



AGH UNIVERSITY OF SCIENCE  
AND TECHNOLOGY

## **Seminar in *Artificial Intelligence***

**Marcin Zajac, Dominik Koza, Lukasz Gorczyca**

Department of Telecommunications

01.04.2019

# Agenda

- 1. Intro, presentation plan.
- 2. What is regression?
- 3. What regression is used for?
- 4. Types of regression.
- 5. Simple Linear regression.
- 6. Multiple dimension extension.
- 7. Ordinary least squares.
- 8. Gradient descent in linear regression.
- 9. Regularization - Ridge and Lasso
- 10. Logistic regression.
- 11. Polynominal regression.
- 12. QA
- 13. Quiz.

## 2. What is regression?

- looks for the relationship between two or more variables.

### 3. What regression is used for?

- used in: forecasting, MS Excel :D, machine learning...

## 4. Types of regression

- linear regression
- logistic regression
- polynomial regression
- stepwise regression
- ridge regression
- lasso regression
- elasticNet regression

## 5. Linear regression

- First known research in this area - method of least squares published by Legendre in 1805 and by Gauss in 1809
- The representation is a linear equation that combines a specific set of input values  $x$  the solution to which is the predicted output for that set of input values  $y$ . As such, both the input values  $x$  and the output value are numeric.

## 5. Simple Linear Regression

- Simple linear regression is a linear regression model with a single independent variable
- Model for single dimension

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (1)$$

- Naming:
  - The unknown parameters -  $\beta$
  - The independent variables -  $X$  or  $x$
  - The dependent variable -  $Y$  or  $y$
  - Introduced error -  $\epsilon$

## 6. Multiple dimension extension.

- When there is a single input variable  $x$ , the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.
- Model for  $n$  dimension

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_n x_{i,n} + \epsilon_i \quad (2)$$

- To matrix representation

$$y_i = x_i^T + \epsilon_i \quad (3)$$

$$Y = X\beta + \epsilon \quad (4)$$



## 7. Ordinary least squares

Method for estimating the unknown parameters in a linear regression model

$$\hat{\beta} = \operatorname{argmin}_{\beta} S(\beta) \quad (5)$$

$$S(\beta) = \sum_{i=1}^n \left| y_i - \sum_{j=1}^p x_{ij} \beta_j \right|^2 \quad (6)$$

$$\begin{bmatrix} n & \sum_{i=1}^n x_{2i} & \dots & \sum_{i=1}^n x_{ki} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{2i}^2 & \dots & \sum_{i=1}^n x_{2i} x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ki} & \sum_{i=1}^n x_{ki} x_{2i} & \dots & \sum_{i=1}^n x_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{2i} y_i \\ \vdots \\ \sum_{i=1}^n x_{ki} y_i \end{bmatrix}$$

## 7. Ordinary least squares (cont.)

Which leads to:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \hat{\beta} = [\mathbf{X}'\mathbf{X}]^{-1} \mathbf{X}'\mathbf{y}$$

## 8.Gradient descent in linear regression

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function.

Denote  $E$  as squared mean error in example for linear regression:

$$E = 1/n * \sum_{i=0}^n (y_i - \beta_1 * x_i + \beta_0) \quad (7)$$

Lets assume that:

$$a = \beta_1 \quad (8)$$

$$b = \beta_0 \quad (9)$$

$$D_a = \frac{d}{da} (E) \quad (10)$$

$$D_b = \frac{d}{db} (E) \quad (11)$$

## 8.Gradient descent in linear regression (cont.)

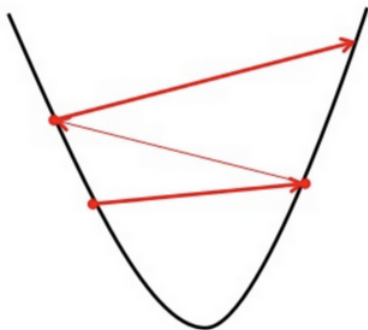
Now find that  $a$  and  $b$  where function  $E$  will reach minimum or be small enough. Take some  $L$  as learning rate and iterative find values of  $a$  and  $b$  from equations:

$$a = a - L * D_a \quad (12)$$

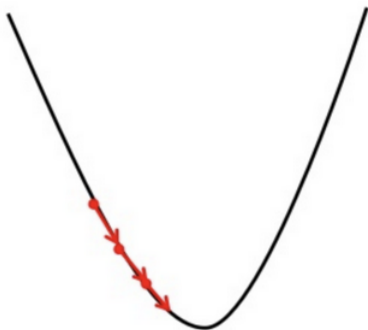
$$b = b - L * D_b \quad (13)$$

## 8.Gradient descent in linear regression (cont.)

Big learning rate



Small learning rate



Training rate  $L$  has to be small enough to avoid skipping minimum.

## 9.Regularization

Regularization methods provide a means to control our regression coefficients, which can reduce the variance and decrease our of sample error.

