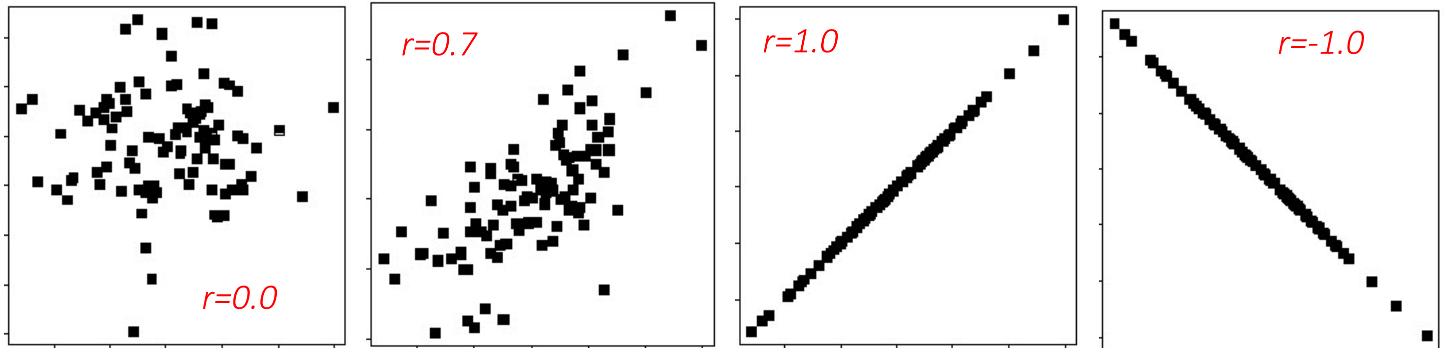
- Covariance

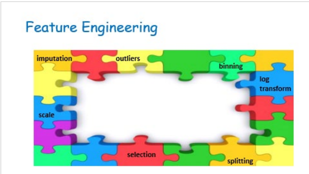
$$Cov(X,Y) = \sum_{i=1}^{N} \frac{(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- Correlation

