Machine Learning

Linear methods

Karol Przystalski

March 20, 2024

Department of Information Technologies, Jagiellonian University

Agenda

1. Linear regression
2. Logistic regression
3. Linear Discriminate Analysis

Linear regression

Correlation

The most popular correlation measure is Pearson correlation. It is about how one feature depend on second feature. We can say that a dogs size is highly correlated to its weight. So we have two features: size and weight of a dog and we know that a higher dogs is usually heavier.

Correlation

Correlation is a value from -1 to 1 and represents the dependency of two values (features) like shown in table.

correlation value	correlation
0	no correlation between variables
$0 - 0.3$	low correlation
$0.3 - 0.5$	mid correlation
$0.5 - 0.7$	mid-high correlation
$0.7 - 0.9$	high correlation
above 0.9	very high correlation
1	total correlation

Table 1: Correlation dependencies

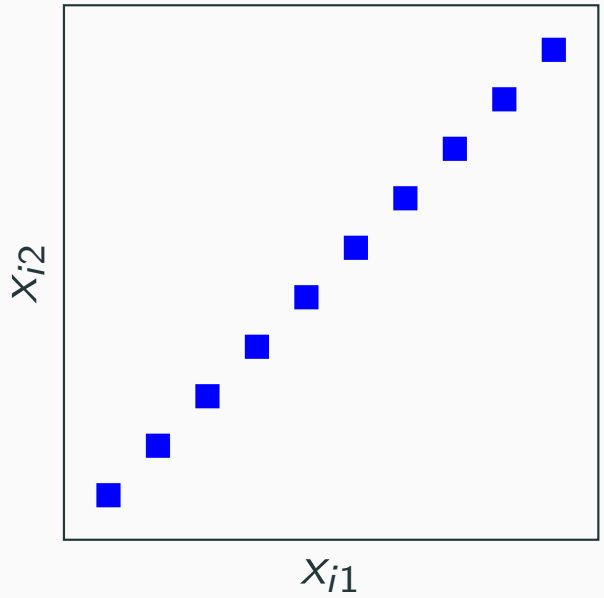
Presented values are positive and we have the same correlation for negative values.

Correlation

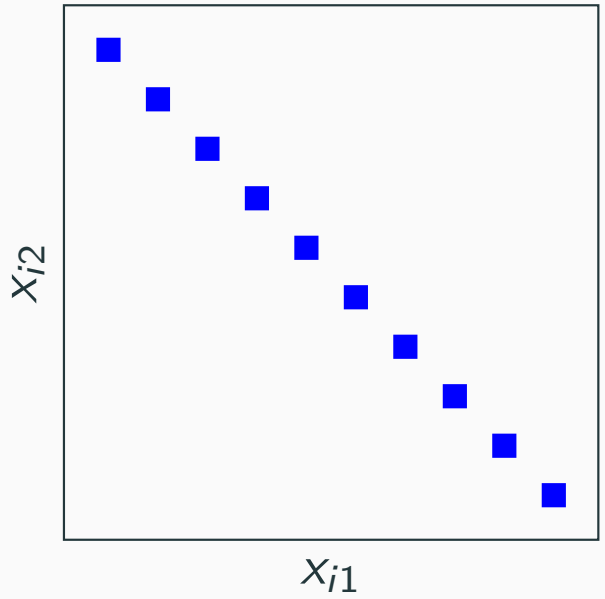
Correlation for two features can be calculated as follows:

$$r^2 = \frac{\sum_{i=1}^{n}(x_{i1} - \overline{x_{i1}})(x_{i2} - \overline{x_{i2}})}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \overline{x_{i1}})^2 \sum_{i=1}^{n}(x_{i2} - \overline{x_{i2}})^2}}. \tag{1}$$

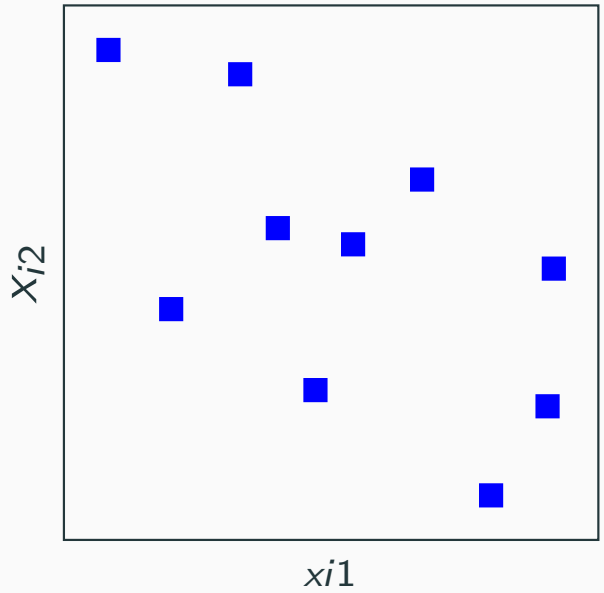
Correlation example where r=1.0



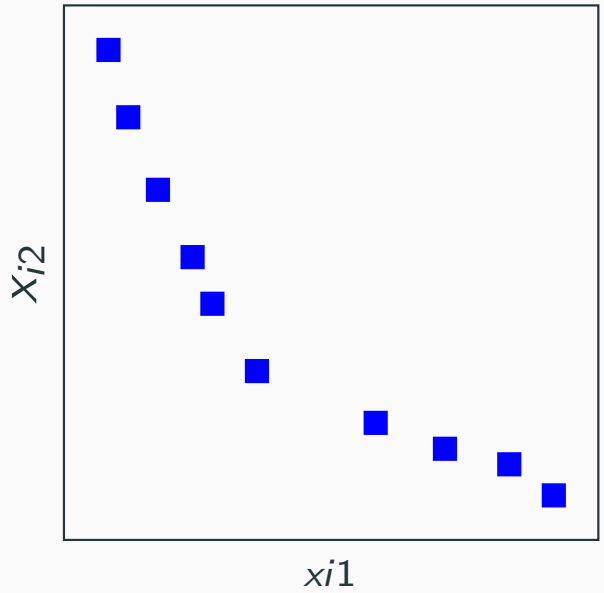
Correlation example where r=-1.0



Correlation example where r=0.2



Correlation example where r≈ -0.7



Correlation – students example

What is the correlation between hours spent on learning and final note of an exam? Lets assume that we have five degrees from A to F. We have a range of points that correspond to each note:

- A – 100-91,
- B – 90-81,
- C – 80-71,
- D – 70-61,
- E – 60-51,
- F – 50-0.

Lets assume we have the averages of hours spent on learning and points collected during the exam looks like presented in table.

