Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression

Krisztian Buza

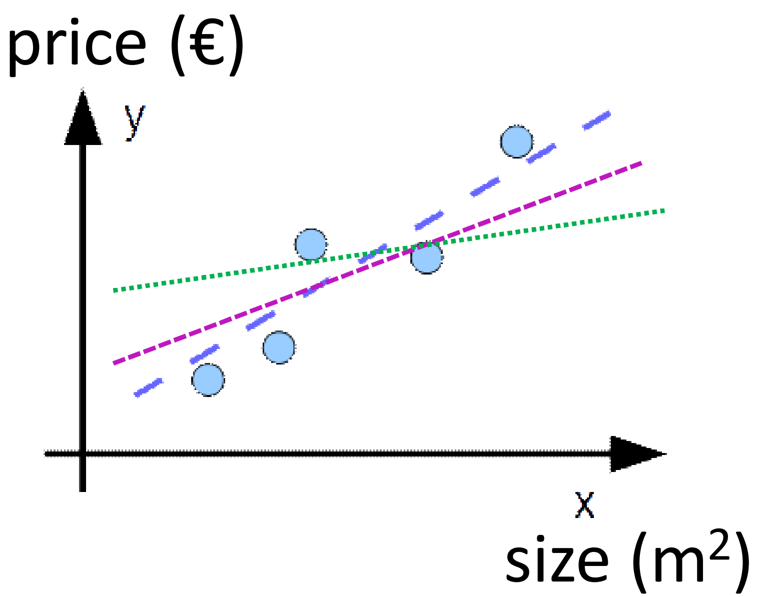
Department of Artificial Intelligence
Eötvös Loránd University
Budapest, Hungary

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Motivating Example

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Example: Estimation of House Prices



Which line fits the data best?

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Basic Concepts and Notations

m features

header

$x^{(1)}$	$x^{(2)}$	...	$x^{(m)}$	y
$x_1^{(1)}$	$x_1^{(2)}$	...	$x_1^{(m)}$	y_1
$x_2^{(1)}$	$x_2^{(2)}$	...	$x_2^{(m)}$	y_2
...	...	...	...	...
$x_i^{(1)}$	$x_i^{(2)}$	...	$x_i^{(m)}$	y_i
...	...	...	...	...
$x_n^{(1)}$	$x_n^{(2)}$	...	$x_n^{(m)}$	y_n

i-th instance

n

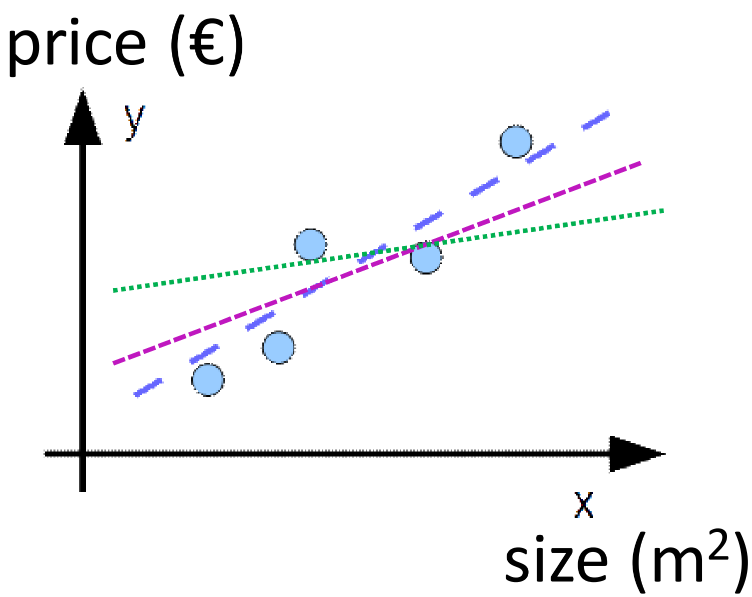
label

- Data table
- instance (observation, object, row)
- feature (attribute, column, variable)
- label (class label, target)

- labeled data
- unlabeled data

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Example: Estimation of House Prices



We may use RMSE to measure how well a model fits the data:

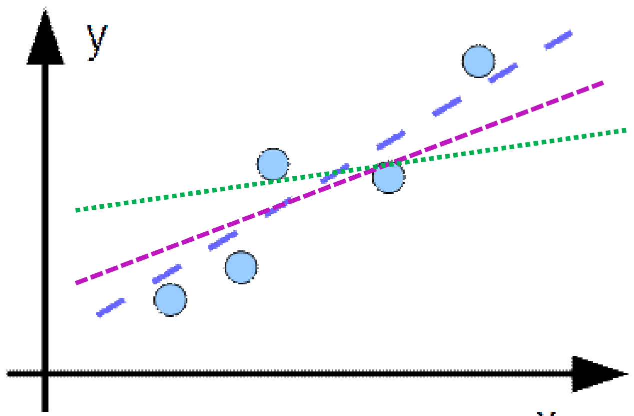
$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}}$$

$\hat{y}_i$ = predicted label of the i-th instance

y_i = true label of the i-th instance

n = number of instances

Example: Estimation of House Prices



Models we consider:

$\hat{y} = w_0 + w_1\, x$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{n}}$$

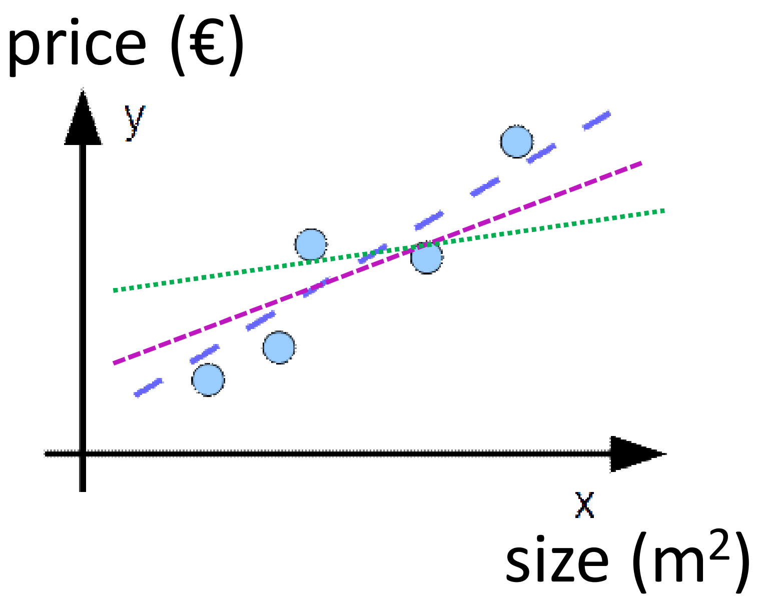
Once the data is given, the error (RMSE) is the function of w_0 and w_1 .

