Анализ данных с использованием языка программирования R

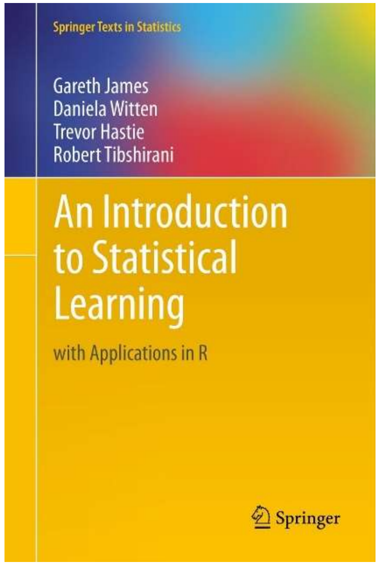
Тема 4
Основы машинного обучения. Регрессионный анализ данных

Минюкович Екатерина Александровна
к.э.н., доцент

miniukovich@bsu.by

Reference

An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, http://www-bcf.usc.edu/~gareth/ISL/ (available online for free)



Reference

Introduction to Machine Learning with R
by Dr. Dimitrios Gouliermis

http://www.mpia.de/homes/dgoulier/MLClasses/Course%20-%20Introduction%20to%20Machine%20Learning%20for%20Scientists%20with%20R.html

Reference

H2O documentation

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html

Supervised vs. Unsupervised Learning

Supervised	Unsupervised
Data: 1) n observations; 2) p variables X1, X2, . . .,Xp, measured on each observation; 3) response Y measured on same n observations Y Continuous: Regression Discrete: Classification	Data: 1) n observations; 2) p variables X1, X2, . . .,Xp, measured on each observation Clustering…

Regression / Classification Problem

Steps to solve

- *Working with data*

- *Modeling*

Modeling

- Choose a class of model
- Fit the model to data
- Validate the model and optimize hyperparameters
- Predict for unknown data

