

# Basics of Machine Learning

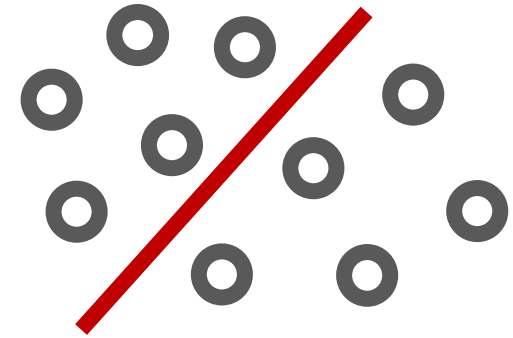
Dmitry Ryabokon, [github.com/dryabokon](https://github.com/dryabokon)





# Lesson 08

## Regression methods



# Linear regression

# Linear Regression

## Variable importance

Before starting linear regression check assumptions

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

# Linear Regression

## Variable importance

Following are the methods of variable selection you can use:

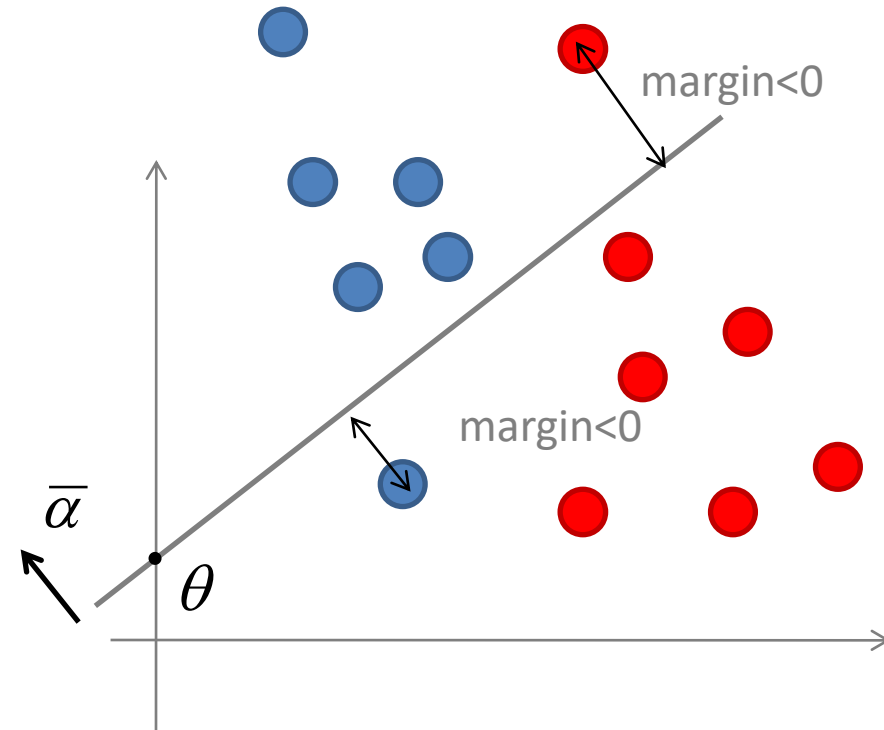
- Remove the correlated variables prior to selecting important variables
- Use linear regression and select variables based on p values
- Use Forward Selection, Backward Selection, Stepwise Selection
- Use Random Forest, Xgboost and plot variable importance chart
- Use Lasso Regression
- Measure information gain for the available set of features and select top n features accordingly.

# Logistic regression

# Logistic regression

**Target function: empirical risk = number of errors**

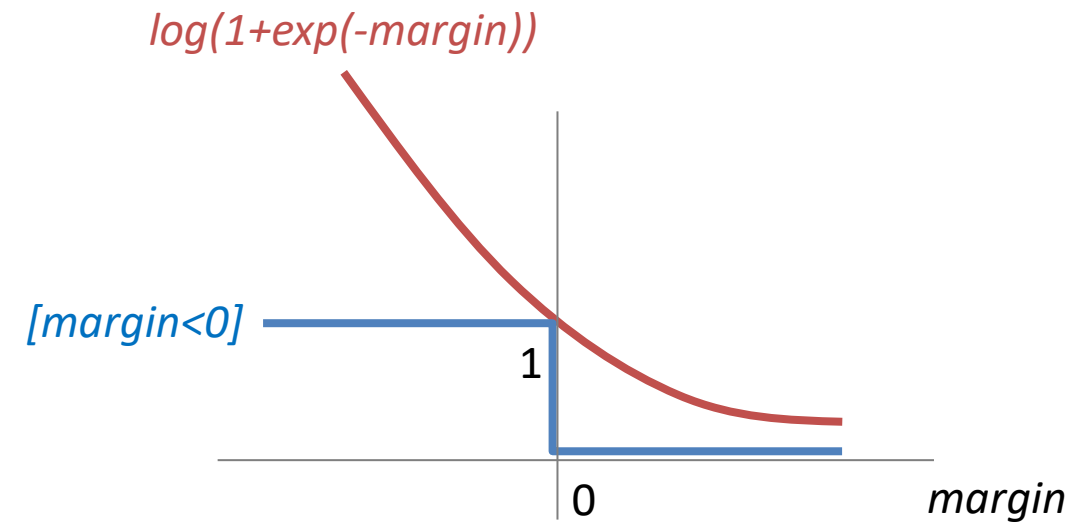
$$\#errors = \sum_{i=1}^n [margin(x_i, y_i) < 0] = \sum_{i=1}^n [y_i \cdot (a, x_i) < 0] \rightarrow \min$$



# Logistic regression

## Target function: upper bound

$$\sum_{i=1}^n [y_i \cdot (a, x_i) < 0] < \sum_{i=1}^n \log(1 + \exp(-y_i \cdot (a, x_i))) \rightarrow \min$$





# Logistic regression

## Relationship with $P(y|x)$

$$\sum_{i=1}^n [y_i \cdot \langle a, x_i \rangle < 0] < \sum_{i=1}^n \log(1 + \exp(-y_i \cdot \langle a, x_i \rangle)) \rightarrow \min$$
$$\sum_{i=1}^n \log\left(\frac{1}{1 + \exp(-y_i \cdot \langle a, x_i \rangle)}\right) \rightarrow \max$$

$$\sum_{i=1}^n \log(\text{sigmoid}(y_i \cdot \langle a, x_i \rangle)) \rightarrow \max$$

$$\sum_{i=1}^n \log(p(y_i|x_i)) \rightarrow \max$$

*model*

$$p(y_i|x_i) = \text{sigmoid}(y_i \cdot \langle a, x_i \rangle)$$

# Logistic regression

## Logistic function: the sigmoid

$$\text{sigmoid}(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

