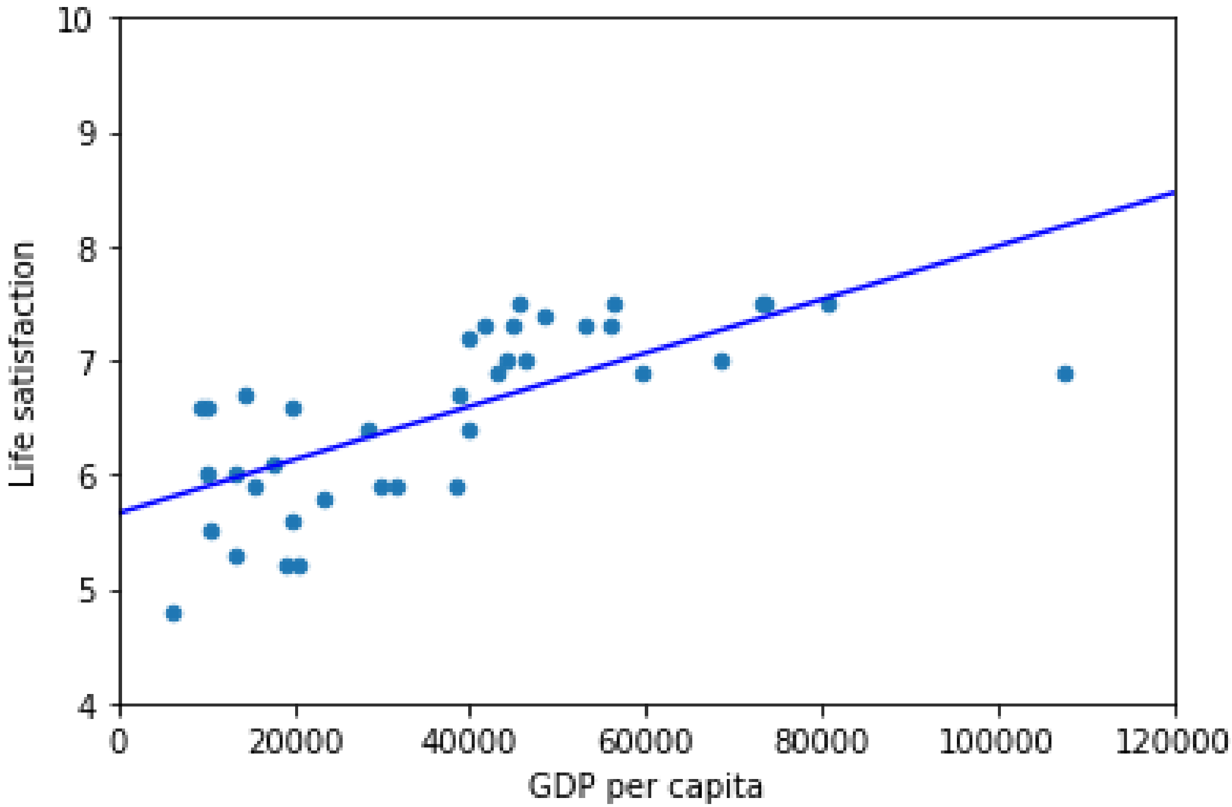
Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Training Machine Learning Models

Andrey Sozykin
Andrey.Sozykin@urfu.ru

Linear Regression



$y = a * x + b,$

a, b – parameters of the model

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

What is training?

Training a model means setting its parameters so that the model best fits the training set.

Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Model Metrics

Training a model means setting its parameters so that the model **best fits** the training set.