Average hours spent on learning	10	22	30	38	44	48
Average points collected	40	52	61	75	88	95

Table 2: Correlation between hours spent on learning and exam note example data

Correlation – students example

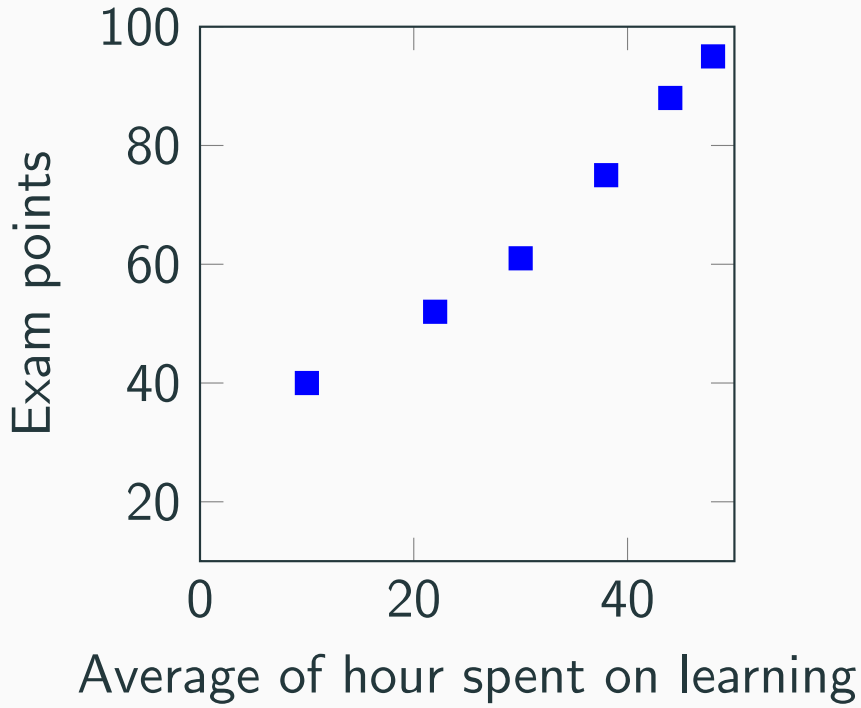


Figure 1: Average hours spent on learning compared to the final exam

Correlation – example result

We can easily say based on the figure that the correlation is positive curvilinear. To be sure we can calculate the correlation value like following:

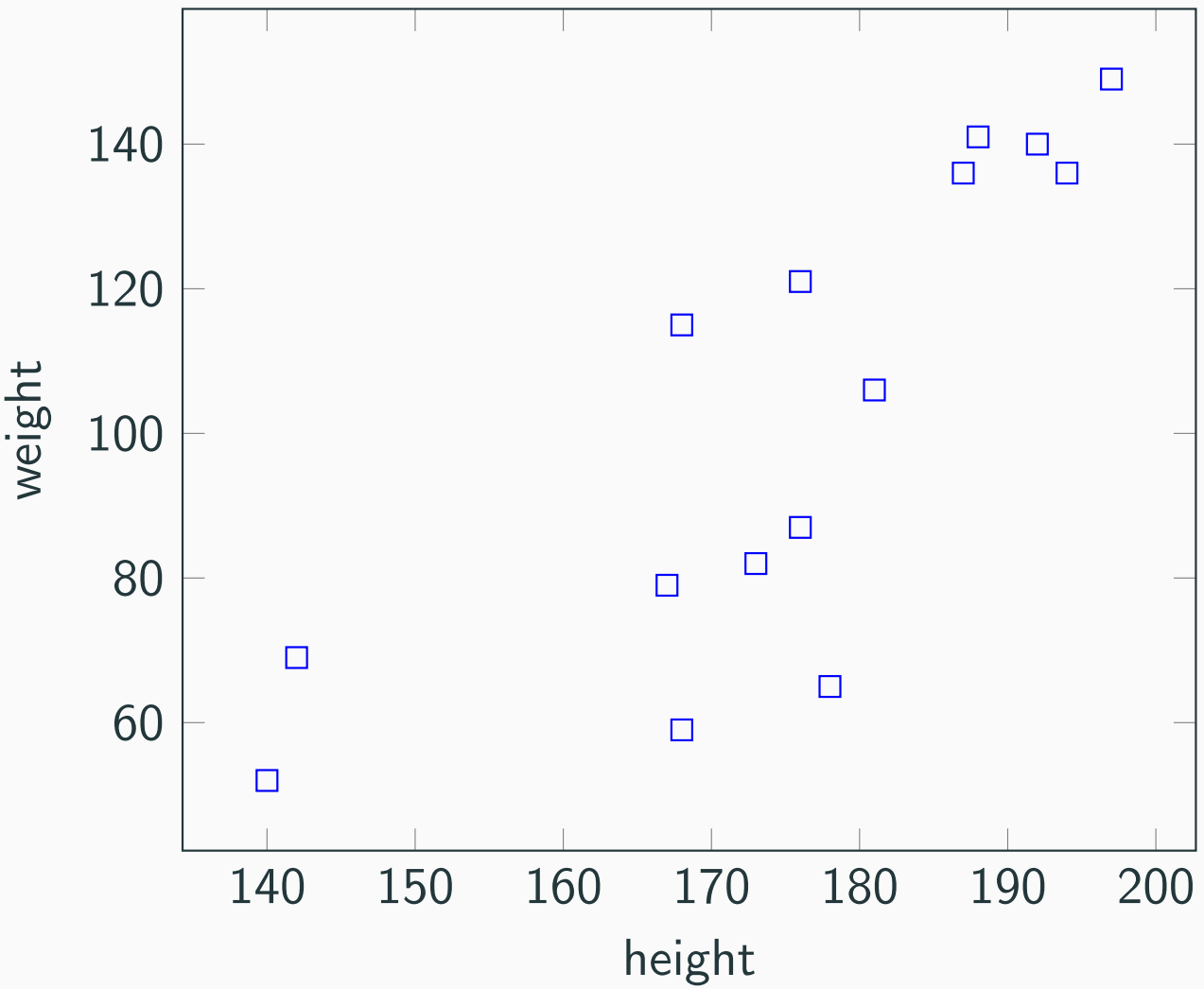
$$r = \frac{627 + 165 + 15 + 39 + 234 + 424}{\sqrt{1024 \cdot 2265.5}} \approx 0.9874.$$

This proves that there is a high correlation between hours that we spent on learning and the final exam note that we get.