$$Cor(X,Y) = r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$



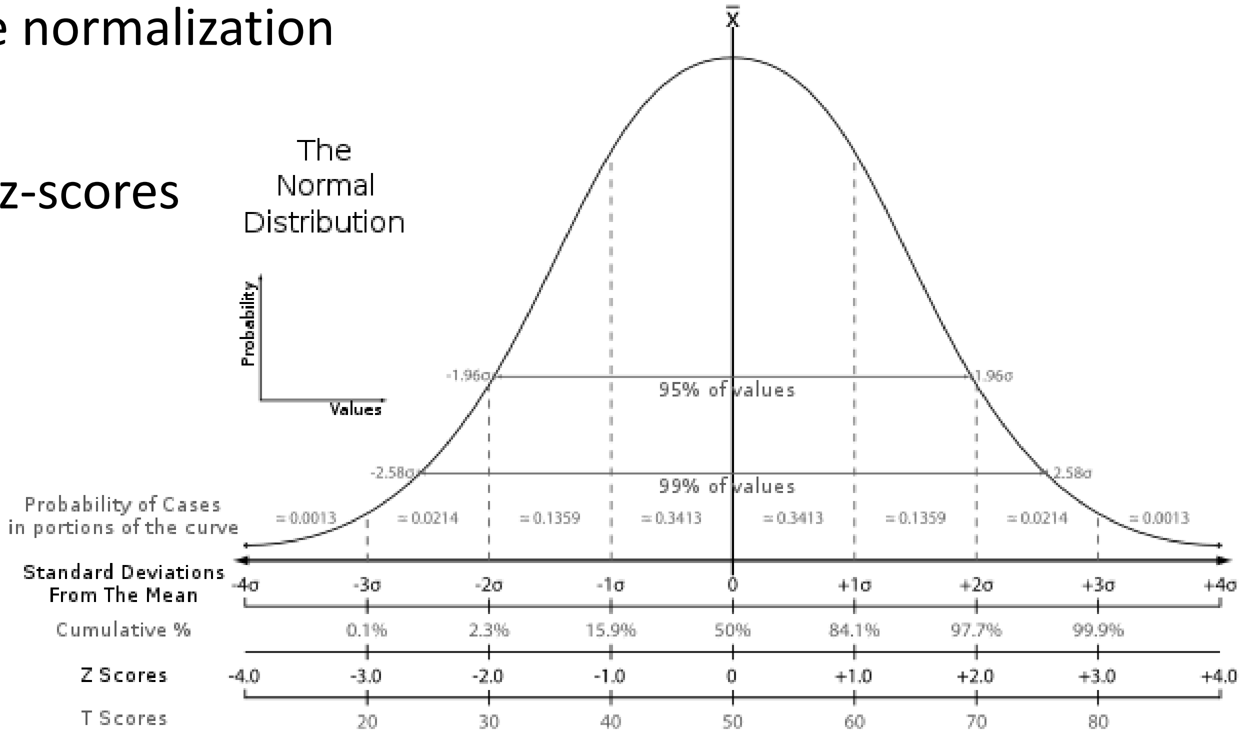


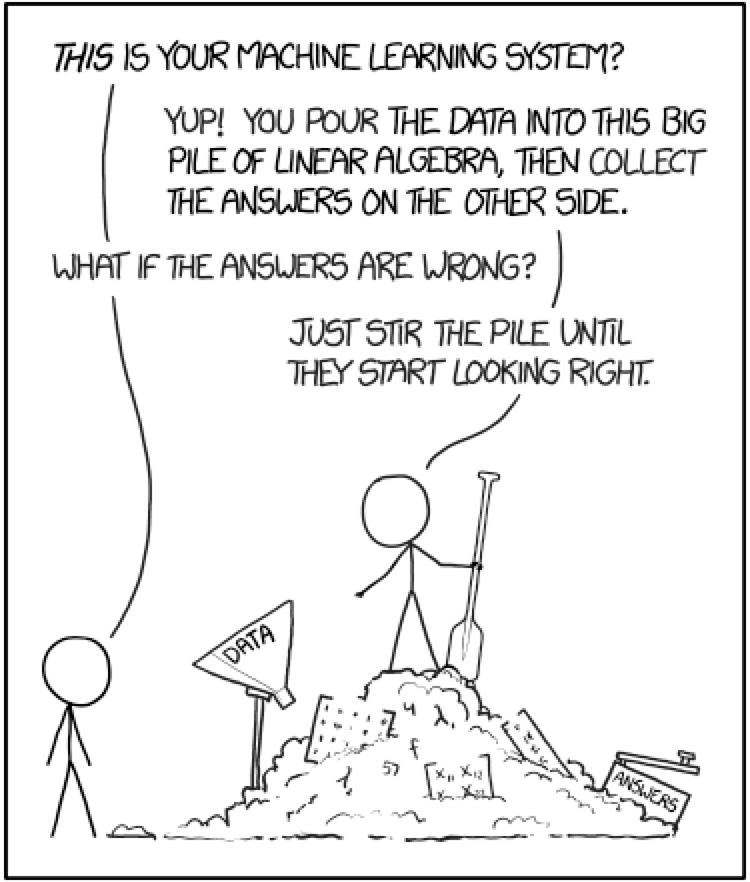


Scale

- We discussed feature normalization

- Another approach is z-scores (standard scores)




THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

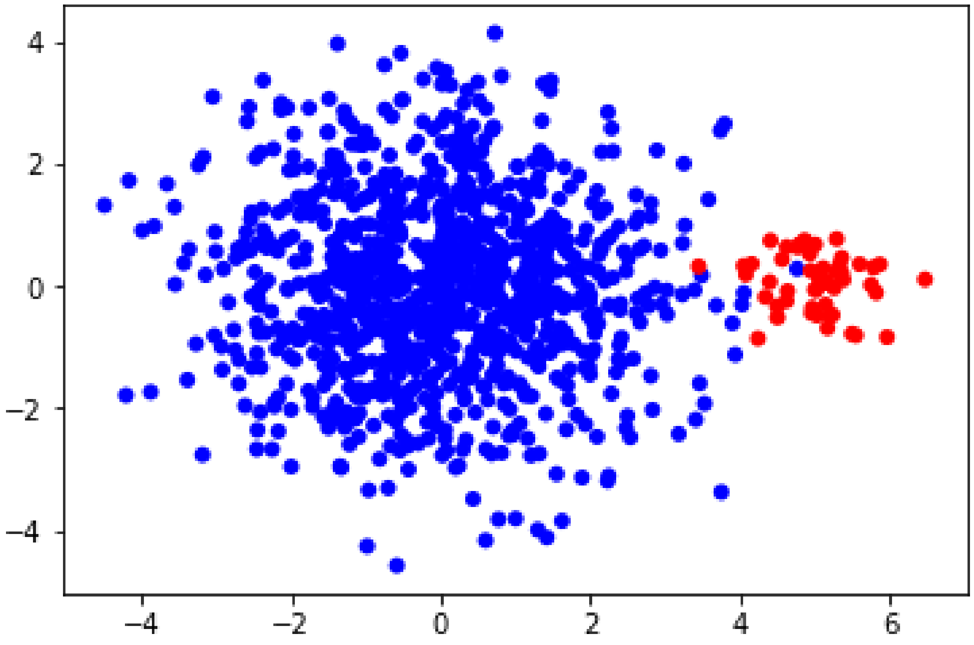
Handling Class Imbalance

"I achieved that 99.9% fraud detection accuracy you requested."

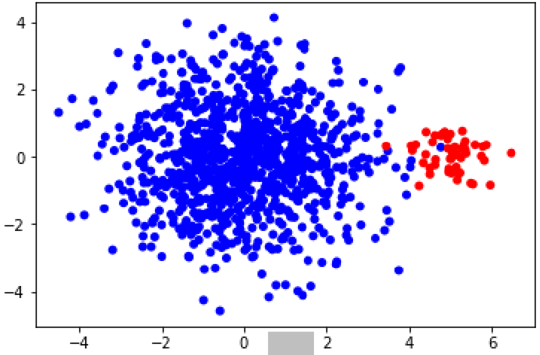
The catch?

Imbalanced class distributions - misleading results

Imbalanced class distributions

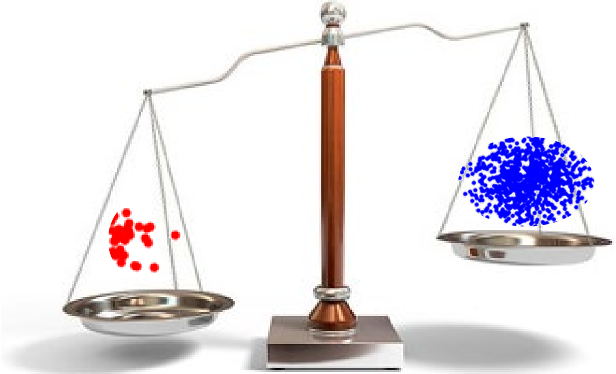


Imbalanced class distributions - misleading results

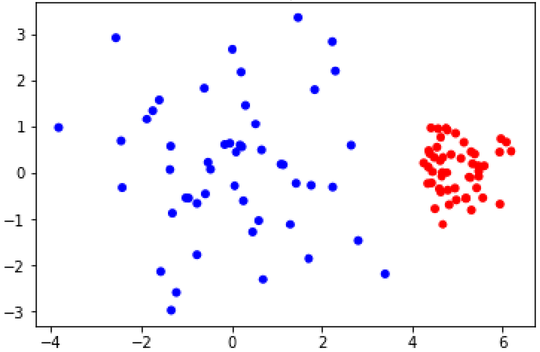


Imbalanced data

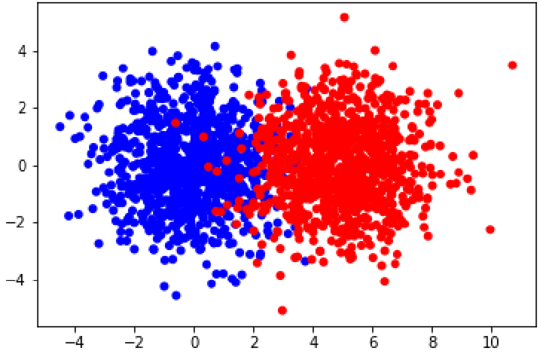
Weight



Under-sample



Over-sample



Standard binary classification

TASK: BINARY CLASSIFICATION

Given:

1. An input space $\mathcal{X}$
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$
3. A training set D sampled from $\mathcal{D}$

Compute: A function f minimizing: $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(x) \neq y\right]$