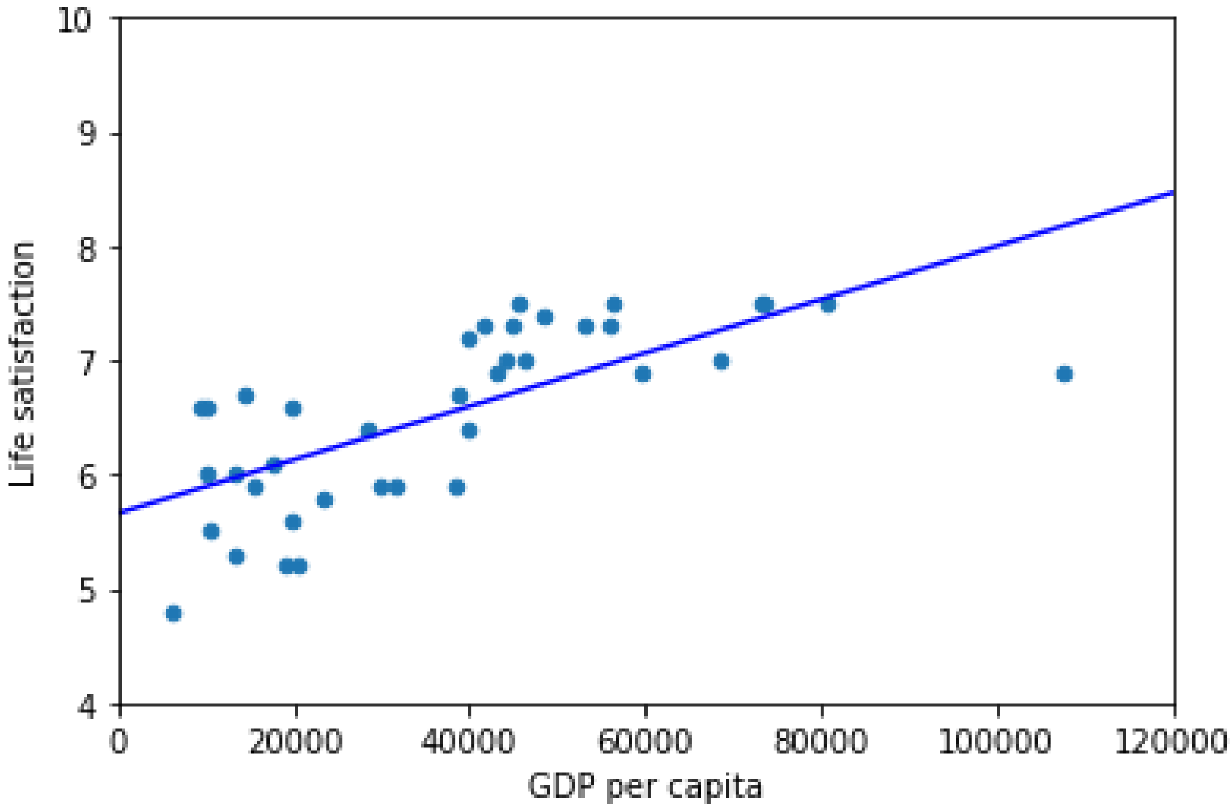
Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Mean Square Error

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - c_i)^2$$

Linear Regression



y = *a* * x + *b*,

a, *b* – parameters of the model

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

MSE for Linear Regression

$$MSE(X) = \frac{1}{m}\sum_{i=1}^{m}(a \cdot x_i + b - c_i)^2$$

Ural Federal University named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

MSE for Linear Regression

$$y = w_0 \cdot 1 + w_1 \cdot x_1 + \cdots + w_n \cdot x_m$$

$$y = \boldsymbol{w} \cdot \boldsymbol{x}$$

$$y = \boldsymbol{w}^T \boldsymbol{x}$$

$$MSE(X) = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}^T \boldsymbol{x}_{\boldsymbol{i}} - \ c_i)^2$$

How to find **w**?

$$MSE(X) = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^T \boldsymbol{x_i} - c_i)^2$$