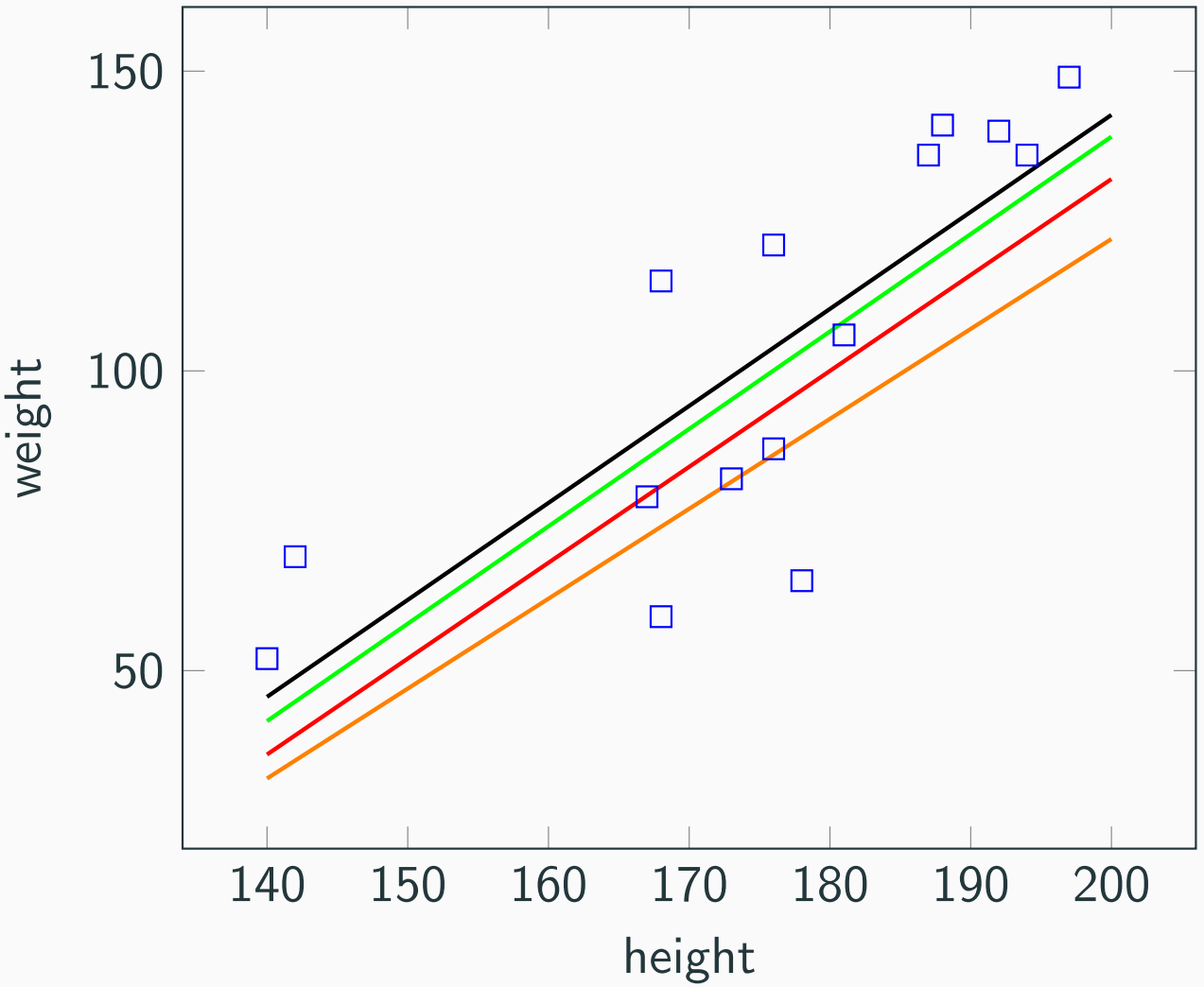
Regression

Regression is about prediction the future values of one feature that depends on a second feature. The first feature is called as explanatory variable and the second that depends on it is called response variable. To simplify let x be explanatory variable and $\hat{y}$ a response variable.

