UNIVERSIDAD
POLITÉCNICA
DE MADRID

Spark Project

18/12/2018 - Massively Parallel Machine Learning

Prof. Alberto Mozo

Filippo Calzavara
Nicolae Righeriu

Data Preprocessing

- Cleaning
    - Removed 57th column
- Sum and transposition
    - Create a structure to mantain the sum and transpose the data
- Keying and shuffling
    - To each sample is assigned a random id. The structure is the sorted by id (shuffling)
- Structure returned:

(int sample_id, boolean train_or_test, Float[ ] X, boolean target)

Normalization and Split

- Calculation of the average
  - For each column

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Calculation of the variance
  - For each column

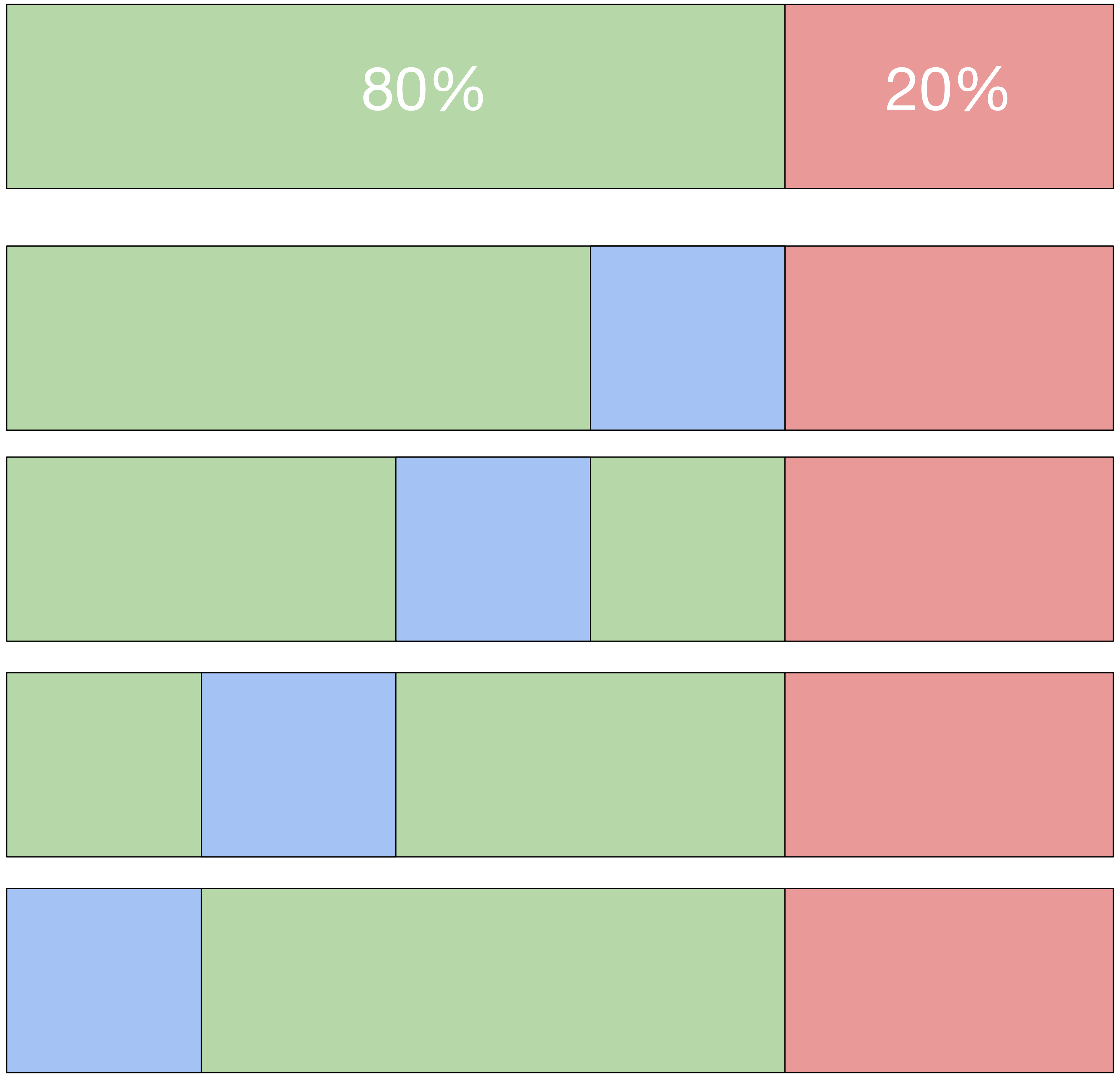
$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Apply normalization
  - To each column