How to find a vector ***w*** that minimize *MSE(x)?*

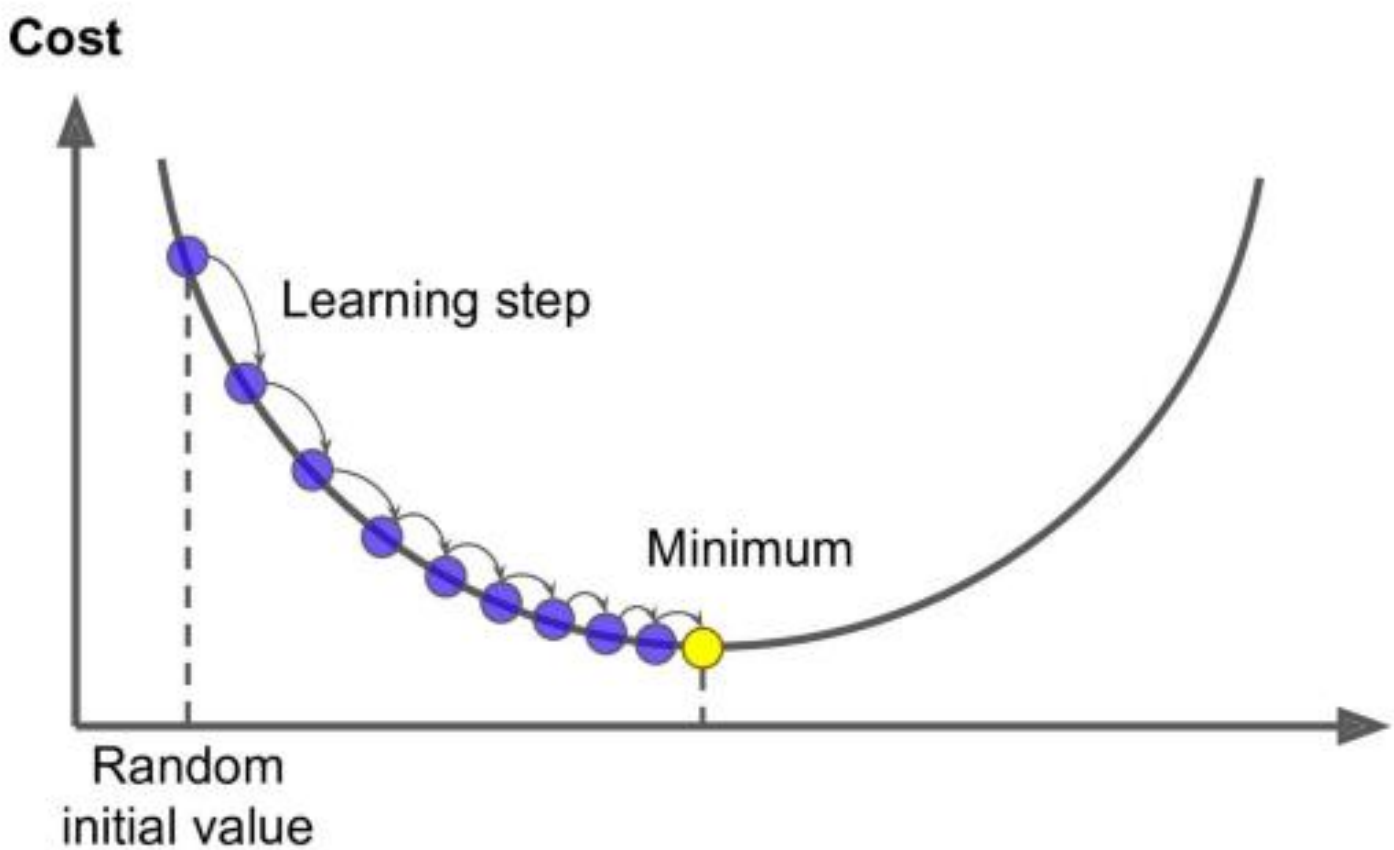
Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