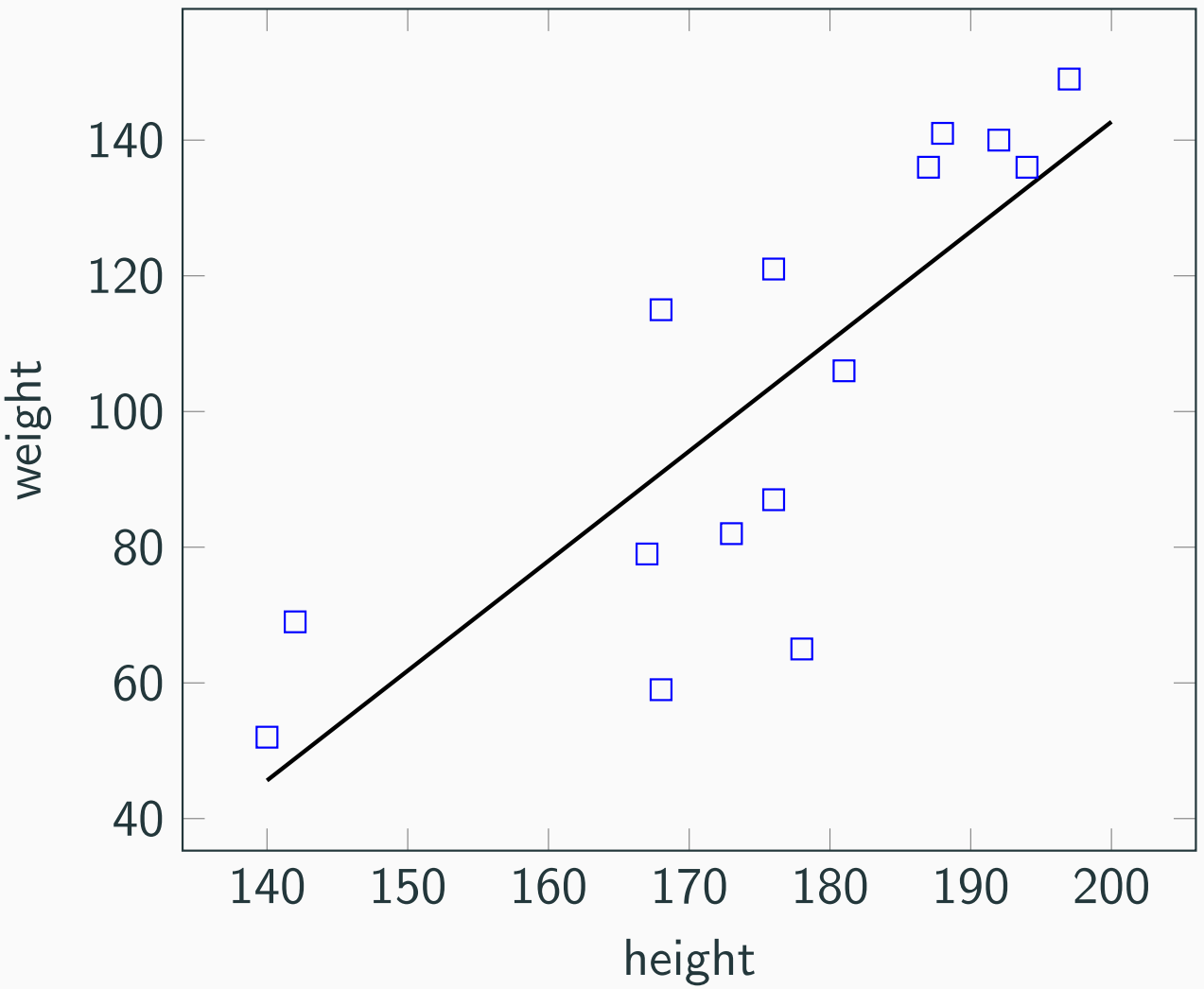
Linear regression – Idea



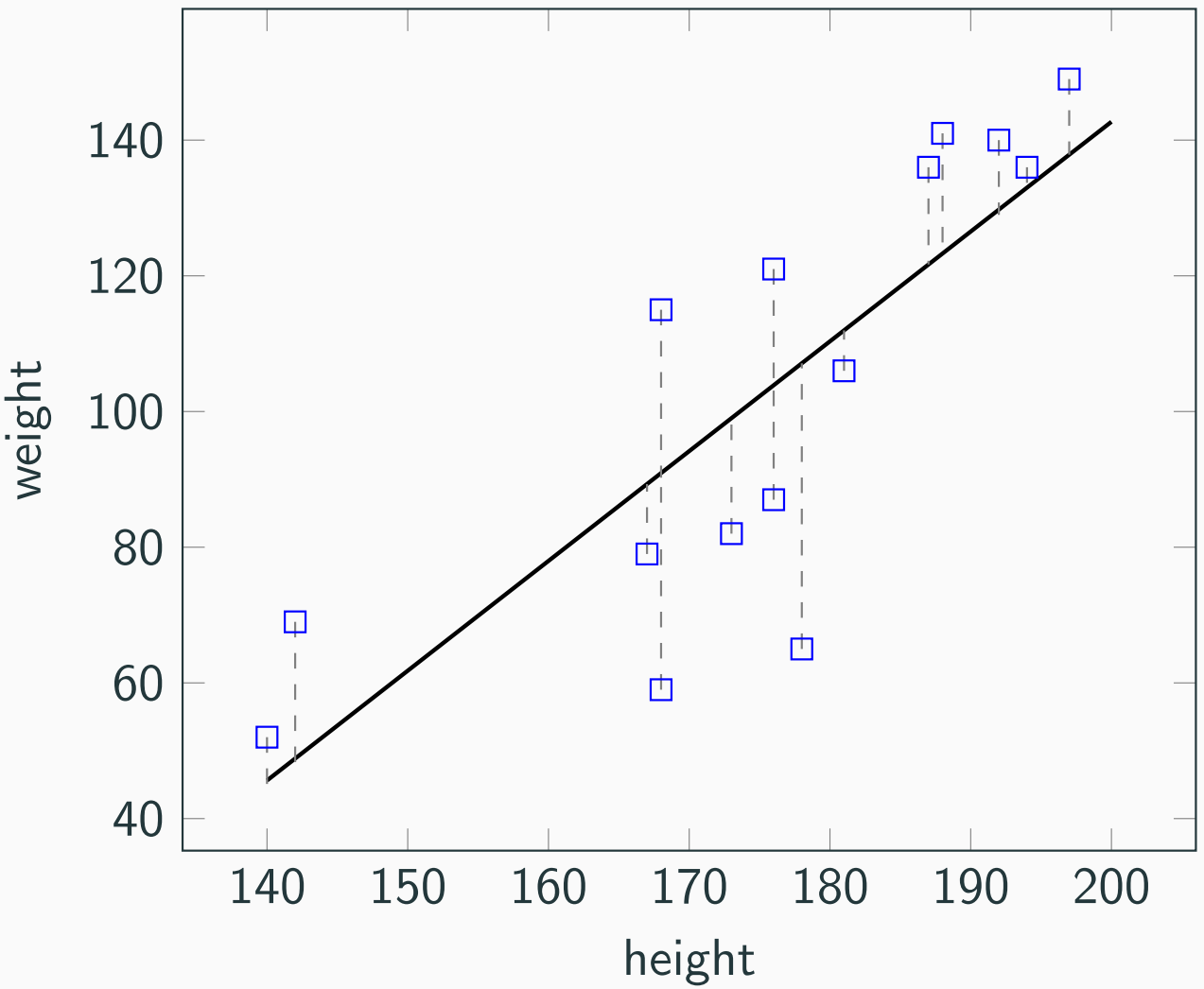
Linear regression – Idea



Linear regression – Idea



Linear regression – How to fit?

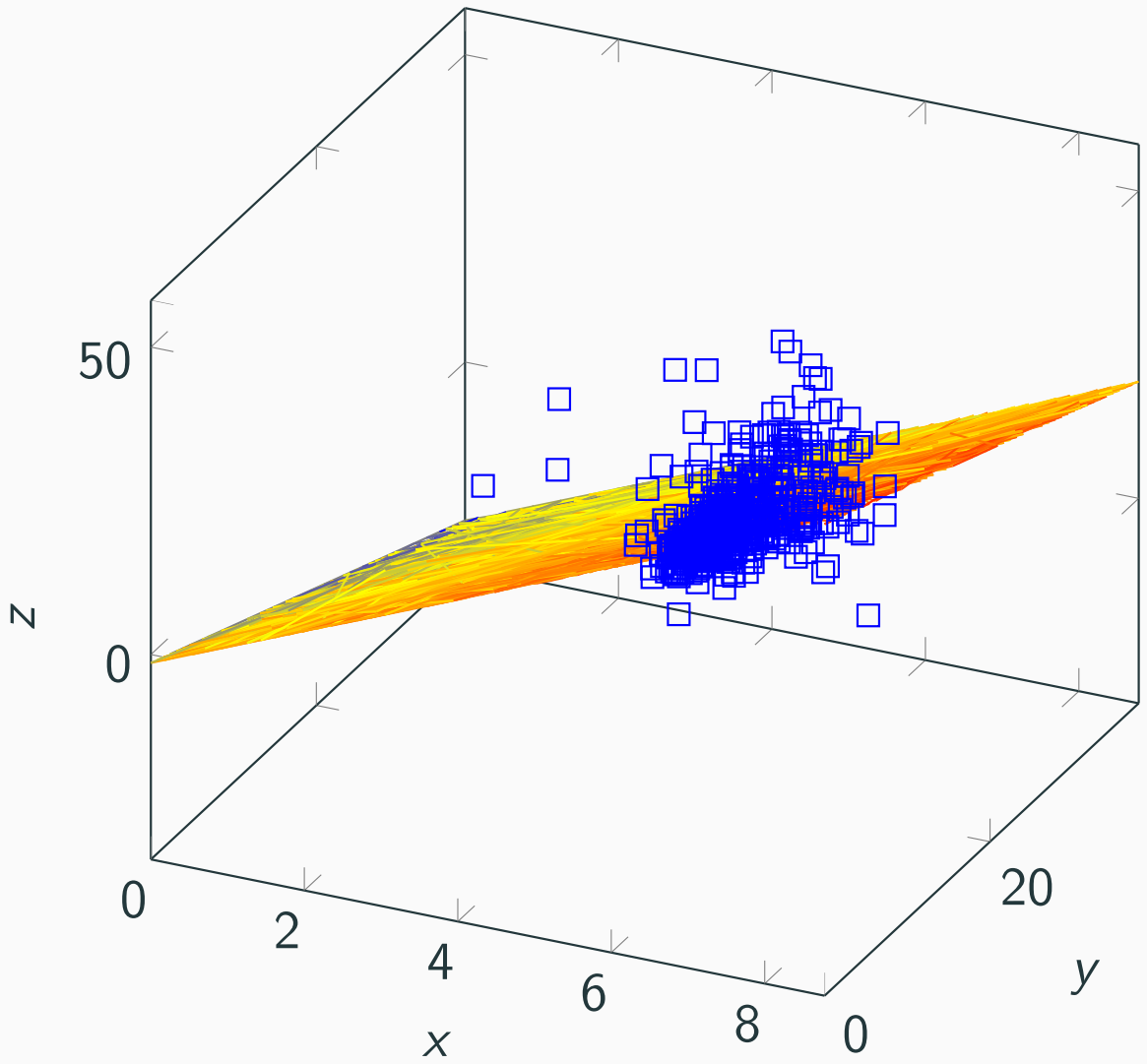


Linear regression – MSE

One of the most popular method to find the proper weights is Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2. \tag{2}$$

Linear regression – a bit more complex example



Linear regression – Equation

It can be calculated as follow:

$$\hat{y} = ax_i + b, \tag{3}$$

where

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4}$$

and

$$b = \overline{y} - a\overline{x}. \tag{5}$$

Variable a is known as slope and is calculated with ordinary least square method that is similar to correlation.