$$x_i' = \frac{x_i - \mu}{\sigma}$$

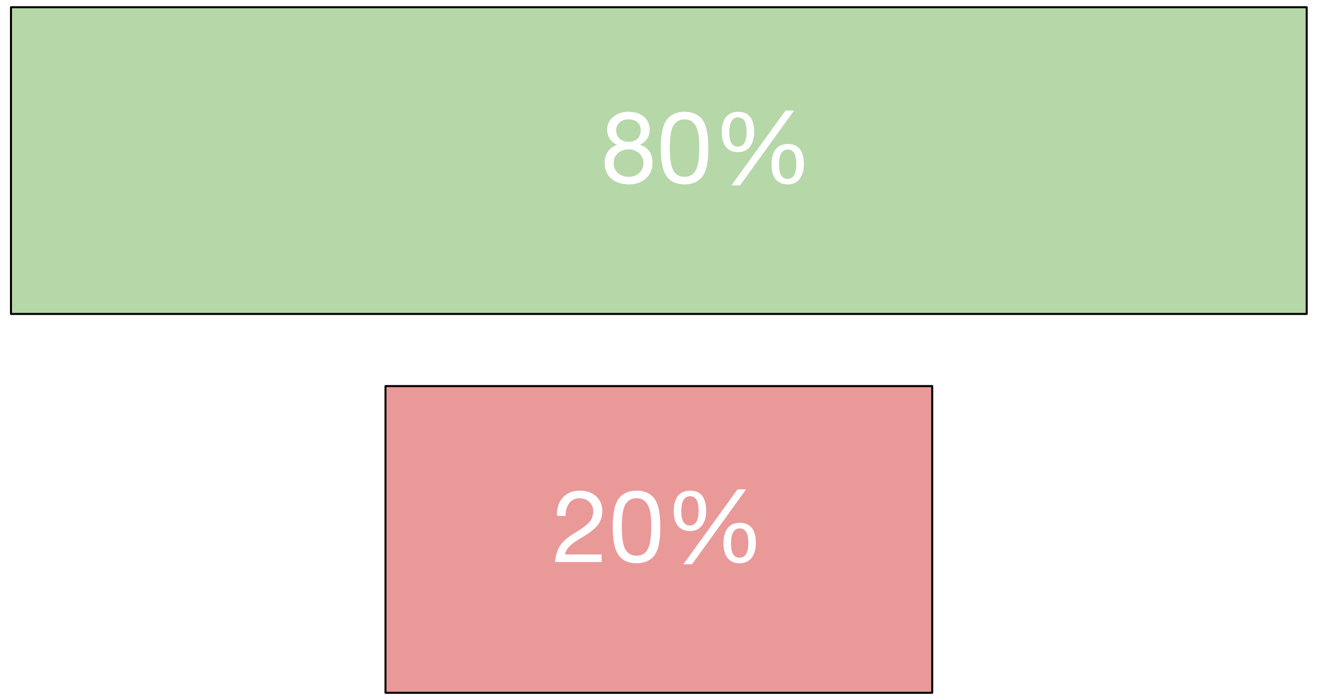
Training

- Holdout split 80-20%
  - Train/Test
- Train on (k-1)-folds
  - Performing
- Test on the validation fold
  - To each column



Grid

- A grid was implemented
  - To select the best parameters
- Once best parameters are found
  - Performing the training on the train dataset
- Evaluation on the test partition



Parallelization

- Preprocessing, labelling data as train/test
- Calculating residuals for normalization and shuffling
- Splitting data for k-Fold CV
- Predicting labels and computing weights in gradient descent
- Filtering predictions to compute confusion matrix
- Computing Gradient descent cost