The Normal Equation

$$w = (X^T X)^{-1} X^T c$$

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Gradient Descend



Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

Derivative of MSE

$$MSE'(X) = \frac{1}{m}\sum_{i=1}^{m}(a \cdot x_i + b \ - \ c_i)^2$$

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Chain Rule

$$MSE'(X) = \frac{1}{m}\sum_{i=1}^{m}(a \cdot x_i + b - c_i)^2$$

$$f'(g(x)) = f'(g(x))g'(x)$$

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Derivative of MSE

$$MSE'(X) = \frac{1}{m}\sum_{i=1}^{m}(a \cdot x_i + b \; - \; c_i)^2$$

$$f'(g(x)) = f'(g(x))g'(x)$$

$$MSE'(X) = \frac{2}{m}\sum_{i=1}^{m}(a \cdot x_i + b \; - \; c_i)x_i$$

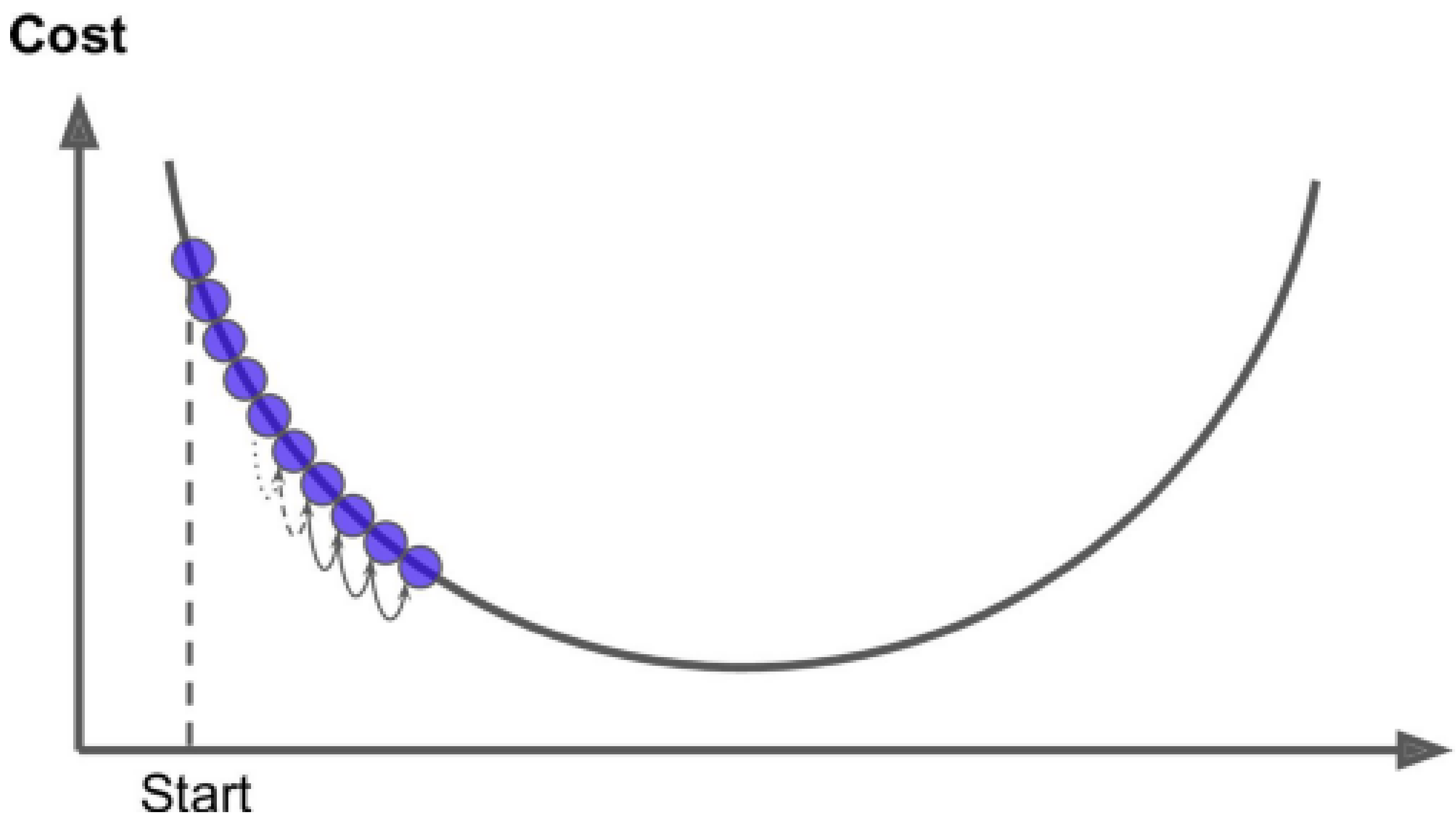
Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Gradient Descent