Mathematical model

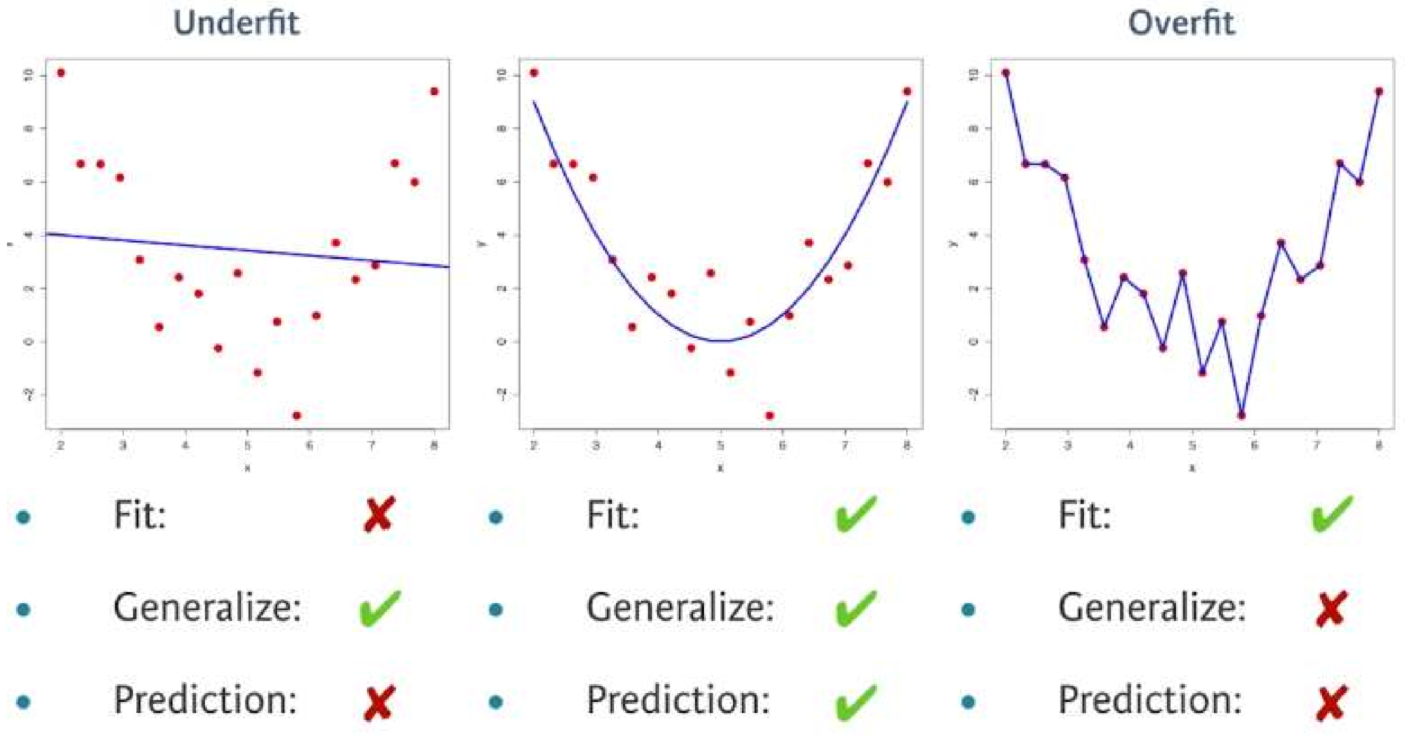
$$\boxed{Y = f(X) + \epsilon}$$

f is some fixed but unknown function of *X*1, . . . , *Xp*, and *e* is a random *error term*, which is independent of *X* and has mean zero. In this formulation, *f* represents the *systematic* information that *X* provides about *Y* .

We can predict *Y* using our estimate for *f*

$$\boxed{\hat{Y} = \hat{f}(X)}$$

Bias-Variance Trade-Off



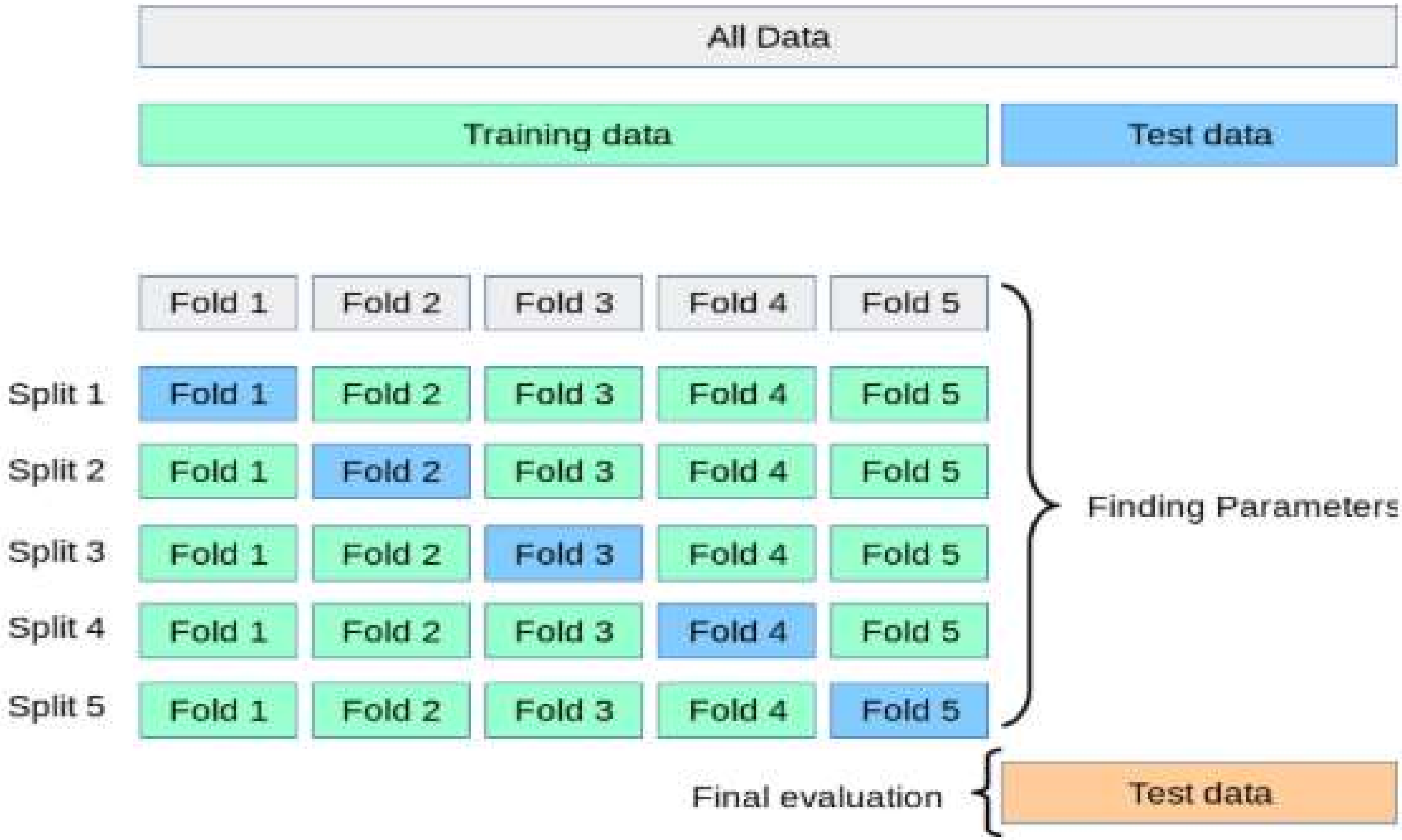
Underfitting *(high bias)* - algorithm is missing the relevant relations between features and target outputs