Example: Estimation of House Prices

„hypothesis“ (or hypothesis function)

Models we consider:

$$\hat{y} = w_0 + w_1 x$$



$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

Once the data is given, the error (RMSE) is the function of w_0 and w_1 .

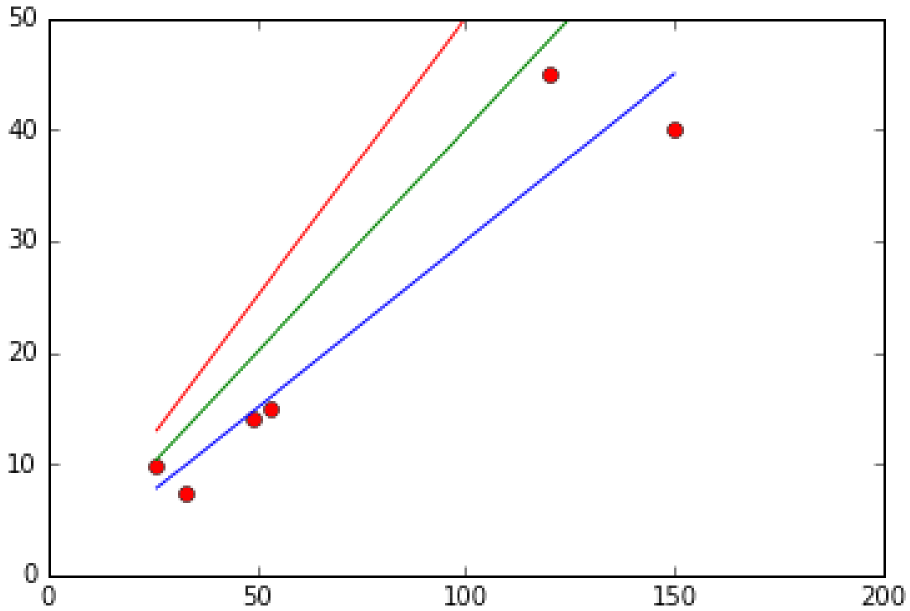
value of the hypothesis function for the *i*-th instance

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

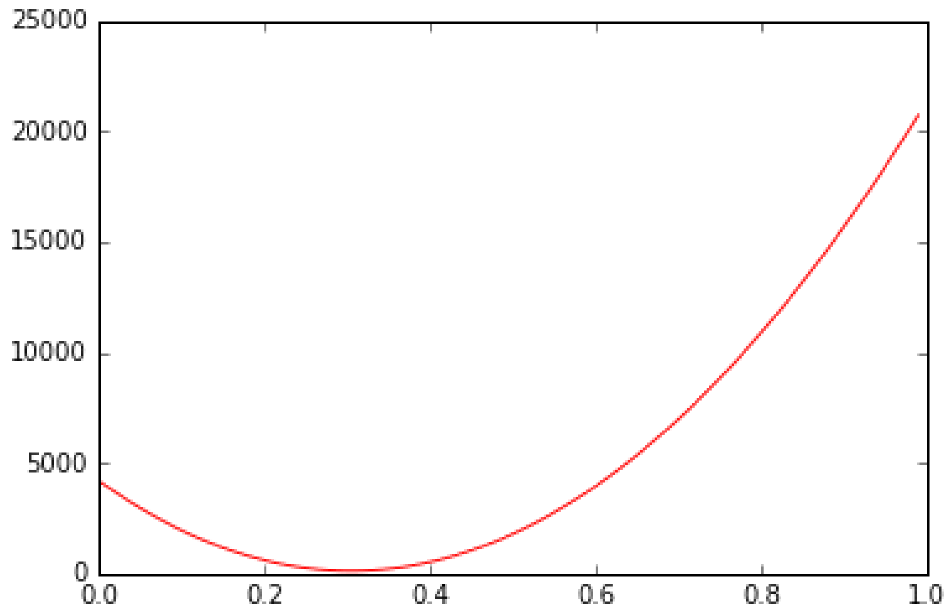
Linear Regression with One Variable

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Example: A Simple Linear Model



Models of the form
price = w * size
(with various values of w)