$$w_{new} = w - \eta \cdot MSE'(x)$$

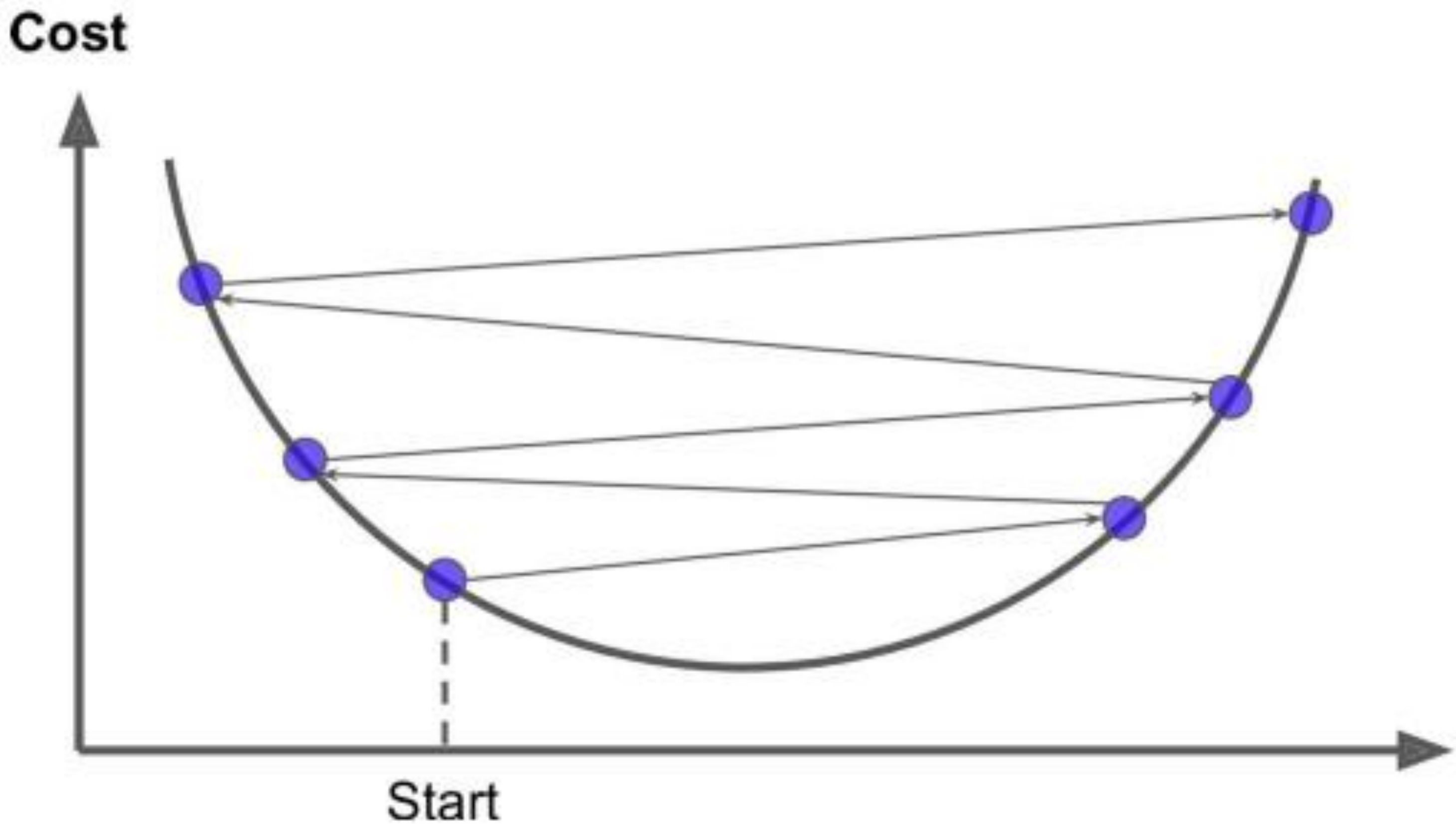
Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Learning Rate



Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

Learning Rate



Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Multiple Features

$$\nabla_w MSE(w) = \begin{bmatrix} \frac{\partial}{\partial w_0} MSE(w) \\ \frac{\partial}{\partial w_1} MSE(w) \\ \frac{\partial}{\partial w_2} MSE(w) \\ \dots \\ \frac{\partial}{\partial w_n} MSE(w) \end{bmatrix} = \frac{2}{m} X^T (Xw - c)$$