Overfitting *(high variance)* - modeling the random noise in the training data, rather than the intended outputs.

Model validation

- **Data**
  - train + test (e.g. 75% + 25%)
  - train + valid + test (e.g. 60% + 20% + 20%)
  - train with cross-validation + test (e.g. 80% + 20%)
- **Metrics**
  - Regression: **R^2**, MSE, MAE,…

Model validation via cross-validation



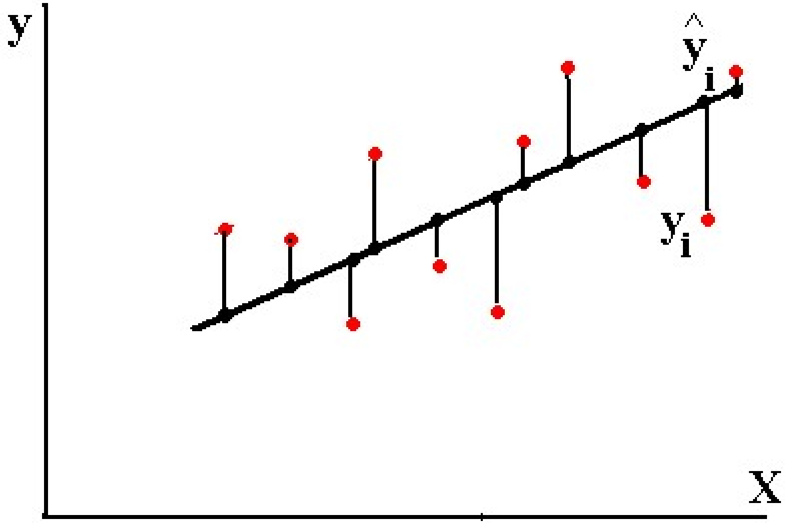
Some models for Regression in h2o

- Generalized Linear Model (GLM)

examples in regression_1.R, regression_2.R

- Ensemble methods
  - Distributed Random Forest (DRF)
  - Gradient Boosting Machine (GBM)
  - Stacked Ensembles

Linear Regression with one variable



(x_i, y_i), $i=1,n$ – number of observations (red points)

$$\hat{y} = ax + b$$

$$\hat{y} = \theta_0 + \theta_1 x_1 = \theta_0 x_0 + \theta_1 x_1, \qquad x_0 = 1$$