Sum of squared error of the model as function of w

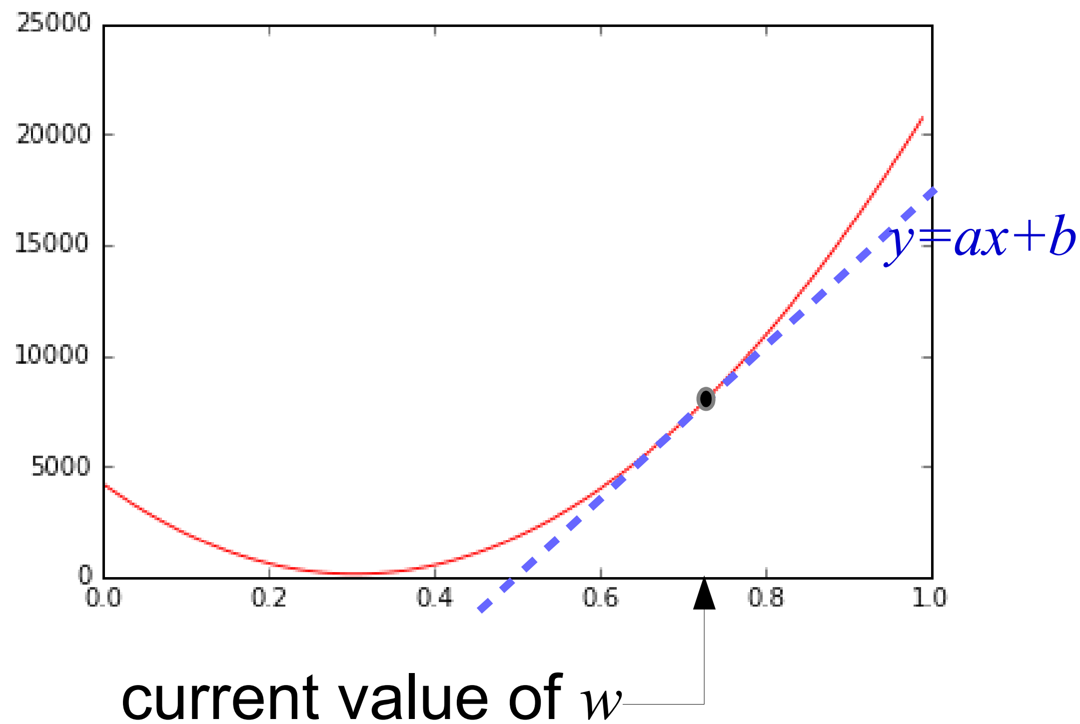
Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- General procedure:
  - **Step 1** Set w to some random value
  - **Step 2** Increase or decrease the value of w a bit so that the error decreases
  - **Step 3** Repeat Step 2 as long as you can decrease the error

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Observations about the Error

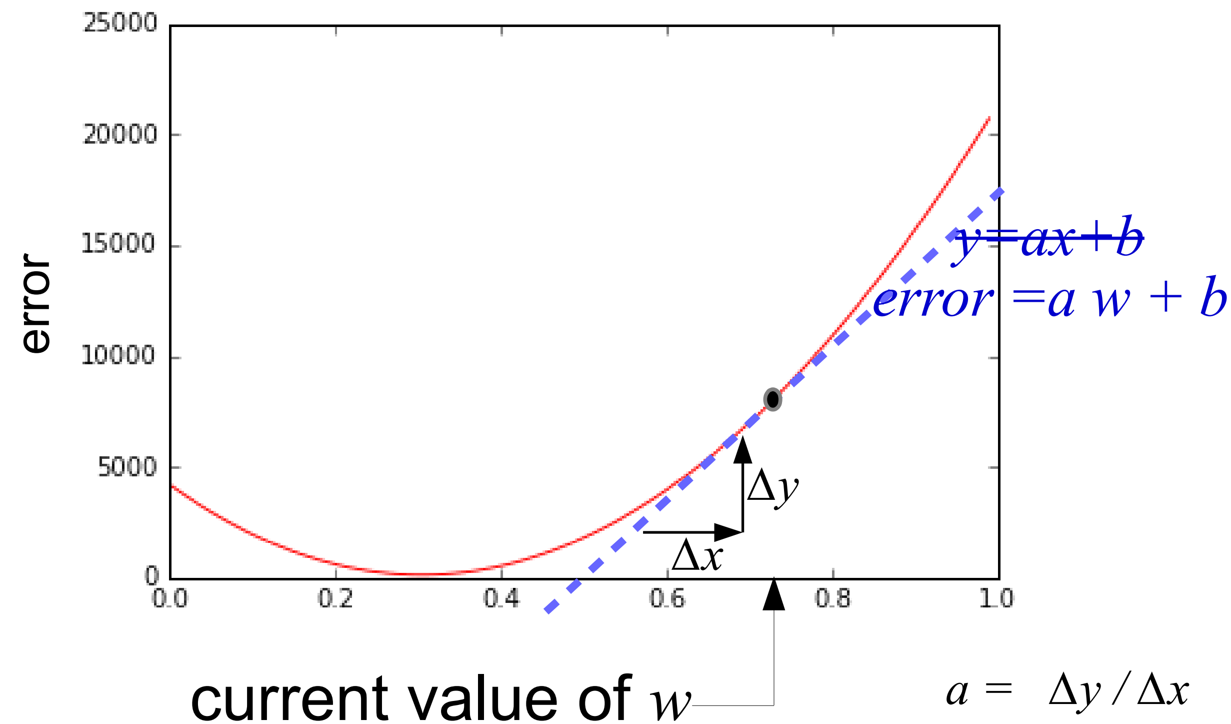


- **Observation 1**
  positive slope ($a > 0$) $\rightarrow$ w should be decreased
  negative slope ($a < 0$) $\rightarrow$ w should be increased

- **Observation 2**
  close to the minimum, the slope is very low $\rightarrow$
  if the absoulte value of the slope is high, you can decrease of increase the value of w a bit more

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Observations about the Error



- **Observation 1**
  positive slope ($a > 0$) $\rightarrow$ w should be decreased
  negative slope ($a < 0$) $\rightarrow$ w should be increased

- **Observation 2**
  close to the minimum, the slope is very low $\rightarrow$
  if the absoulte value of the slope is high, we can decrease or increase the value of w a bit more

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value
  - **Step 2** $w \leftarrow w - \varepsilon\,s$ where
    s is the slope and
    ε is a small number such as $0.00001 = 10^{-5}$
  - **Step 3** Repeat Step 2 as long as you can decrease the error