Multiple Features

$$\boldsymbol{w}_{new} = \boldsymbol{w} - \eta \cdot \nabla_w MSE(w)$$

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

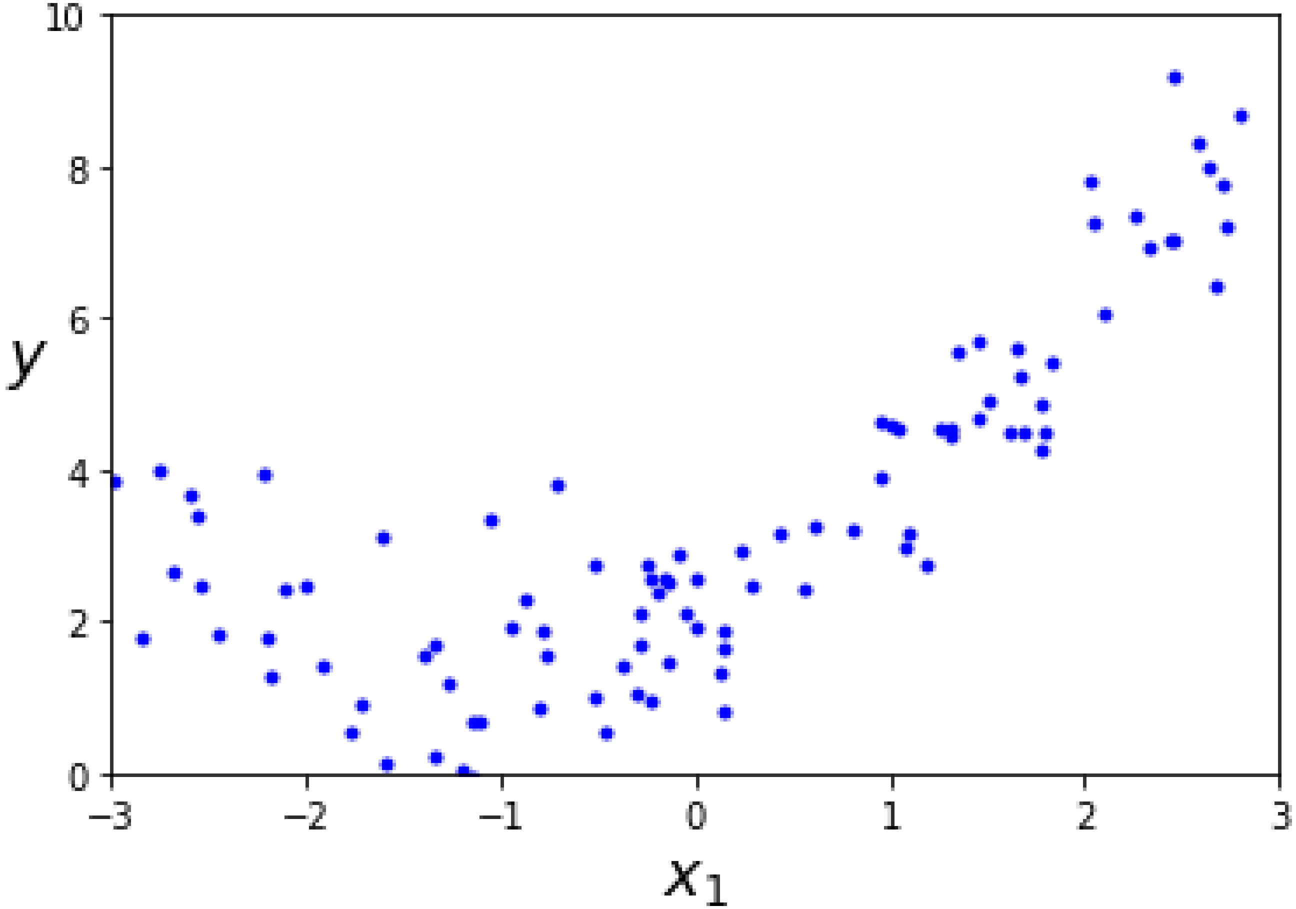
Types of Gradient Descent

Batch Gradient Descent

Stochastic Gradient Descent

Mini-Batch Gradient Descent

How to deal with non-linear data?



Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Polynomial Regression

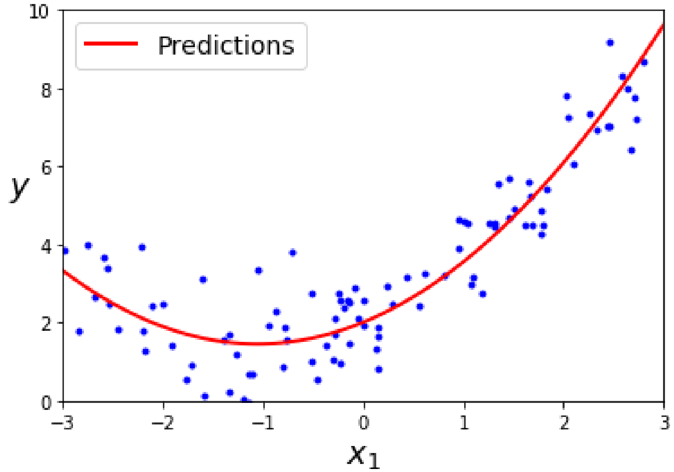
Add powers of existing features as a new features