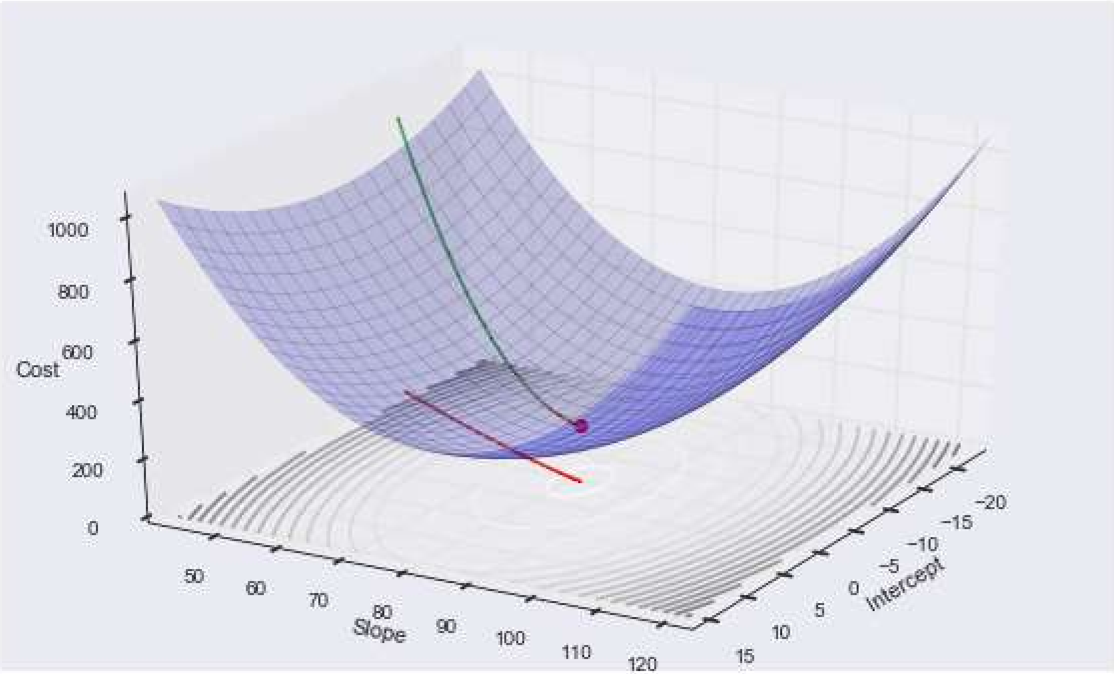
θ_0 - *intercept*, θ_1 - *slope*

The method of least squares

$$Cost = J(\theta_0, \theta_1) = \sum_{i=1}^{n} (\widehat{y^i} - y^i)^2 = \sum_{i=1}^{n} (\theta_0 x_0^i + \theta_1 x_1^i - y^i)^2$$

Our aim - $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Gradient descent to find $\min_{\theta_0,\theta_1} J(\theta_0, \theta_1)$