Performances

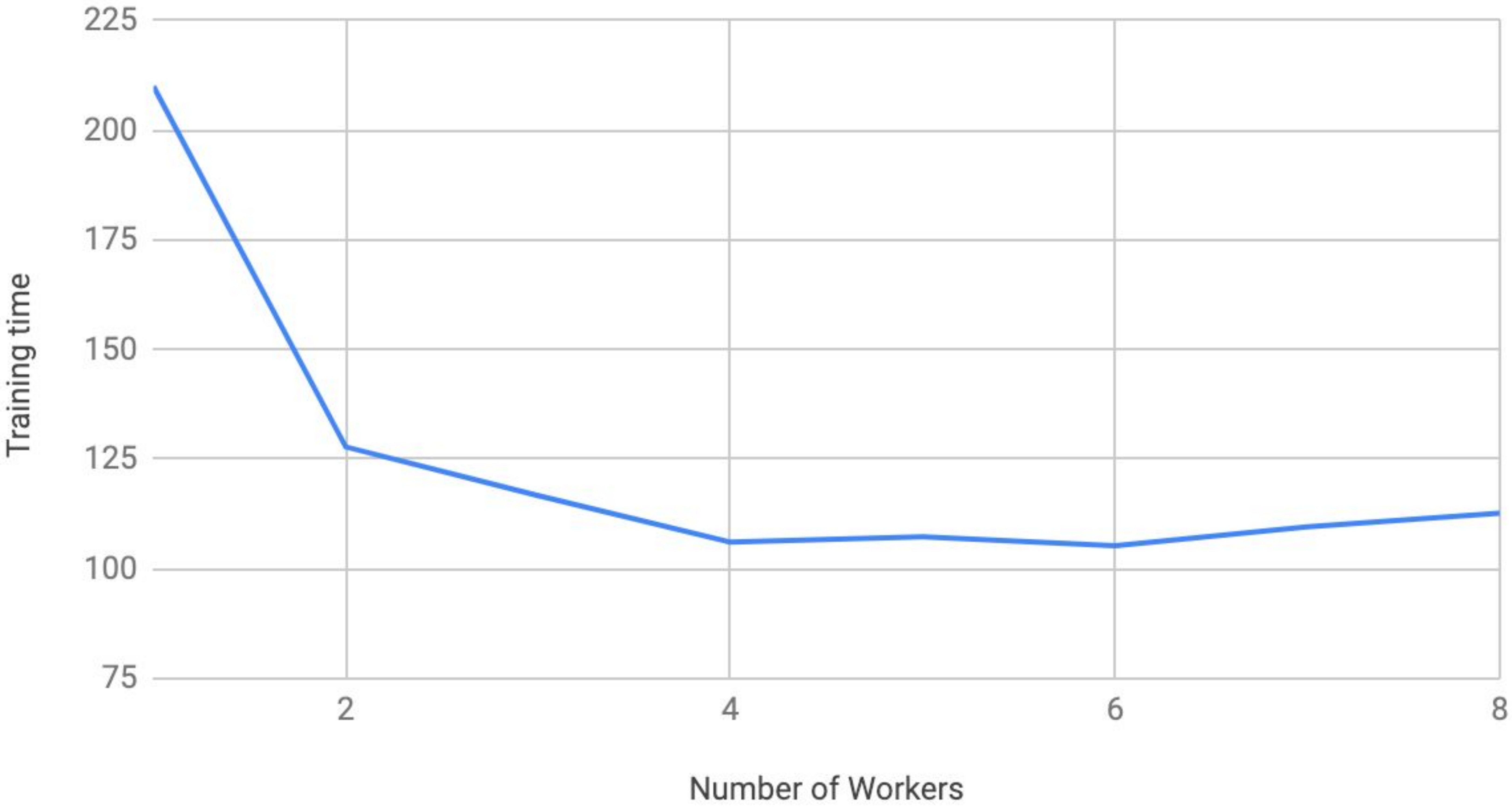
Algorithm Performance

True Positive 23.20%	False Positive 1.41%
False Negative 17.90%	True Negative 57.48%