- One feature: x
- New features: x^2, x^3, x^4, *etc.*

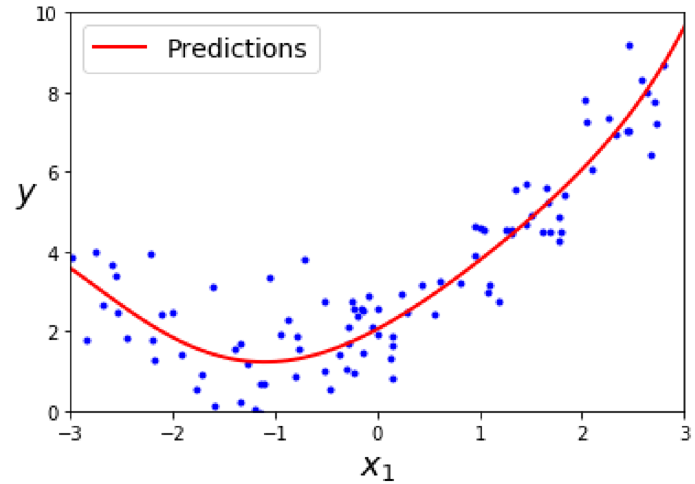
Train a linear model on the extended set of features

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

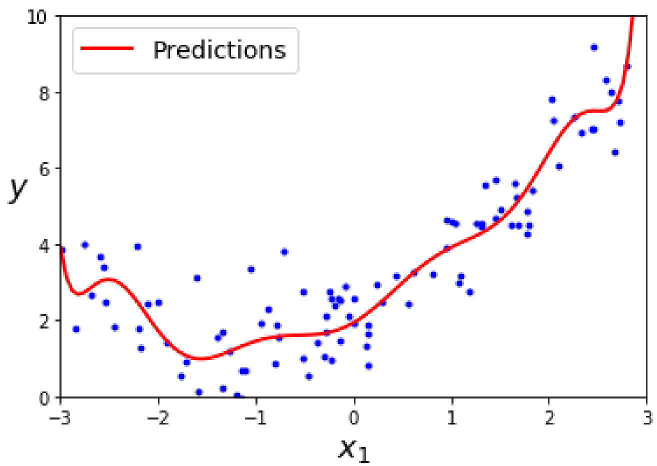
Overfitting



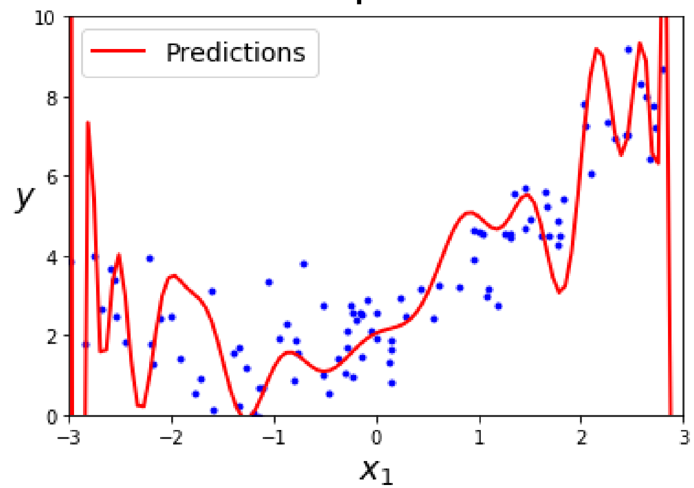
p=2



p=5



p=10



p=30

Overfitting

Overfitting – the models adapts to the training set instead of searching the general patterns in data

Poor generalization – the model works well with data from the training set, but bad with additional data

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Conclusions

Training a model means setting its parameters so that the model best fits the training set.

Metrics (cost functions) are used to determine how good the model fits the data

Overfitting – the models adapts to the training set instead of searching the general patterns in data

Ural Federal University
named after the first President of Russia B.N.Yeltsin
Institute of radioelectronics and information technologies

Thank you!