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

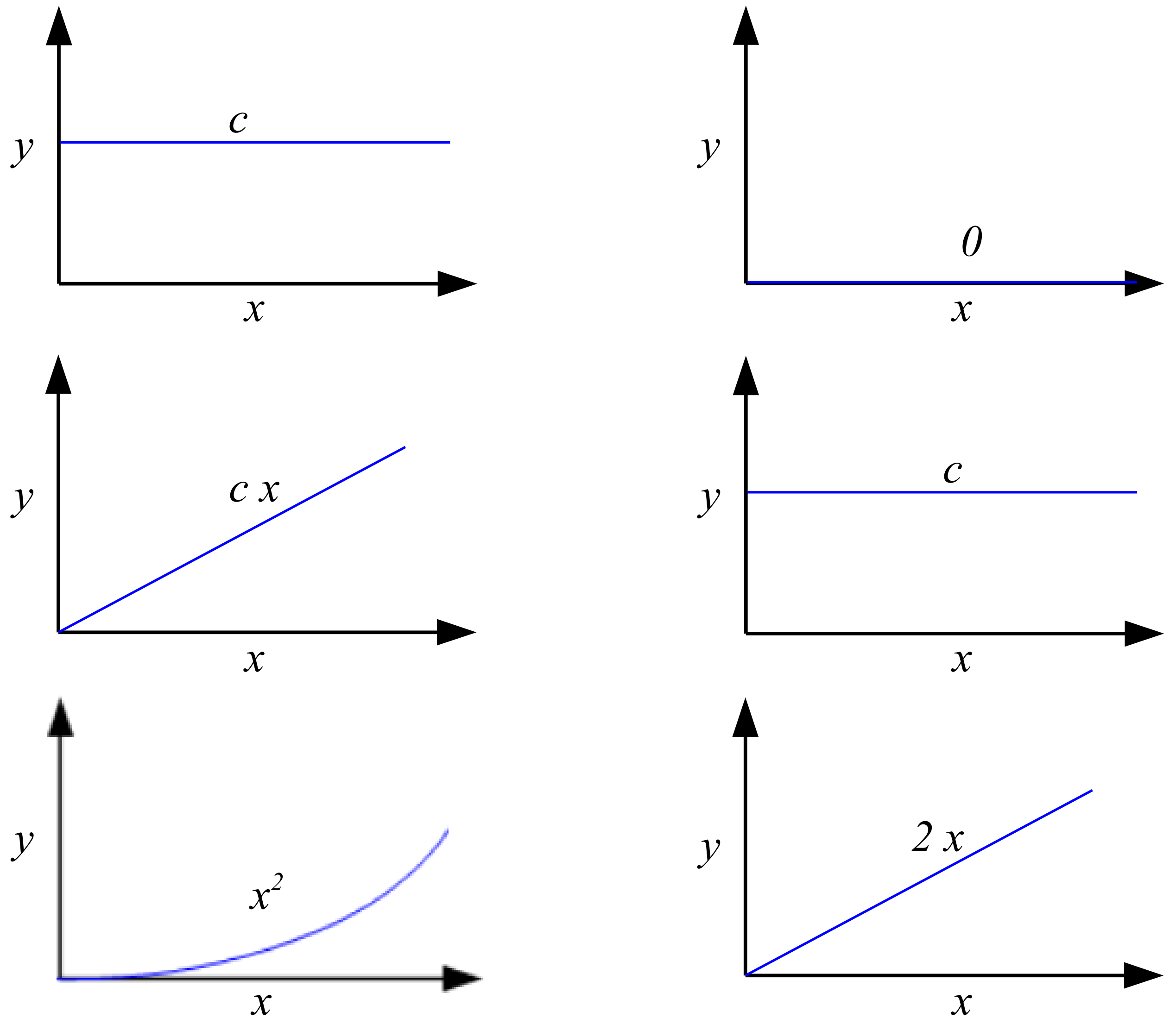
Derivatives

- Mathematically, the slope of function f (or $f(x)$) corresponds to its derivative, denoted as $f\,'$, $\frac{d}{dx} f(x)$ or $\frac{\partial f}{\partial x}$ (in case of multiple variables)

f	$f\,'$
c (constant)	0
$c\,x$ (where c is a constant)	c
x^2	$2x$
x^c (where c is a constant)	cx^{c-1}
$h(x) + g(x)$	$h'(x) + g'(x)$
$h(g(x))$	$h'(g(x))\, g'(x)$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Illustration: Some Functions and Their Derivatives



Can you justify that the derivative of x^2 is really $2x$?

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Calculation of the Slope of the Error

- We consider models of the form $\hat{y} = w\,x$
- Data

x	y
33	20
45	32
61	35

- Sum of squared errors (SSE) for this particular data:

$$(33w - 20)^2 + (45w - 32)^2 + (61\,w - 35)^2$$

- The derivative of the SSE is:

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Calculation of the Slope of the Error

- We consider models of the form $\hat{y} = w\,x$
- Data

x	y
33	20
45	32
61	35

- Sum of squared errors (SSE) for this particular data:

$$z^{2} + (45w - 32)^2 + (61\,w - 35)^2$$

- The derivative of the SSE is:

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Calculation of the Slope of the Error

- We consider models of the form $\hat{y} = w\,x$
- Data

x	y
33	20
45	32
61	35

- Sum of squared errors (SSE) for this particular data:

$$z^2 + (45w - 32)^2 + (61\,w - 35)^2$$

- The derivative of the SSE is:

$$2\,z$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Calculation of the Slope of the Error

- We consider models of the form $\hat{y} = w\,x$
- Data

x	y
33	20
45	32
61	35

- Sum of squared errors (SSE) for this particular data:

$$(33w - 20)^2 + (45w - 32)^2 + (61\,w - 35)^2$$

- The derivative of the SSE is:

$$2\,(33w - 20)$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Calculation of the Slope of the Error

- We consider models of the form $\hat{y} = w\,x$
- Data

x	y
33	20
45	32
61	35

- Sum of squared errors (SSE) for this particular data:

$$(33w - 20)^2 + (45w - 32)^2 + (61\,w - 35)^2$$

- The derivative of the SSE is:

$$2\,(33w - 20)\,33$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Calculation of the Slope of the Error

- We consider models of the form $\hat{y} = w\,x$
- Data

x	y
33	20
45	32
61	35

- Sum of squared errors (SSE) for this particular data:

$$(33w - 20)^2 + (45w - 32)^2 + (61\,w - 35)^2$$

- The derivative of the SSE is:

$$2\,(33w - 20)\,33 + 2\,(45w - 32)\,45 + 2\,(61w - 35)\,61$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value, e.g. 0.3
  - **Step 2** The new value of w is $0.3 - \varepsilon\, s$

