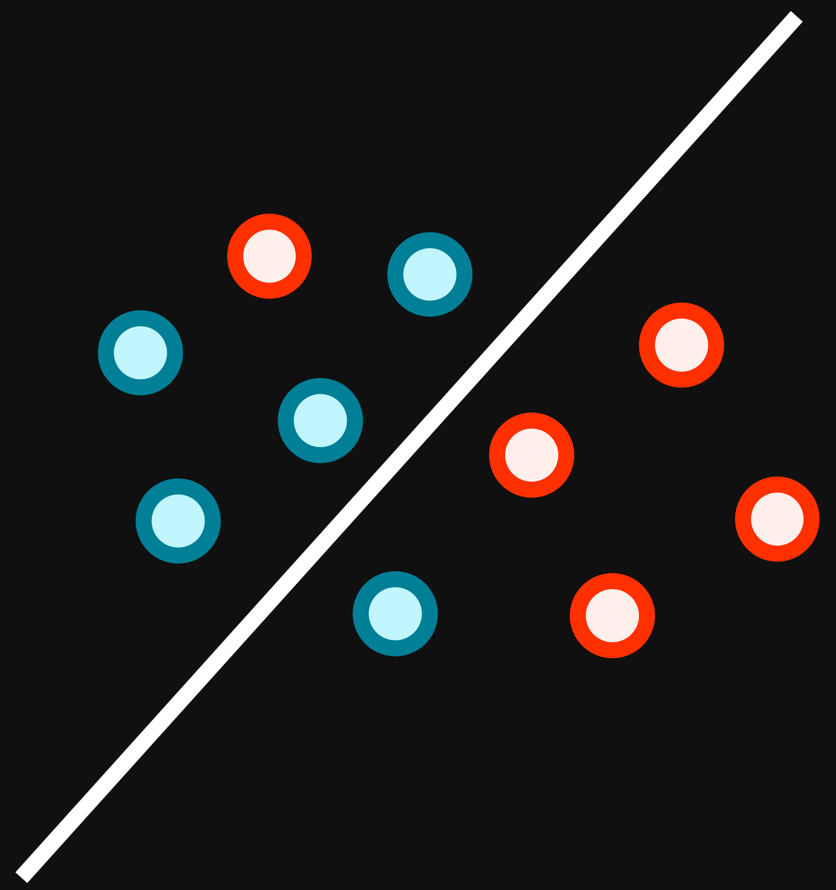
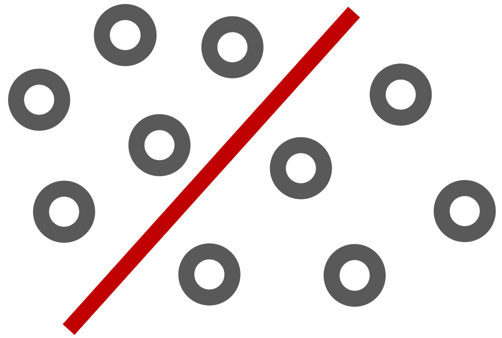
Basics of Machine Learning

Dmitry Ryabokon, github.com/dryabokon



Lesson 08
Regression methods



Linear regression

Variable importance

Before starting linear regression check assumptions

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

Variable importance

Following are the methods of variable selection you can use:

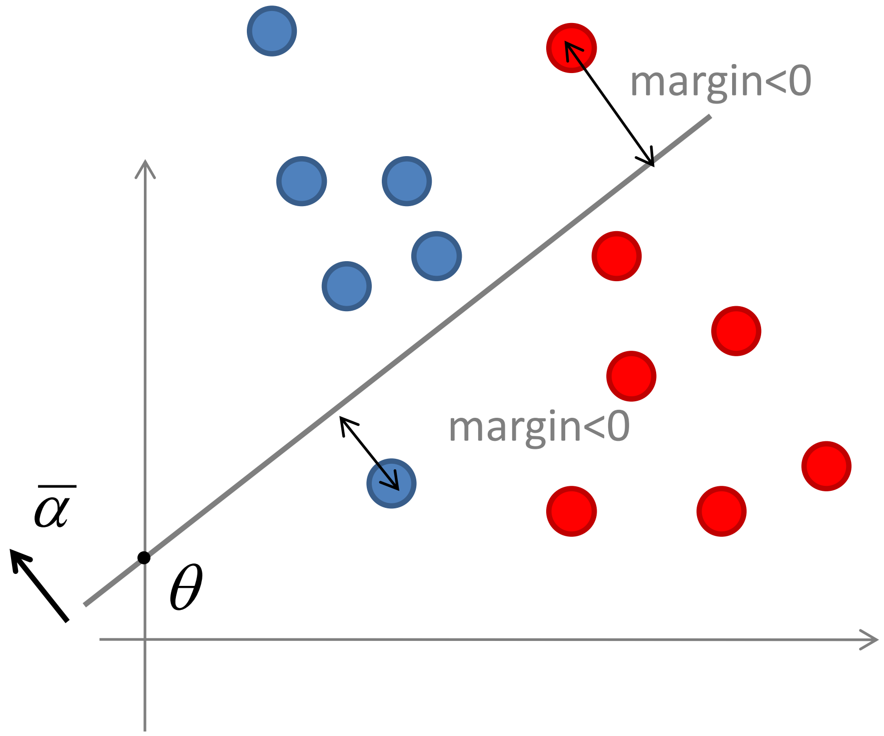
- Remove the correlated variables prior to selecting important variables
- Use linear regression and select variables based on p values
- Use Forward Selection, Backward Selection, Stepwise Selection
- Use Random Forest, Xgboost and plot variable importance chart
- Use Lasso Regression
- Measure information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

Logistic regression

Logistic regression

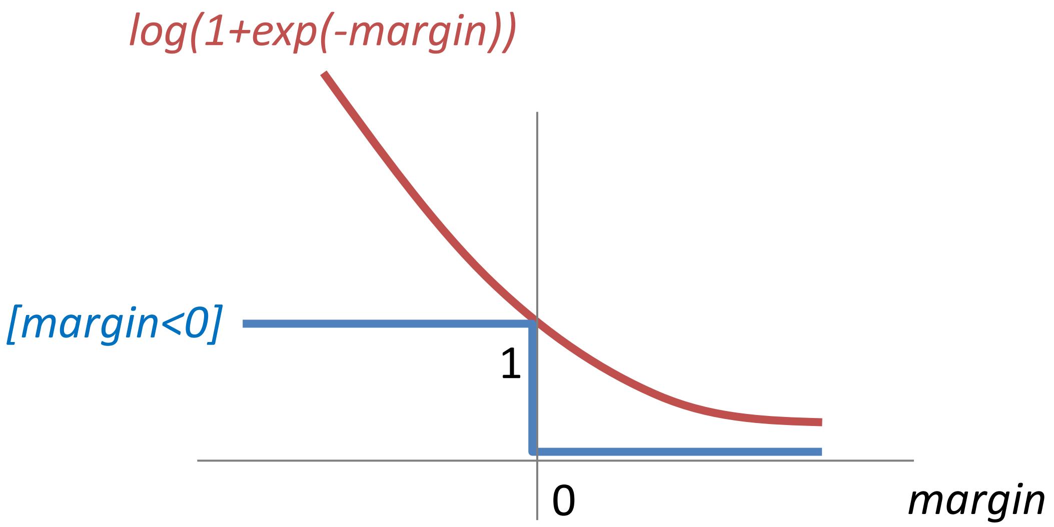
Target function: empirical risk = number of errors

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



Target function: upper bound

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



Logistic regression

Relationship with P(y|x)

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

$$\sum_{i=1}^{n}\log\left(\frac{1}{1+\exp(-y_i \cdot \langle a, x_i\rangle)}\right) \to \max$$

model

$$\boxed{p\left(y_i \middle| x_i\right) = sigmoid\left(y_i \cdot \langle a, x_i\rangle\right)}$$

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

$$\sum_{i=1}^{n}\log\left(p\left(y_i \middle| x_i\right)\right) \to \max$$

Logistic regression

Logistic function: the sigmoid

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