Standard binary classification

TASK: α-WEIGHTED BINARY CLASSIFICATION

Given:

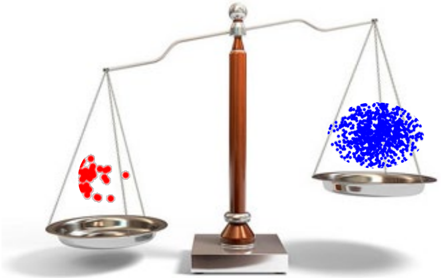
1. An input space $\mathcal{X}$
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$
3. A training set D sampled from $\mathcal{D}$

Compute: A function f minimizing: $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\alpha^{y=1}\left[f(\boldsymbol{x}) \neq y\right]\right]$

How use this for imbalanced class distribution?

- Fraudulent transactions, 0.01% data
  - Weight = 0.0001
- Normal transactions
  - Weight = 0.9999

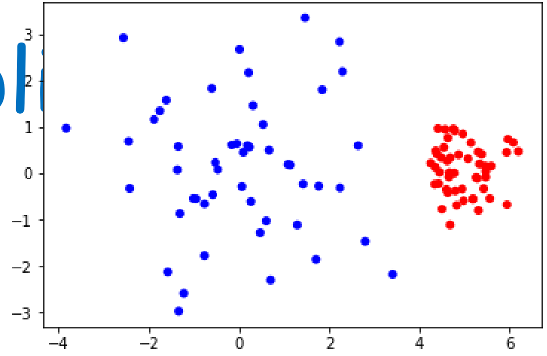
Importance weight



- Entropy(S) = $-p_+ \log_2 p_+ - p_- \log_2 p_-$

- WeightedEntropy(S) = $(\alpha) -p_+ \log_2 p_+ - (1-\alpha)\, p_- \log_2 p_-$
- Gain(S,A) = Entropy(S) - $\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \mathrm{Entropy}(S_v)$
- WeightedGain(S,A) = Entropy(S) - $\sum_{v \in Values(A)} \frac{\alpha|S_v^+|+(1-\alpha)|S_v^-|}{\alpha|S^+|+(1-\alpha)|S^-|}$

Sampling (undersampl



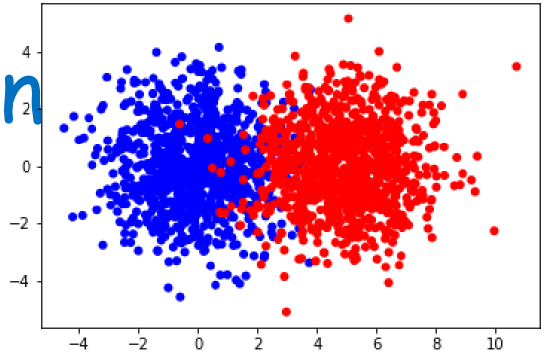
Algorithm 11 $\textsc{SubsampleMap}(\mathcal{D}^{weighted}, \alpha)$

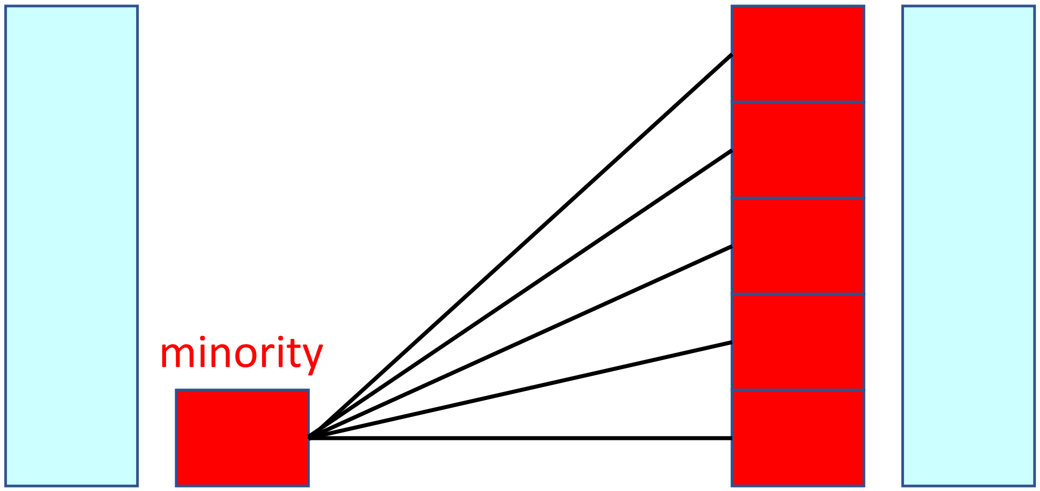
```
while true do
    (x, y) ~ D^weighted                // draw an example from the weighted distribution
    u ~ uniform random variable in [0, 1]
    if y = +1 or u < 1/α then
        return (x, y)
    end if
end while
```