    $\varepsilon = 0.00001$

    $s = 2\,(33*0.3 - 20)\,33 + 2\,(45*0.3 - 32)\,45 +$
    $\quad + 2\,(61*0.3 - 35)\,61 = -4369$

    that is: $w \leftarrow 0.34369$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value, e.g. 0.3
  - **Step 2** The new value of w is $0.34369 - \varepsilon\, s$

    $\varepsilon = 0.00001$

    $s = 2\,(33*0.34369 - 20)\,33 + 2\,(45*0.34369 - 32)\,45 +$
    $+ 2\,(61*0.34369 - 35)\,61 = -3771.76$

    that is: $w \leftarrow 0.3814$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value, e.g. 0.3
  - **Step 2** The new value of w is $0.3814 - \varepsilon\, s$
    $\varepsilon = 0.00001$
    $s = 2\,(33{*}0.3814 - 20)\,33 + 2\,(45{*}0.3814 - 32)\,45 +$
    $+ 2\,(61{*}0.3814 - 35)\,61 = -\,3256.26$
    that is: $w \leftarrow 0.4140$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value, e.g. 0.3
  - **Step 2** The new value of w is $0.4140 - \varepsilon\, s$

    $\varepsilon = 0.00001$

    $s = 2\,(33*0.4140 - 20)\,33 + 2\,(45*0.4140 - 32)\,45 +$
    $+ 2\,(61*0.4140 - 35)\,61 = -\,2810.62$

    that is: $w \leftarrow 0.4421$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value, e.g. 0.3
  - **Step 2** The new value of w is $0.4421 - \varepsilon\, s$

    $\varepsilon = 0.00001$

    $s = 2\,(33*0.4421 - 20)\,33 + 2\,(45*0.4421 - 32)\,45 +$
    $\quad + 2\,(61*0.4421 - 35)\,61 = -\,2426.49$

    that is: $w \leftarrow 0.4664$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value, e.g. 0.3
  - **Step 2** The new value of w is $0.4664 - \varepsilon\, s$

    $\varepsilon = 0.00001$

    $s = 2\,(33{*}0.4664 - 20)\,33 + 2\,(45{*}0.4664 - 32)\,45 + {}$
    $+ 2\,(61{*}0.4664 - 35)\,61 = -\,2094.31$

    that is: $w \leftarrow 0.4873$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We
- Co
- Pro
  - S
  - S

...and so on...

s = 2 (33*0.4664 – 20) 33 + 2 (45*0.4664 – 32) 45 + + 2 (61*0.4664 – 35) 61 = - 2094.31

that is: $w \leftarrow 0.4873$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Calculation of the Slope of the Error

- We consider models of the form $\hat{y} = w\,x$
- Data

	x	y
	x_1	y_1
	x_2	y_2
	…	…
i-th instance	x_i	y_i
	…	…
	x_n	y_n

n

- Sum of squared errors (SSE):

$$\sum_{i=1}^{n} \left(\underset{\underset{w\,x_i}{\uparrow}}{\hat{y}_i} - y_i \right)^2$$

- The derivative of the SSE:

$$\sum_{i=1}^{n} 2x_i(\hat{y}_i - y_i)$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error

- Instead of RMSE, we can minimize the sum of squared errors.
- We use gradient descent to minimize the error
- Consider models of the form $\hat{y} = w\,x$
- Procedure:
  - **Step 1** Set w to some random value
  - **Step 2** $w \leftarrow w - \varepsilon \sum_{i=1}^{n} 2x_i(\hat{y}_i - y_i)$
  - **Step 3** Repeat Step 2 as long as you can non-negligibly decrease the error

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with Two Variables

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with Two Variables

- We consider models of the form $\hat{y} = w^{(1)}x^{(1)} + w^{(2)}x^{(2)}$
- Data

$x^{(1)}$	$x^{(2)}$	y
$x_1^{(1)}$	$x_1^{(2)}$	y_1
$x_2^{(1)}$	$x_2^{(2)}$	y_2
...	...	...
$x_i^{(1)}$	$x_i^{(2)}$	y_i
...		...
$x_n^{(1)}$	$x_n^{(2)}$	y_n

i-th instance

n

- Sum of squared errors (SSE):

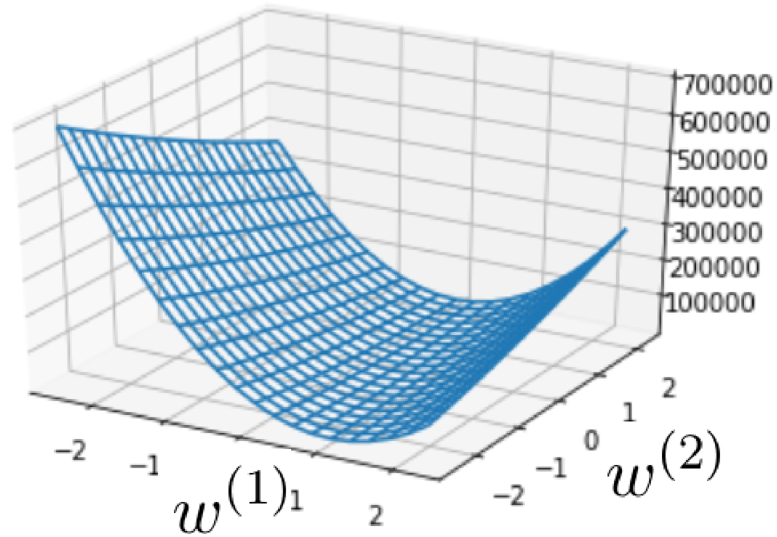
$$\sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)^2$$

$$\hat{y}_i = w^{(1)}x_i^{(1)} + w^{(2)}x_i^{(2)}$$

- Given a dataset, SSE is a function of $w^{(1)}$ and $w^{(2)}$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Error as Function of $w^{(1)}$ and $w^{(2)}$