Two popular examples of regularization procedures for linear regression are:

- Lasso Regression - called L1 regularization.
- Ridge Regression - called L2 regularization.

These methods are effective to use when there is collinearity in your input values and ordinary least squares would overfit the training data.

## 9.Regularization

### Ridge Regression

Ridge should be used if we want to remain all parameters.

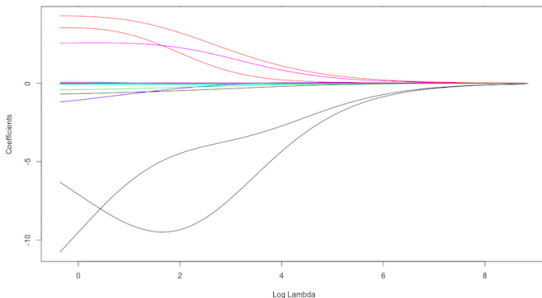
$$E = 1/n * \sum_{i=0}^n (y_i - \beta_1 * x_i + \beta_0) \quad (14)$$

Adding Ridge penalty:

$$\text{minimize}(E + \lambda \sum_{j=1}^p \beta_j^2) \quad (15)$$

## 9.Regularization (cont.)

### Ridge Regression

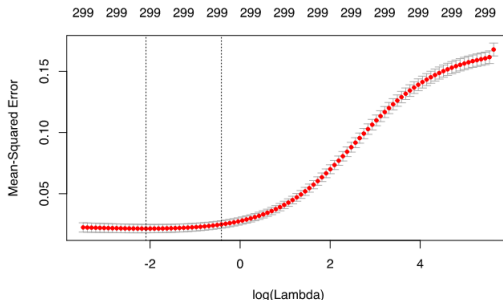


We see that Ridge remain all variables and with incrementing  $\lambda$  all are forced to 0. To find  $\lambda$  we can also perform CV. Example:



## 9.Regularization (cont.)

### Ridge Regression



In this case Ridge is not providing improvements. What about Lasso?

## 9.Regularization

### Lasso Regression

Lasso allow to get rid of some parameters.

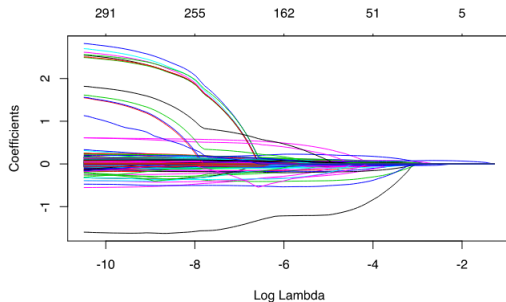
$$E = 1/n * \sum_{i=0}^n (y_i - \beta_1 * x_i + \beta_0) \quad (16)$$

Adding Lasso penalty:

$$\text{minimize}(E + \lambda \sum_{j=1}^p |\beta_j|) \quad (17)$$

## 9.Regularization (cont.)

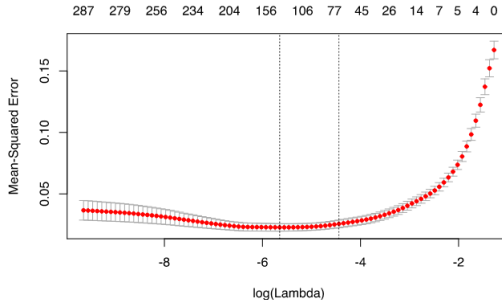
### Lasso Regression



How to find  $\lambda$ ? We can perform cross-validation. Example:

## 9.Regularization (cont.)

### Lasso Regression



There is improvement but some coefficients are equals to 0 which indicades that some parameters will not be taken into consideration.

## 10. Logistic regression

- Logistic regression was developed by statistician David Cox in 1958
- Logistic Regression is one of the basic and the most popular algorithm to solve a classification problem
- The main idea of logistic regression is to find a relationship between features and probability of particular outcome

## 10. Logistic regression

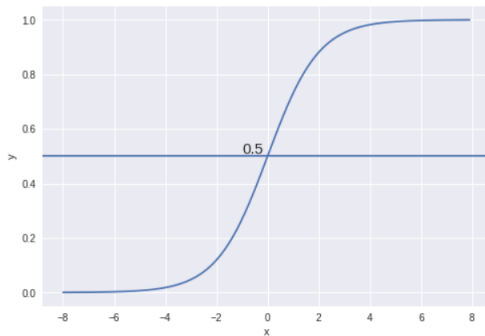
- In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, etc.) or 0 (FALSE, failure, etc.)
- Linear regression predicts values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)

## 10. Logistic regression

Function graph:

- For the beginning arguments values are nearly 0 or 1
- After reaching critical value we see dynamically increase/decrease of function value
- For the final arguments values are nearly 1 or 0 (opposite to the beginning)