Need to choose

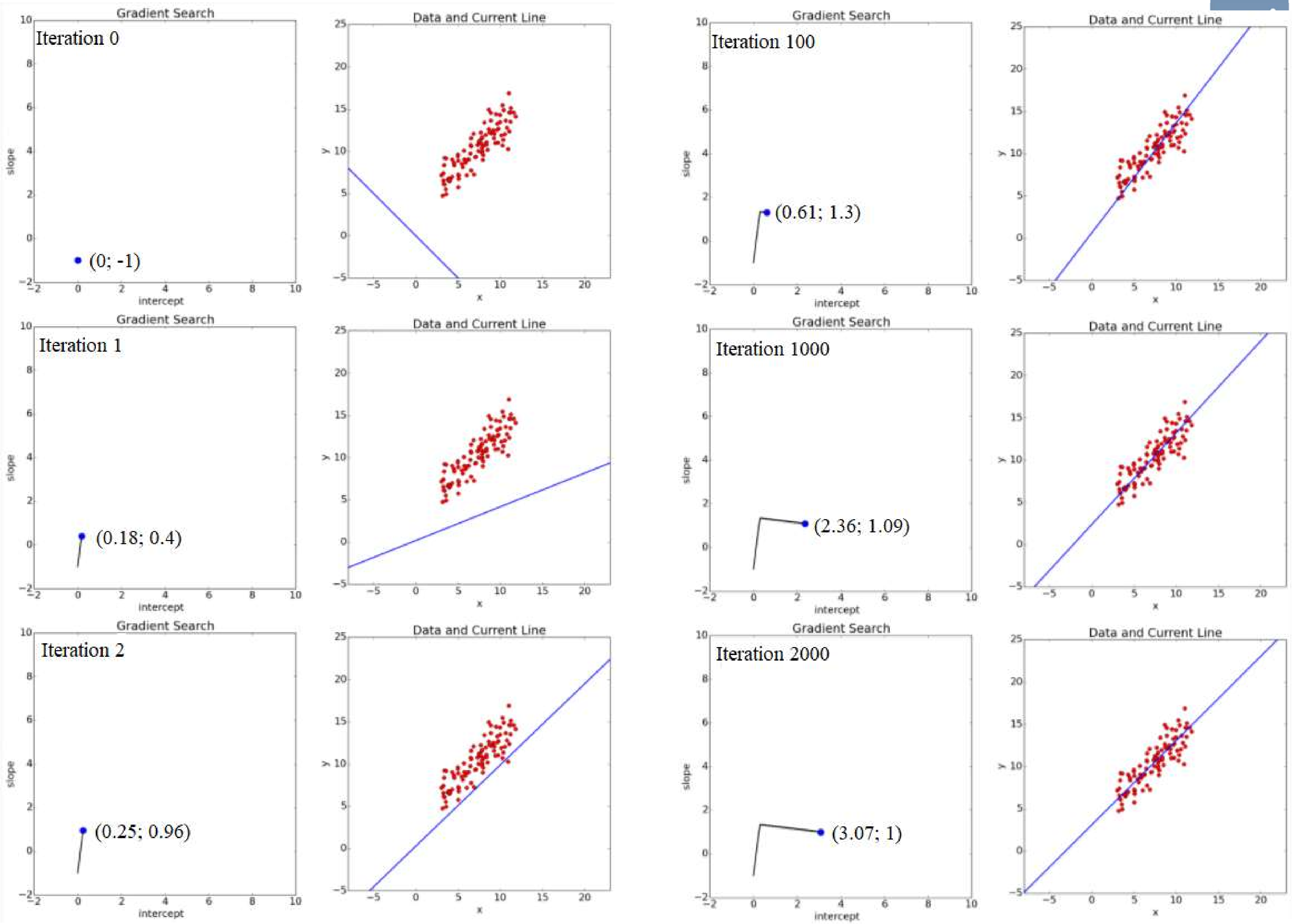
$\boldsymbol{\alpha}$ – *learning rate (step size)*
$(\boldsymbol{\theta_0}, \boldsymbol{\theta_1})$ - *start point*