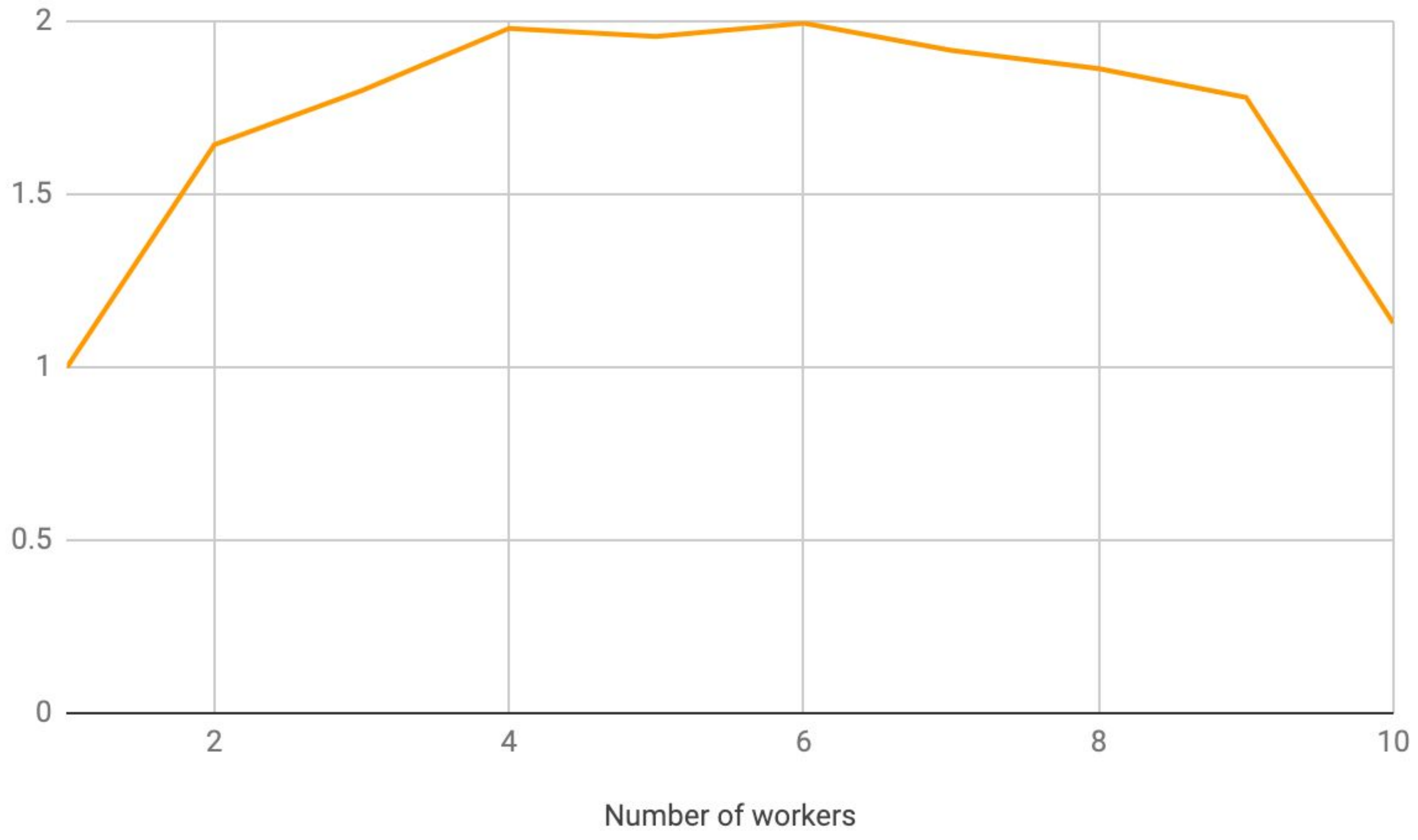
- Precision: 94.25%
- Recall: 56.44%
- F1-Score: 70.61%
- Accuracy: 80.68%

Holdout: 0.8 | (learning_rate, lambda_reg) = (0.36, 0.196) | Iterations: 50 | Threshold: 0.5 | Workers: 8