## 10. Logistic regression (cont.)



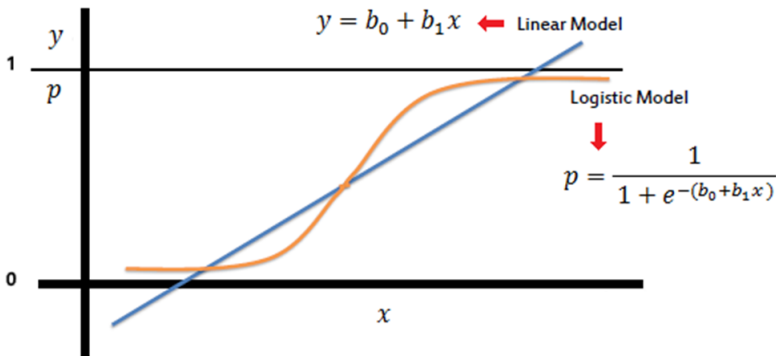
sigmoid function



## 10. Logistic regression

- The biggest difference between linear and logistic regression is how the line is fit to the data
- Logistic regression uses maximum likelihood estimation (MLE) to get the model coefficients

## 10. Logistic regression (cont.)



## 10. Logistic regression

Why should we use logistic regression instead of linear?

- Let  $\mathbf{x}$  be some feature and  $\mathbf{y}$  be the output which can be either 0 or 1.
- The probability that the output is 1 can be represented as:

$$P = (y = 1 | \mathbf{x}) \quad (18)$$

- Using linear regression we will get:

$$p(X) = \beta_0 + \beta_1 X \quad (19)$$

- Logistic regression is expressed by logit function, therefore:

$$p(X) = e^{\beta_0 + \beta_1 X} / (1 + e^{\beta_0 + \beta_1 X}) \quad (20)$$

## 10. Logistic regression

Different types of logistic regression:

- 1. Binary: The categorical response has only two possible outcomes (e.g.: Spam or Not)
- 2. Multinomial: Three or more categories without ordering. (e.g.: Predicting which food is preferred more (Veg, Non-Veg, Vegan))
- 3. Ordinal: Three or more categories with ordering. (e.g.: Movie rating from 1 to 5)

## 10. Logistic regression

### Summing up

- Logistic regression's ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular machine learning method.

## 11. Polynominal Regression

- Occurs when regression equation has independent variable in power higher than 1.
- General equation

$$y = \beta_0 + \beta_1 * x_i + \beta_2 * x_i^2 + \beta_3 * x_i^3 + \dots + \beta_m * x_i^m + E, i = 1, 2, 3 \dots \quad (21)$$

- Example (variable in 2 power)

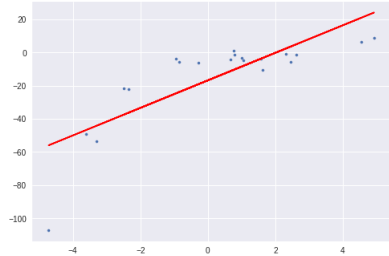
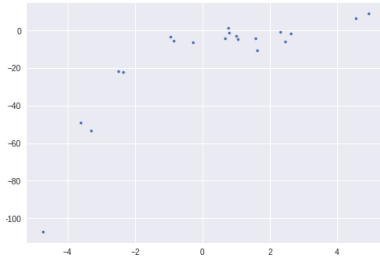
$$y = a + b * x^2 \quad (22)$$

- the best fit is rather curve not a straight line

## 11. Polynomial Regression

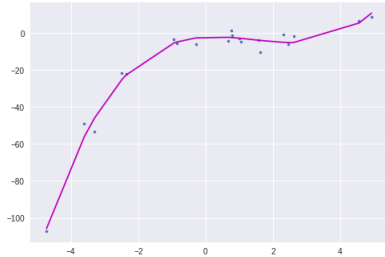
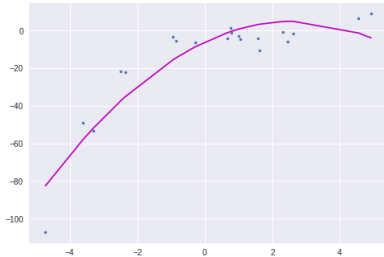
- polynomial with higher degree can give us lower error rate
- if degree will be too high then overfitting will occur
- curve should fit the nature of the problem (trend) not every single sample
- nonlinear "relationship" of variables, but it is considered as linear model due to the coefficients/weights associated with the features are still linear.

# 11. Polynominal Regression

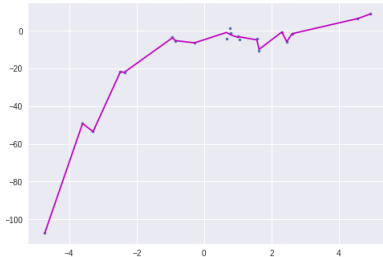




# 11. Polynominal Regression



# 11. Polynominal Regression



**Thank you for your  
attention!**

# Q & A