- The tangent is not a line, but a plane
- Two slopes: w.r.t. $w^{(1)}$ and $w^{(2)}$
- Two partial derivatives: w.r.t. $w^{(1)}$ and $w^{(2)}$
- Calculation of partial derivatives:
  - when calcuating the partial derivative w.r.t. $w^{(1)}$, $w^{(2)}$ should be treated as constant
  - when calcuating the partial derivative w.r.t. $w^{(2)}$, $w^{(1)}$ should be treated as constant

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Partial Derivatives in Case of Two Variables

- Sum of Squared Errors (SSE): $\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2$

$$\hat{y}_i = w^{(1)} x_i^{(1)} + w^{(2)} x_i^{(2)}$$

- Partial Derivative of SSE w.r.t. $w^{(1)}$: $\sum_{i=1}^{n} 2x_i^{(1)}(\hat{y}_i - y_i)$
- Partial Derivative of SSE w.r.t. $w^{(2)}$: $\sum_{i=1}^{n} 2x_i^{(2)}(\hat{y}_i - y_i)$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error (Two Variables)

- We consider models of the form $\hat{y} = w^{(1)}x^{(1)} + w^{(2)}x^{(2)}$
- Procedure:
  - **Step 1** Set $w^{(1)}$ and $w^{(2)}$ to some random values
  - **Step 2**

    $$w^{(1)} \leftarrow w^{(1)} - \epsilon \sum_{i=1}^{n} 2x_i^{(1)}(\hat{y}_i - y_i)$$

    $$w^{(2)} \leftarrow w^{(2)} - \epsilon \sum_{i=1}^{n} 2x_i^{(2)}(\hat{y}_i - y_i)$$

    where $\hat{y}_i = w^{(1)}x_i^{(1)} + w^{(2)}x_i^{(2)}$
  - **Step 3** Repeat Step 2 as long as you can non-negligibly decrease the error

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with Multiple Variables

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with m Variables

- We consider models of the form

$$\hat{y} = w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + \ldots + w^{(m)}x^{(m)} = \sum_{j=1}^{m} w^{(j)}x^{(j)}$$

- Data

m

$x^{(1)}$	$x^{(2)}$	...	$x^{(m)}$	y
$x_1^{(1)}$	$x_1^{(2)}$	...	$x_1^{(m)}$	y_1
$x_2^{(1)}$	$x_2^{(2)}$	...	$x_2^{(m)}$	y_2
...	...	...	...	...
$x_i^{(1)}$	$x_i^{(2)}$	...	$x_i^{(m)}$	y_i
...	...	...	...	...
$x_n^{(1)}$	$x_n^{(2)}$	...	$x_n^{(m)}$	y_n

n

i-th instance

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Dot Product a.k.a. Scalar Product

$$\vec{w} = \mathbf{w} = (w^{(1)}, w^{(2)}, \ldots, w^{(m)})$$

$$\vec{x} = \mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(m)})$$

$$\vec{w}\vec{x} = \mathbf{w}\mathbf{x} = \sum_{j=1}^{m} w^{(j)} x_i^{(j)}$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with m Variables

- We consider models of the form

$$\hat{y} = \vec{w}\vec{x} \quad \text{where} \quad \vec{w} = (w^{(1)}, w^{(2)}, \dots, w^{(m)})$$
$$\vec{x} = (x^{(1)}, x^{(2)}, \dots, x^{(m)})$$

- Data

m

$x^{(1)}$	$x^{(2)}$	...	$x^{(m)}$	y
$x_1^{(1)}$	$x_1^{(2)}$	...	$x_1^{(m)}$	y_1
$x_2^{(1)}$	$x_2^{(2)}$	...	$x_2^{(m)}$	y_2
...	...	...	...	...
$x_i^{(1)}$	$x_i^{(2)}$	...	$x_i^{(m)}$	y_i
...	...	...	...	...
$x_n^{(1)}$	$x_n^{(2)}$	...	$x_n^{(m)}$	y_n

i-th instance

n

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Partial Derivatives in Case of *m* Variables

- Sum of Squared Errors: $E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

$$\hat{y}_i = w^{(1)} x_i^{(1)} + w^{(2)} x_i^{(2)} + \ldots + w^{(m)} x_i^{(m)} = \sum_{j=1}^{m} w^{(j)} x_i^{(j)}$$

- Partial Derivative w.r.t. $w^{(1)}$: $\frac{\partial E}{\partial w^{(1)}} = \sum_{i=1}^{n} 2 x_i^{(1)} (\hat{y}_i - y_i)$

…

- Partial Derivative w.r.t. $w^{(j)}$: $\frac{\partial E}{\partial w^{(j)}} = \sum_{i=1}^{n} 2 x_i^{(j)} (\hat{y}_i - y_i)$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Gradient

$$\nabla E = \left(\frac{\partial E}{\partial w^{(1)}}, \ldots, \frac{\partial E}{\partial w^{(m)}}\right)$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error (m Variables)

- We consider models of the form $\hat{y} = \sum_{j=1}^{m} w^{(j)} x^{(j)}$
- Procedure:
  - **Step 1** Set $w^{(1)}, w^{(2)}, ..., w^{(m)}$ to some random values
  - **Step 2** for j in $1...m$

$$w^{(j)} \leftarrow w^{(j)} - \epsilon \sum_{i=1}^{n} 2 x_i^{(j)} (\hat{y}_i - y_i)$$

where $\hat{y}_i = \sum_{j=1}^{m} w^{(j)} x_i^{(j)}$

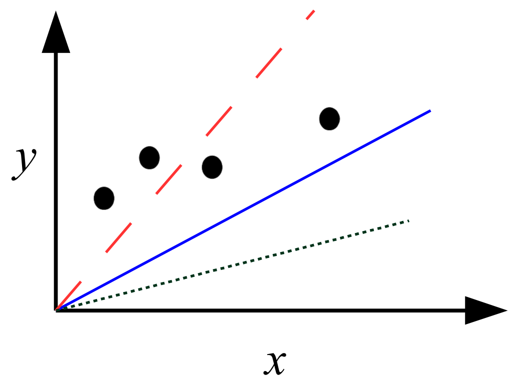
  - **Step 3** Repeat Step 2 as long as you can non-negligibly decrease the error