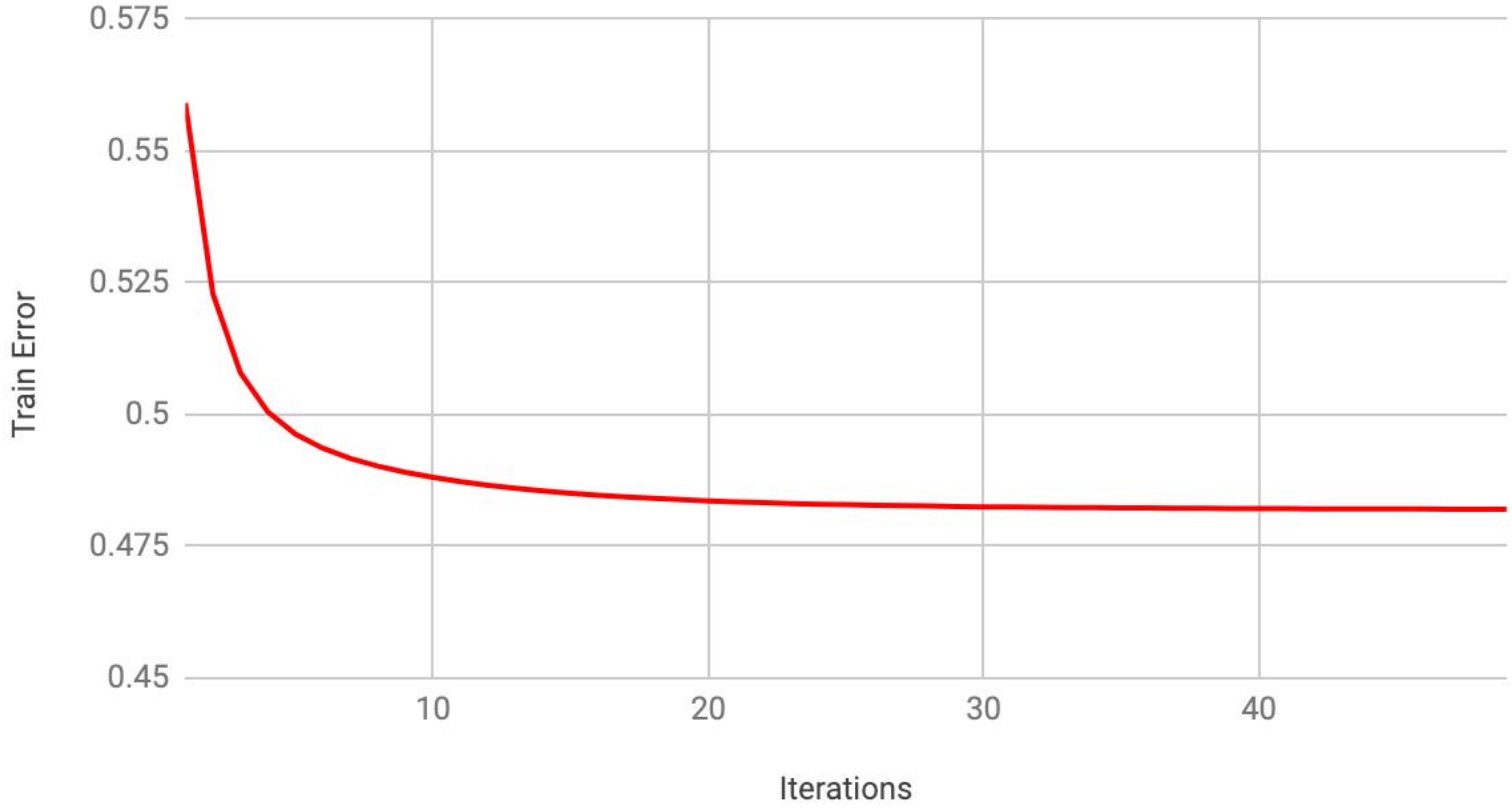
Time Performance



Speedup Curve



Test error



Conclusions

- Strength
    - No external library was used to store structure
    - SGD is well parallelized
    - Good accuracy, and small false positive classification
    - Good logging
- Weakness
    - Cross validation could be parallelized better

Thank you
Any questions?