Repeat until convergence

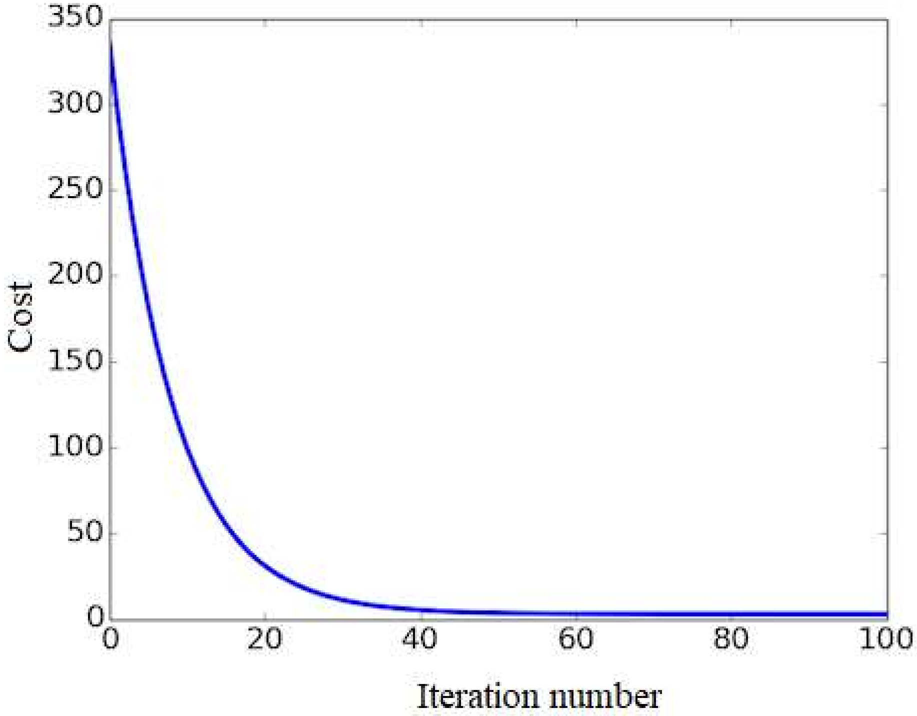
$$\theta_0 = \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \theta_0 - 2\alpha \sum_{i=1}^{n} \left(\theta_0 x_0^i + \theta_1 x_1^i - y^i\right) x_0^i$$

$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \theta_1 - 2\alpha \sum_{i=1}^{n} \left(\theta_0 x_0^i + \theta_1 x_1^i - y^i\right) x_1^i$$

Gradient descent (example)



Gradient descent (example)



Linear Regression with multiple variables

m *variables,* ***n*** *observations*

$$\hat{y} = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m, \qquad x_0 = 1$$

$$X = [1, x_1, \dots, x_m] \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_m \end{bmatrix} \qquad \hat{\boldsymbol{y}} = \boldsymbol{h_\theta}(\boldsymbol{X}) = \boldsymbol{X\theta}$$

Dataset for training: $X^{(i)} = \left[1, x_1^{(i)}, \dots, x_m^{(i)}\right], y^{(i)}, i = 1, \dots, n$

$$\boldsymbol{Cost} = \boldsymbol{J}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\boldsymbol{h_\theta}(\boldsymbol{X}^{(i)}) - \boldsymbol{y}^{(i)})^2 \qquad \textbf{\textit{Our aim}} \text{ - } \min_\theta J(\theta)$$

Repeat until convergence:

for j=0…m

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) = \theta_j - 2\alpha \sum_{i=1}^{n} \left(h_\theta\left(X^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

Cost functions

***m** variables, **n** observations*

$$\hat{y} = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m, \qquad x_0 = 1$$

$$X = [1, x_1, \ldots, x_m] \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_{01} \\ \ldots \\ \theta_m \end{bmatrix} \qquad \hat{\boldsymbol{y}} = \boldsymbol{h_\theta}(\boldsymbol{X}) = \boldsymbol{X\theta}$$

GLM (*Gaussian regression*) $Cost = J(\theta) = \sum_{i=1}^{n} (X^{(i)}\theta - y^{(i)})^2$

Regularization: $Cost = Cost + Penalty$

Ridge (*regularization **l2***) $Penalty = \sum_{j=0}^{m} {\theta_j}^2$

Lasso (*regularization **l1***) $Penalty = \sum_{j=0}^{m} |\theta_j|$

Elastic net (*combines **l1** and **l2***) $Penalty = \lambda * ((1-\alpha) * l2 + \alpha * l1)$

Regression metrics

R^2 score, the coefficient of determination

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2}, \qquad where \;\; \bar{y} = \frac{\sum_{i=1}^{n} y^{(i)}}{n}$$