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with a Bias Term

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

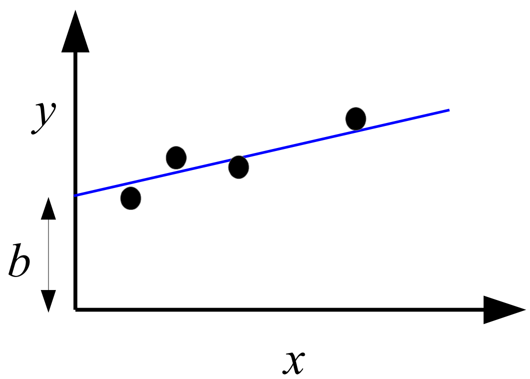
Models without Bias Term



$$\hat{y} = wx$$

$$\hat{y} = \sum_{j=1}^{m} w^{(j)} x^{(j)}$$

Models with a Bias Term



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

$$\hat{y} = b + \sum_{j=1}^{m} w^{(j)} x^{(j)}$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with a Bias Term

- We consider models of the form

$$\hat{y} = b + w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + \ldots + w^{(m)}x^{(m)} = b + \sum_{j=1}^{m} w^{(j)}x^{(j)}$$

- Data

m

	$x^{(1)}$	$x^{(2)}$	...	$x^{(m)}$	y
	$x_1^{(1)}$	$x_1^{(2)}$	...	$x_1^{(m)}$	y_1
	$x_2^{(1)}$	$x_2^{(2)}$	...	$x_2^{(m)}$	y_2
	...	...	...	...	...
i-th instance	$x_i^{(1)}$	$x_i^{(2)}$	...	$x_i^{(m)}$	y_i
	...	...	...	...	...
	$x_n^{(1)}$	$x_n^{(2)}$	...	$x_n^{(m)}$	y_n

n

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with a Bias Term

- We consider models of the form

$$\hat{y} = \underset{w^{(0)}x^{(0)}}{\cancel{b}} + w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + \ldots + w^{(m)}x^{(m)} = \cancel{b} + \sum_{j=\cancel{1}0}^{m} w^{(j)}x^{(j)}$$

- Data

m

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	...	$x^{(m)}$	y
1	$x_1^{(1)}$	$x_1^{(2)}$	...	$x_1^{(m)}$	y_1
1	$x_2^{(1)}$	$x_2^{(2)}$	...	$x_2^{(m)}$	y_2
...	...	...	...	...	...
1	$x_i^{(1)}$	$x_i^{(2)}$	...	$x_i^{(m)}$	y_i
...	...	...	...	...	...
1	$x_n^{(1)}$	$x_n^{(2)}$	...	$x_n^{(m)}$	y_n

n

i-th instance

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Linear Regression with a Bias Term

- We consider models of the form

$$\hat{y} = \cancel{b} + w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + \ldots + w^{(m)}x^{(m)} = \cancel{b} + \sum_{j=\cancel{1}0}^{m} w^{(j)}x^{(j)}$$

$w^{(0)}x^{(0)}$

$$= \vec{w}\vec{x} = \mathbf{wx}$$

- Data

m

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	…	$x^{(m)}$	y
1	$x_1^{(1)}$	$x_1^{(2)}$	…	$x_1^{(m)}$	y_1
1	$x_2^{(1)}$	$x_2^{(2)}$	…	$x_2^{(m)}$	y_2
…	…	…	…	…	…
1	$x_i^{(1)}$	$x_i^{(2)}$	…	$x_i^{(m)}$	y_i
…	…	…	…	…	…
1	$x_n^{(1)}$	$x_n^{(2)}$	…	$x_n^{(m)}$	y_n

i-th instance

n

$$\vec{w} = (w^{(0)}, w^{(1)}, \ldots, w^{(m)})$$

$$\vec{x} = (x^{(0)}, x^{(1)}, \ldots, x^{(m)})$$

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Setting the Learning Rate

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error (m Variables)

- We consdier models of the form $\hat{y} = \sum_{j=1}^{m} w^{(j)} x^{(j)}$
- Procedure:
  - **Step 1** Set $w^{(1)}, w^{(2)}, \ldots, w^{(m)}$ to some random values
  - **Step 2** for j in $1 \ldots m$

$$w^{(j)} \leftarrow w^{(j)} - \epsilon \sum_{i=1}^{n} 2 x_i^{(j)} (\hat{y}_i - y_i)$$

Learning rate

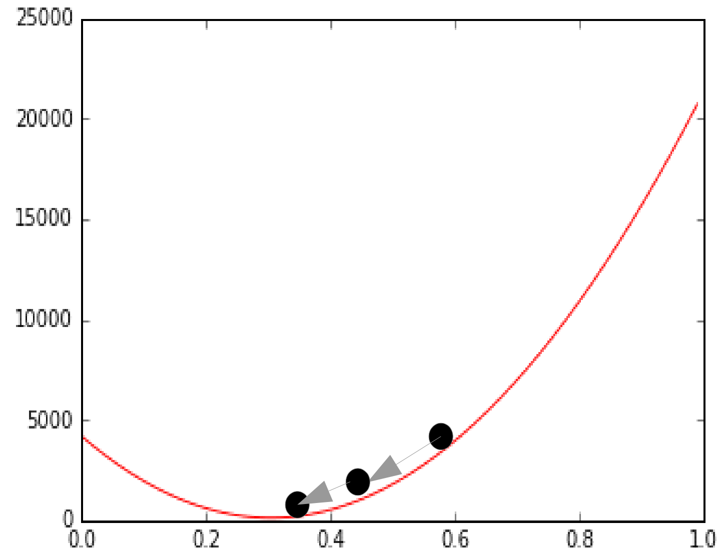
where $\hat{y}_i = \sum_{j=1}^{m} w^{(j)} x_i^{(j)}$

  - **Step 3** Repeat Step 2 as long as you can non-negligibly decrease the error

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Setting the Learning Rate

„just right“



„The model converges“
(slang!)