Algorithm 12 $\textsc{SubsampleTest}(f^{\text{BINARY}}, \hat{x})$

```
return f^BINARY(x̂)
```

Sampling (oversamplin



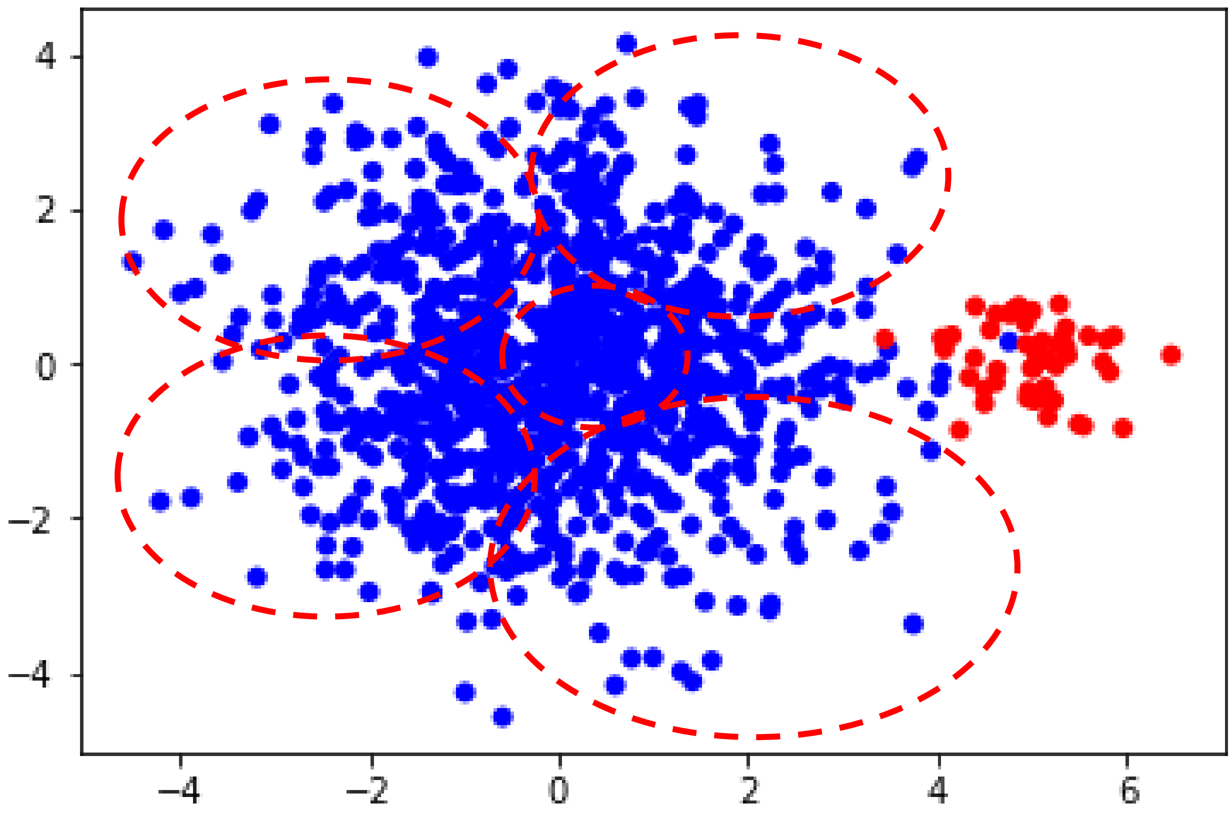


Original dataset

Oversampled dataset

Let's try this out

Decompose



Let's try this out