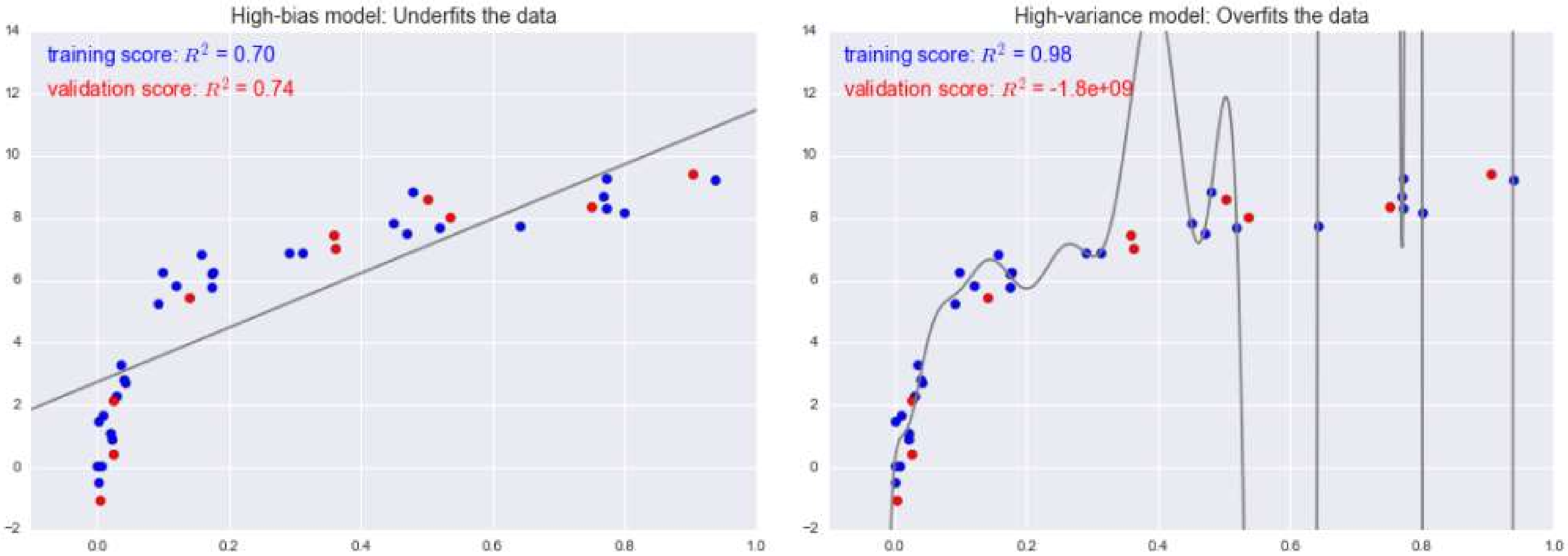
Mean squared error

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

Mean absolute error

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|y^{(i)} - \hat{y}^{(i)}\right|$$

Bias-Variance Trade-Off



- For **high-bias** models, the **performance** of the model on the **validation** set is **similar** to the performance on the **training set** *(but the performance is worse than for appropriate fitting)*.

- For **high-variance** models, the **performance** of the model on the **validation** set is **far worse** than the performance on the **training** set.

Bias-Variance Trade-Off

What to do in case of high-bias or high variance?

Change

- Model complexity (e.g. via regularization)
- Quantity of training samples
- Set of features

Reading

Andrew Ng ***ML: Advice for Applying Machine Learning***

Ways to fix high bias/variance in linear models

High bias (underfitting)	High variance (overfitting)
• Add more features • Add polynomial features	• More training examples • Smaller set of features • Use regularization • Increase regularization strength (coefficient)

Practise

regression_1.R
regression_2.R

Managed Independent Work

pr_regression.R

Choose the best GLM model for Boston ds

Models	R^2	
	train	test
Without regularization		
• for all predictors		
• for predictors with p value <= 0.05		
With regularization (best alpha from grid …)		
Polynomial features of degree 2		
• without regularization		
• with regularization (default)		
• with regularization (best alpha … and lambda … from grid)		