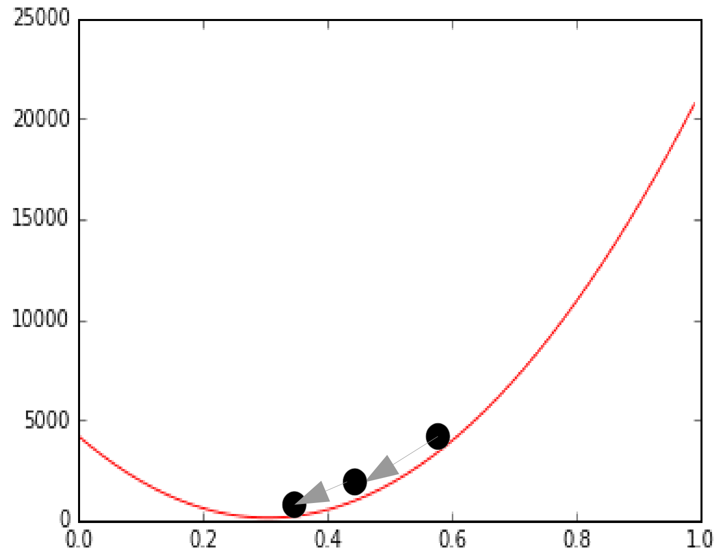
Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Setting the Learning Rate

too low

„just right“

Convergence
may be
very slow



„The model converges“
(slang!)

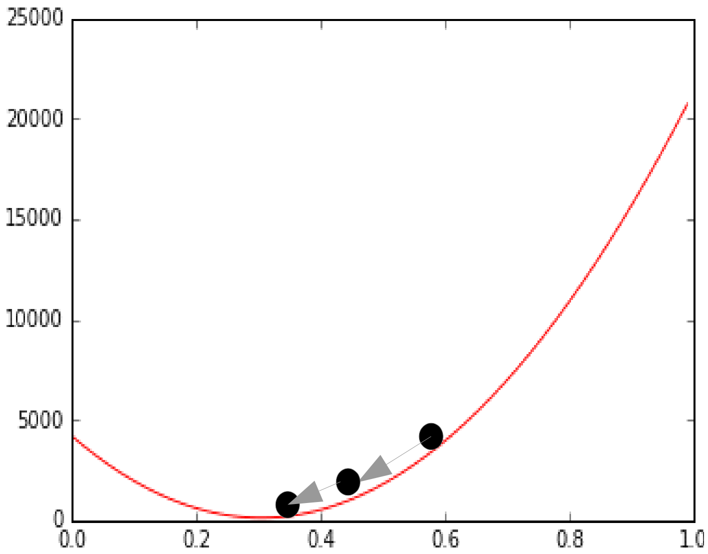
Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Setting the Learning Rate

too low

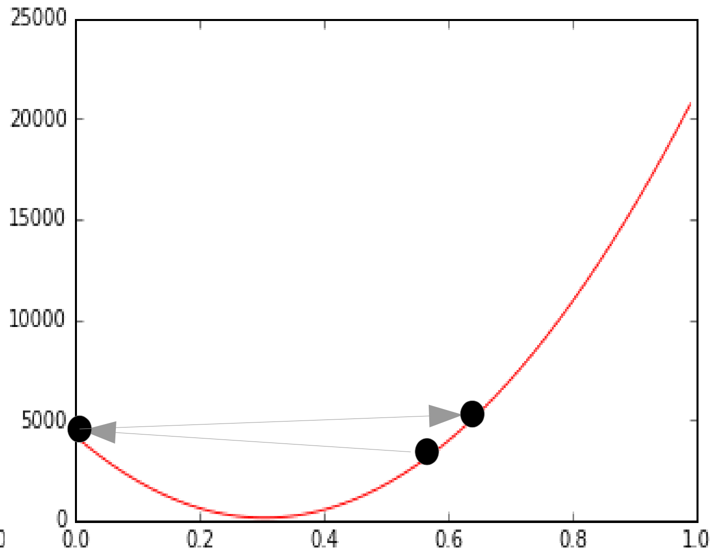
Convergence may be very slow

„just right“



„The model converges“ (slang!)

„too high“



Divergence

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Learning Rate and the Objective Function of the Optimization

- Your model converges with $\varepsilon = 10^{-5}$.
- What if you receive a new dataset containing 1000-times more instances?
- The **sum** of squared errors will be 1000-times larger.
- Optimize the **average** (mean) of squared errors instead of the sum of squared errors.
- This corresponds to a division by n in the partial derviatives and in the update formulas respectively.

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Minimization of the Error (m Variables)

- We consider models of the form $\hat{y} = \sum_{j=1}^{m} w^{(j)} x^{(j)}$
- Procedure:
  - **Step 1** Set $w^{(1)}, w^{(2)}, ..., w^{(m)}$ to some random values
  - **Step 2** for j in $1...m$

$$w^{(j)} \leftarrow w^{(j)} - \epsilon \frac{1}{n} \sum_{i=1}^{n} 2x_i^{(j)} (\hat{y}_i - y_i)$$

where $$\hat{y}_i = \sum_{j=1}^{m} w^{(j)} x_i^{(j)}$$

  - **Step 3** Repeat Step 2 as long as you can non-negligibly decrease the error

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Summary

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Summary

- Revision of mathematical concepts
  - vector, dot product
  - (partial) derivatives, gradient
  - convergence and divergence
- Linear regression: $\hat{y} = \vec{w}\vec{x}$
- Learning as an optimisation task
  - Optimisation technique we considered: gradient descent
  - Further optimisation techniques: stochastic gradient descent, batch gradient descent, ADAM
- Learning rate should be set carefully

Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Essential Concepts

- Vector
- Dot product or Scalar Product
- (Partial) Derivative of a function
- Gradient
- Gradient Descent
- Learning Rate
- Model Equation
- Parameters of a Model
- Parameters of a Model
- Objective Function
- Convergence
- Divergence
- Instance or Observation or Object
- Attribute or Feature
- Target or Label
- RMSE (root mean squared error)