Anscombe data sets

In 1973 Frank Anscombe figured out that we can have multiple of data sets that can give the same results of linear regression.

Anscombe data sets

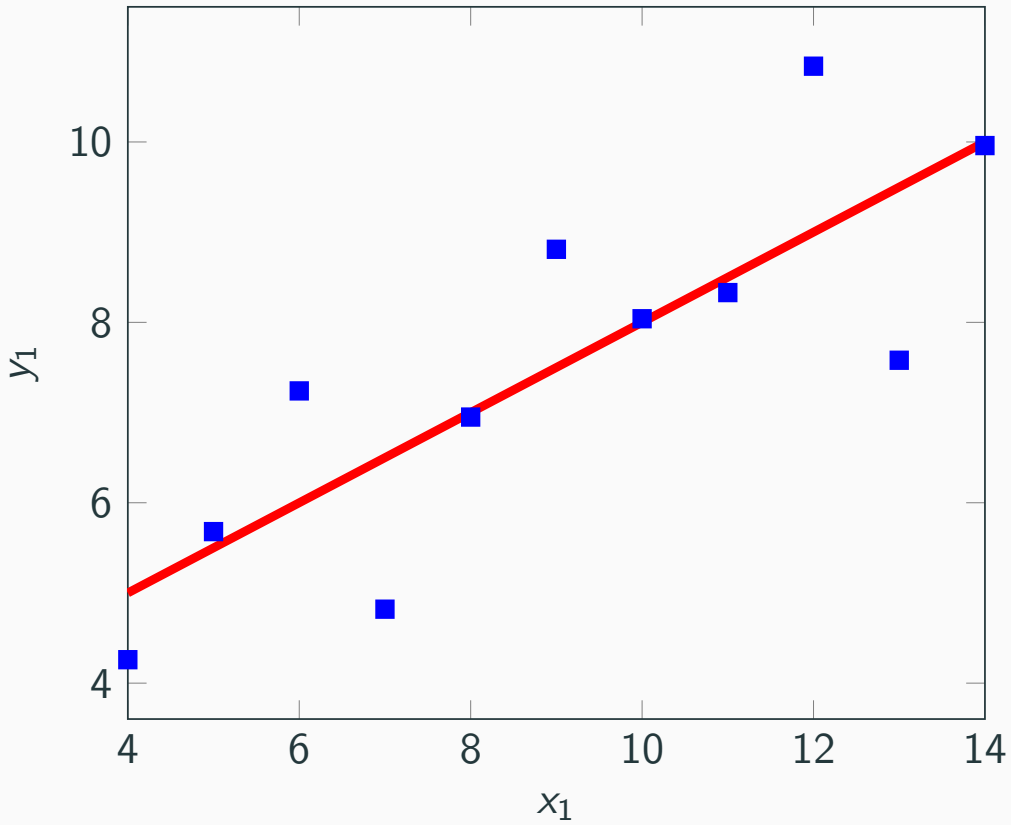
I		II		III		IV	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10,00	8,04	10,00	9,14	10,00	7,46	8,00	6,58
8,00	6,95	8,00	8,14	8,00	6,77	8,00	5,76
13,00	7,58	13,00	8,74	13,00	12,74	8,00	7,71
9,00	8,81	9,00	8,77	9,00	7,11	8,00	8,84
11,00	8,33	11,00	9,26	11,00	7,81	8,00	8,47
14,00	9,96	14,00	8,10	14,00	8,84	8,00	7,04
6,00	7,24	6,00	6,13	6,00	6,08	8,00	5,25
4,00	4,26	4,00	3,10	4,00	5,39	19,00	12,50
12,00	10,84	12,00	9,13	12,00	8,15	8,00	5,56
7,00	4,82	7,00	7,26	7,00	6,42	8,00	7,91
5,00	5,68	5,00	4,74	5,00	5,73	8,00	6,89

Anscombe data sets

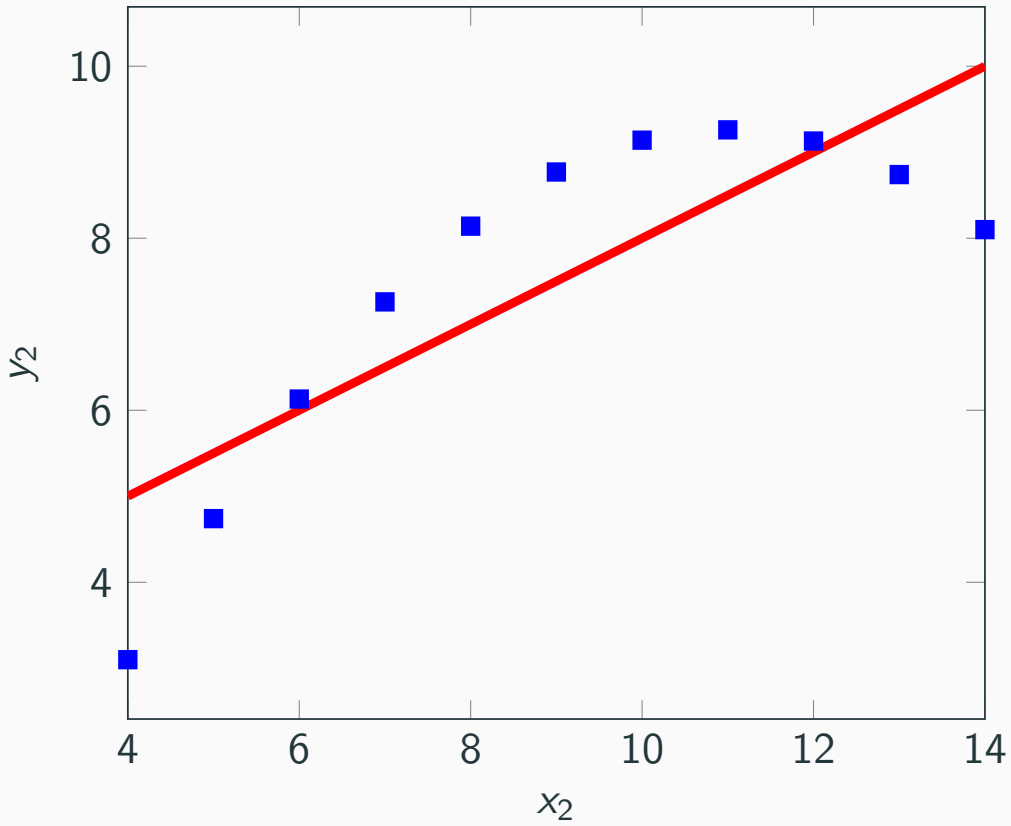
The means for both features are the same and are respectively $\overline{x} = 9$ and $\overline{y} = 7.5$. The regression equation would look for each data set the same and looks like follows:

$$\hat{y} = 3 + \frac{1}{2}x_i.$$

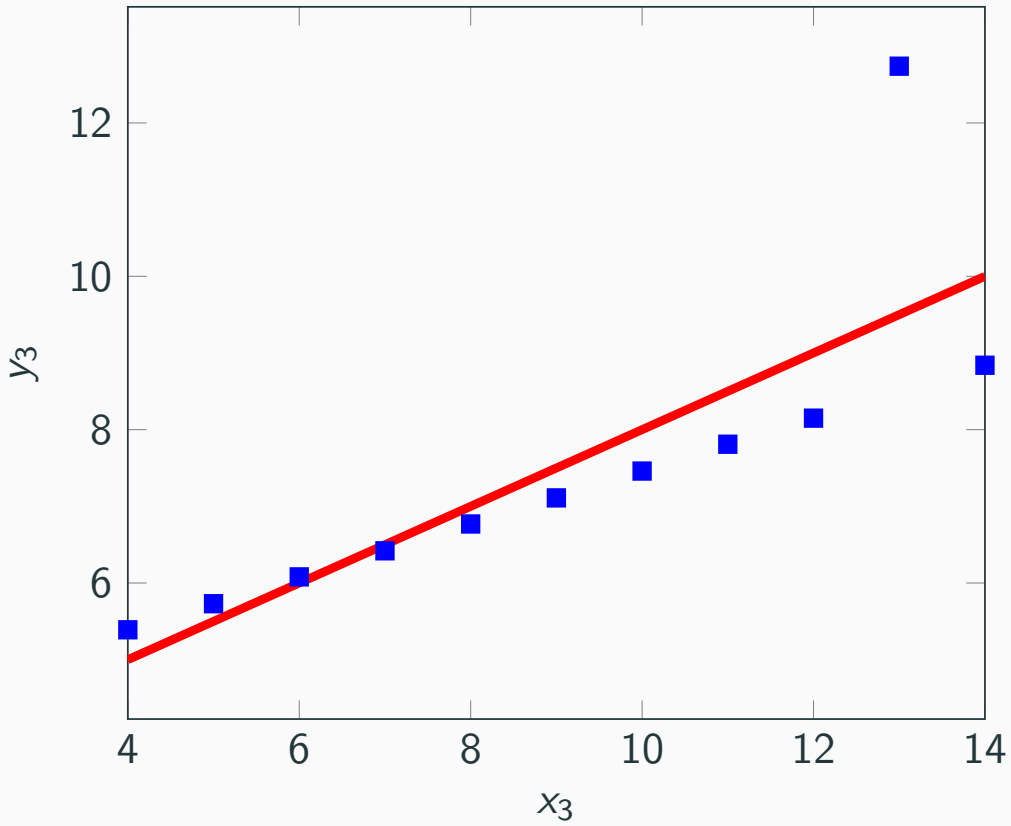
Anscombe data sets – plot



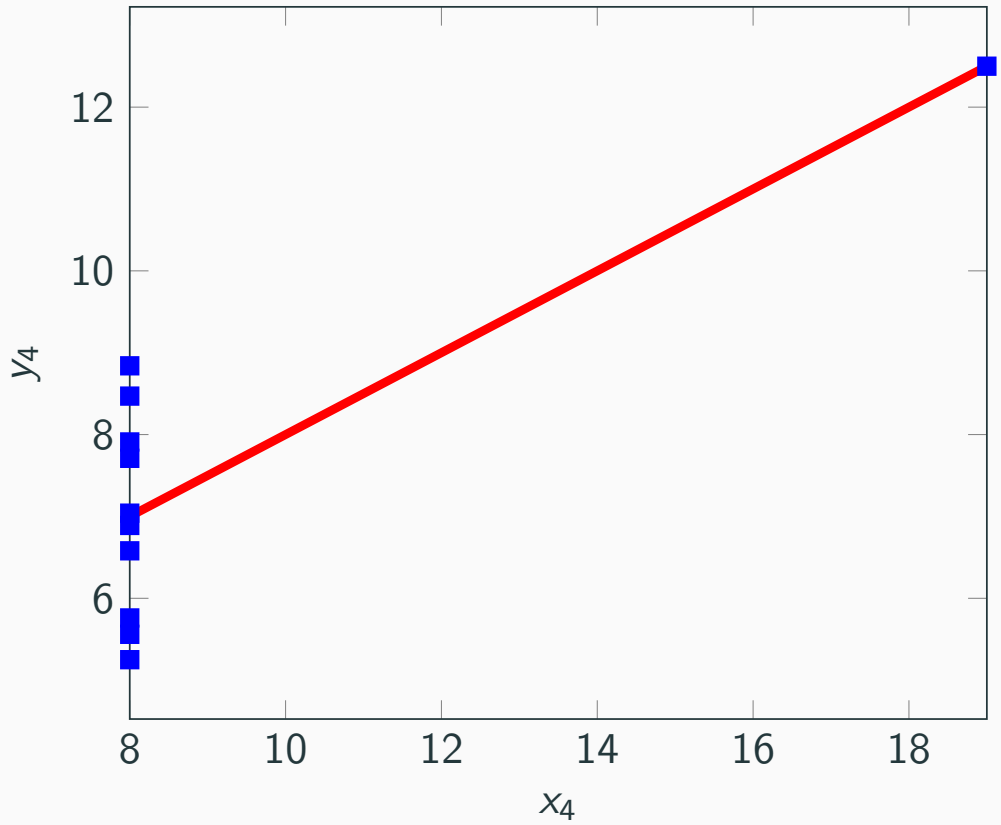
Anscombe data sets – plot



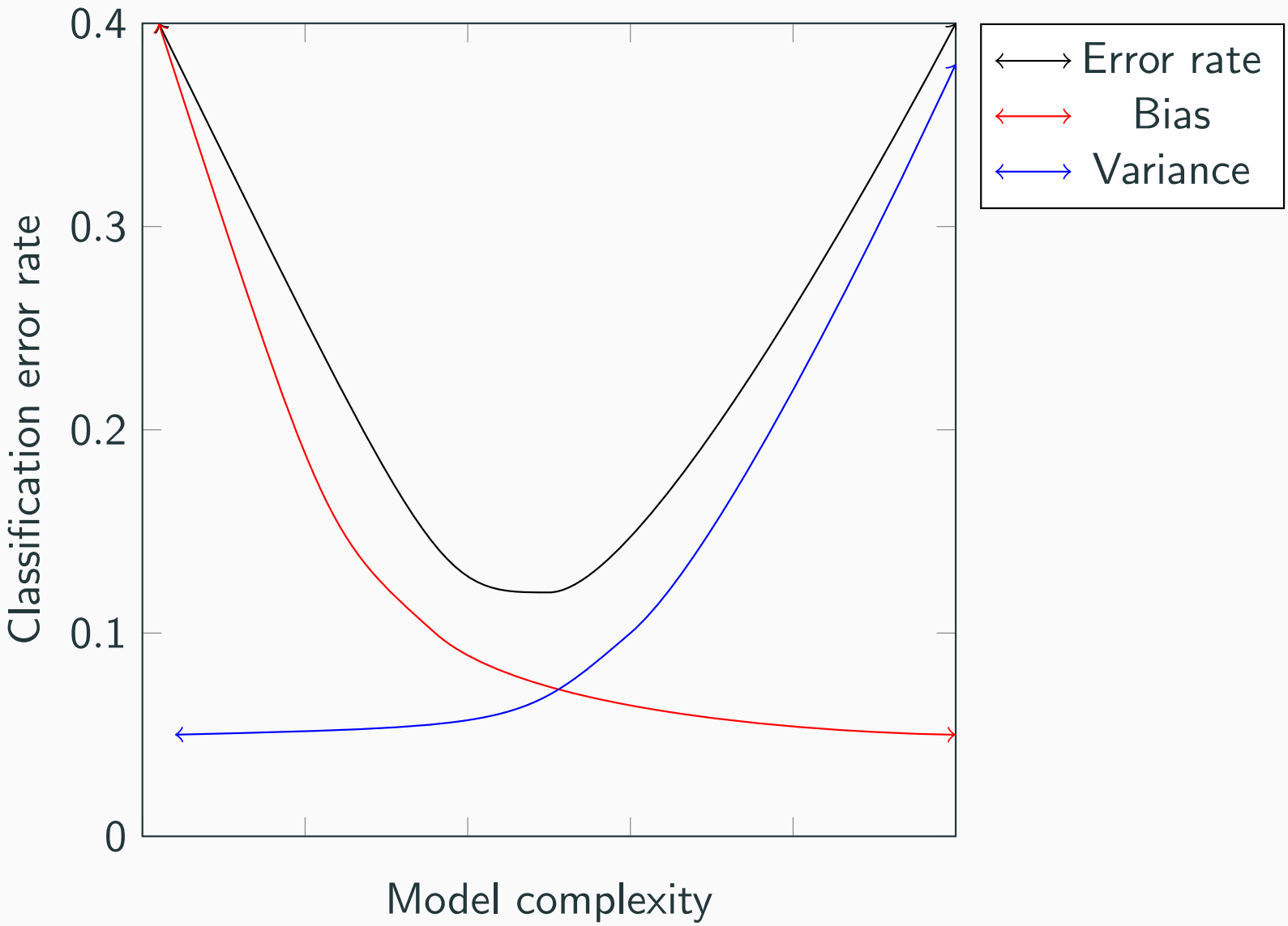
Anscombe data sets – plot



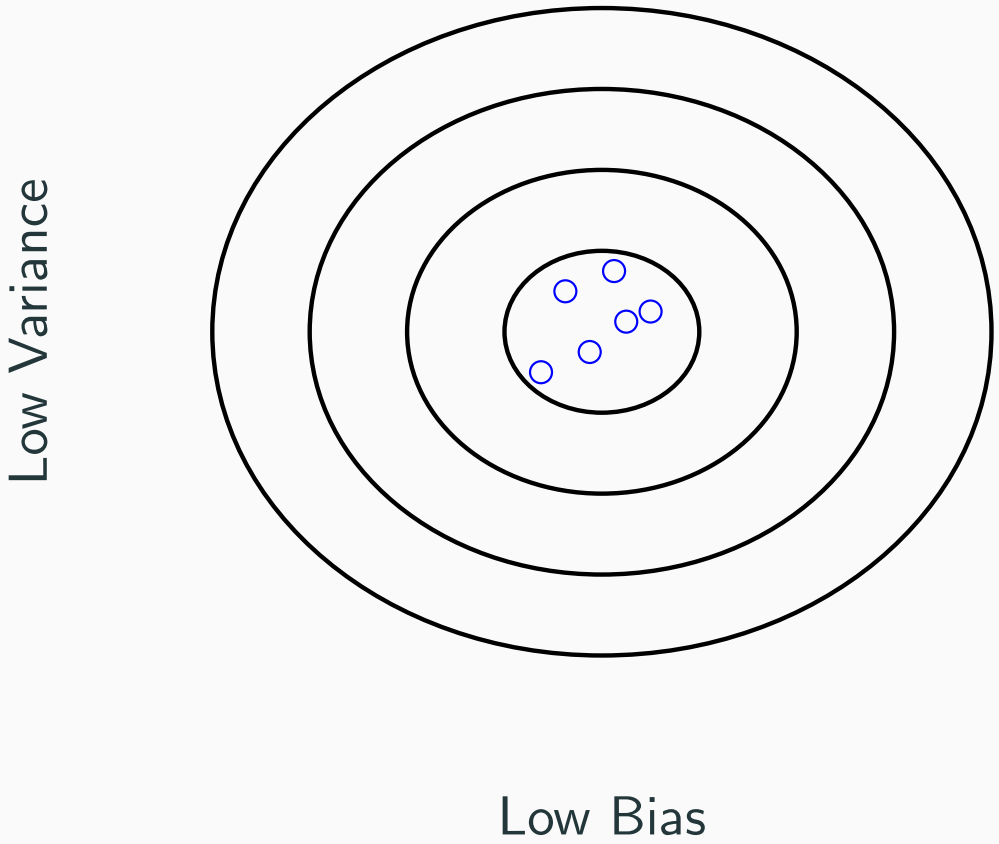
Anscombe data sets – plot



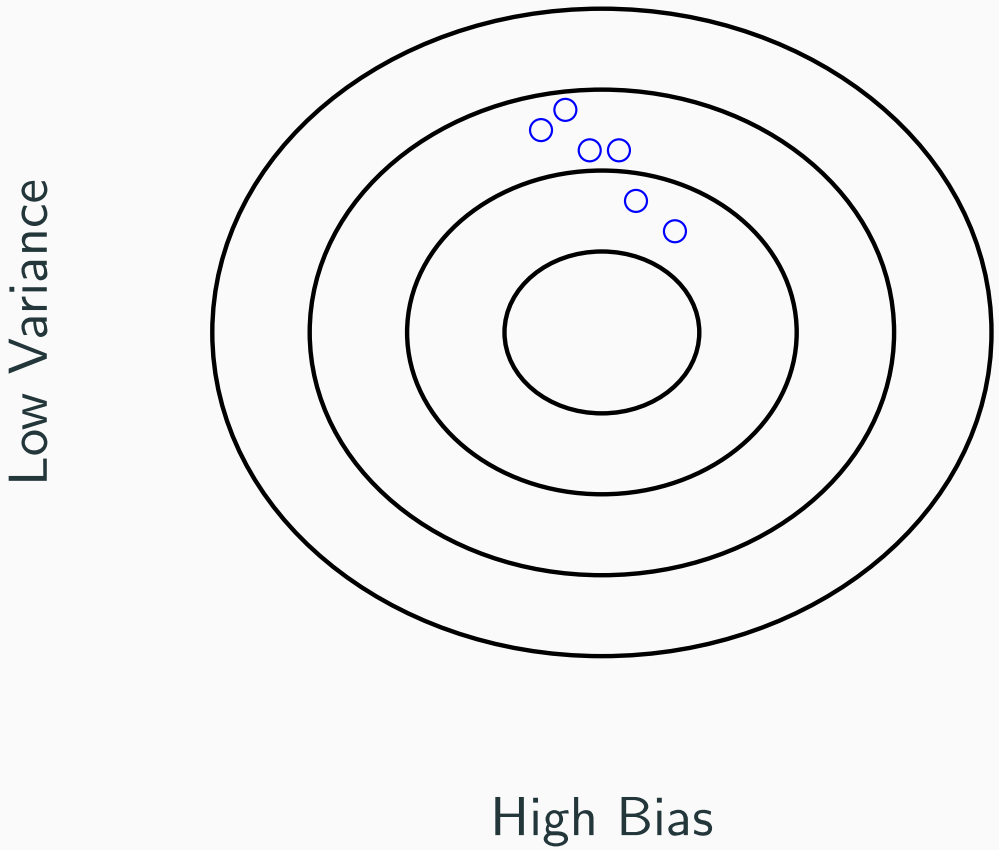
Bias vs. variance



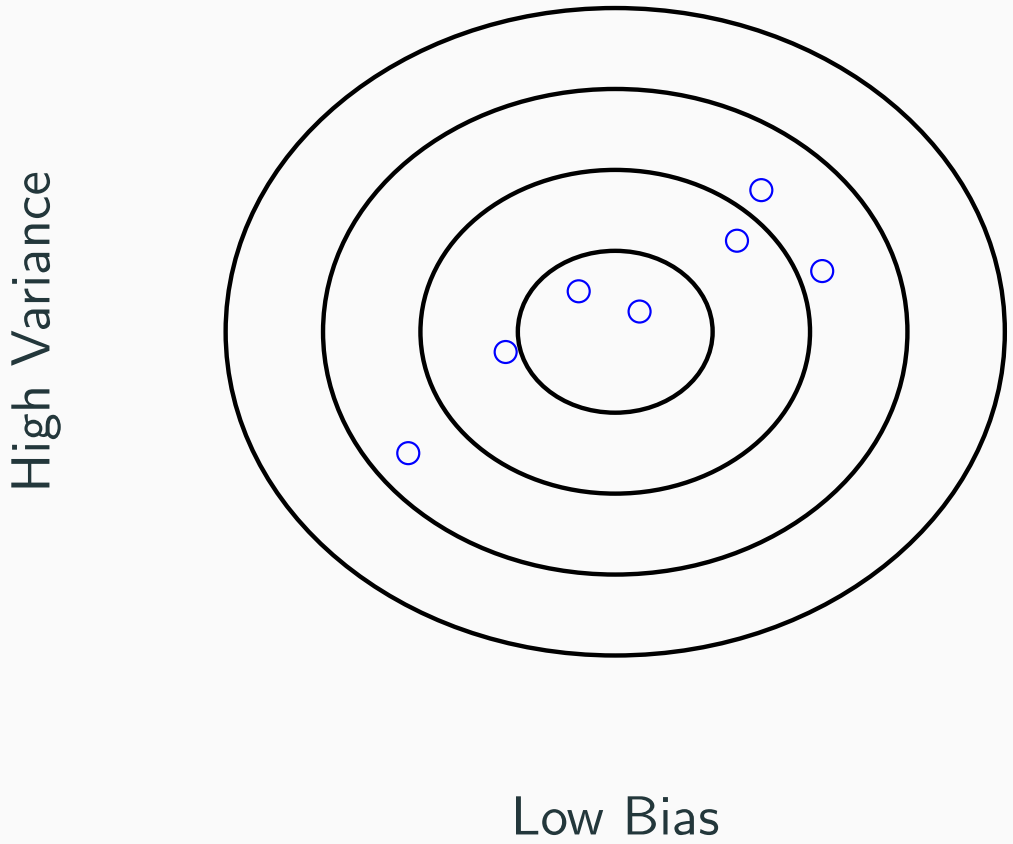
Bias vs. variance



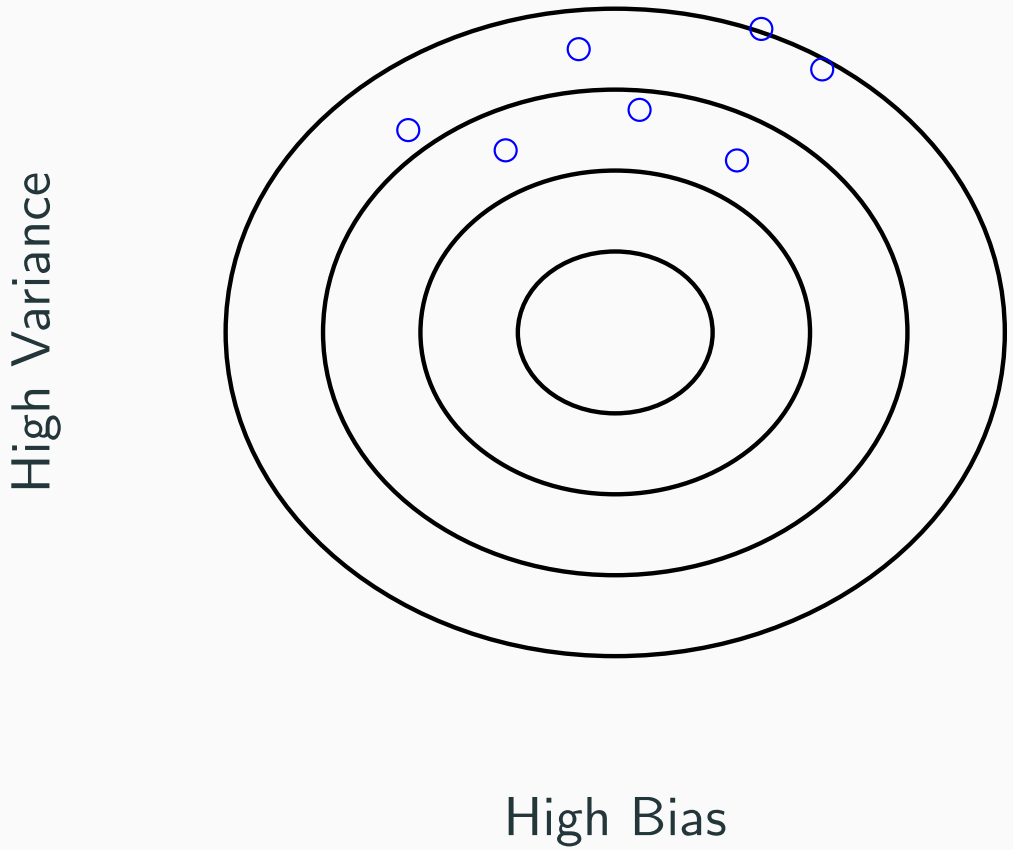
Bias vs. variance



Bias vs. variance



Bias vs. variance



Regularization

A complex model can end with overfitting. There are some methods that can reduce the model complexity like:

- Lasso regression,
- Ridge regression,
- Elastic Net regression.

A cost function in linear regression is:

$$\sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \dot{x}_{ij})^2. \tag{6}$$

L1 regularization – Lasso regression

Lasso stands for Least absolute shrinkage and selection operator. It use the L1 regularizer. We take magnitudes into account:

$$\sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \dot{x}_{ij})^2 + \lambda \sum_{j=0}^{p} |w_j|. \tag{7}$$

For $\lambda = 0$ the formula is a linear regression one. This regularization can make some of the features not to be taken in the final output. It means we can use Lasso to select the features. The λ value:

- higher value means less features,
- lower values means more features selected.

L2 regularization – Ridge regression

Ridge regression is about to shrink the coefficients. The equation of the ridge regression can be written as:

$$\sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \dot{x}_{ij})^2 + \lambda \sum_{j=0}^{p} w_j^2. \tag{8}$$

The λ adds a penalty to the coefficients w. It avoid to have to big values of the coefficients and add a penalty whenever the values are going to be too big. The λ value:

- higher value means more penalty when the coefficients are bigger,
- lower values make it more like regular linear regression,
- higher values makes the variance decreases, and the bias increases.

Elastic Net regression – L1 and L2 regularization together

Elastic Net implements both L1 and L2 regularizators.

The cost function is as: A cost function in linear regression is:

$$\frac{\sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \dot{x}_{ij})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m} w_j^2 + \alpha \sum_{j=1}^{m} |w_j|). \qquad (9)$$

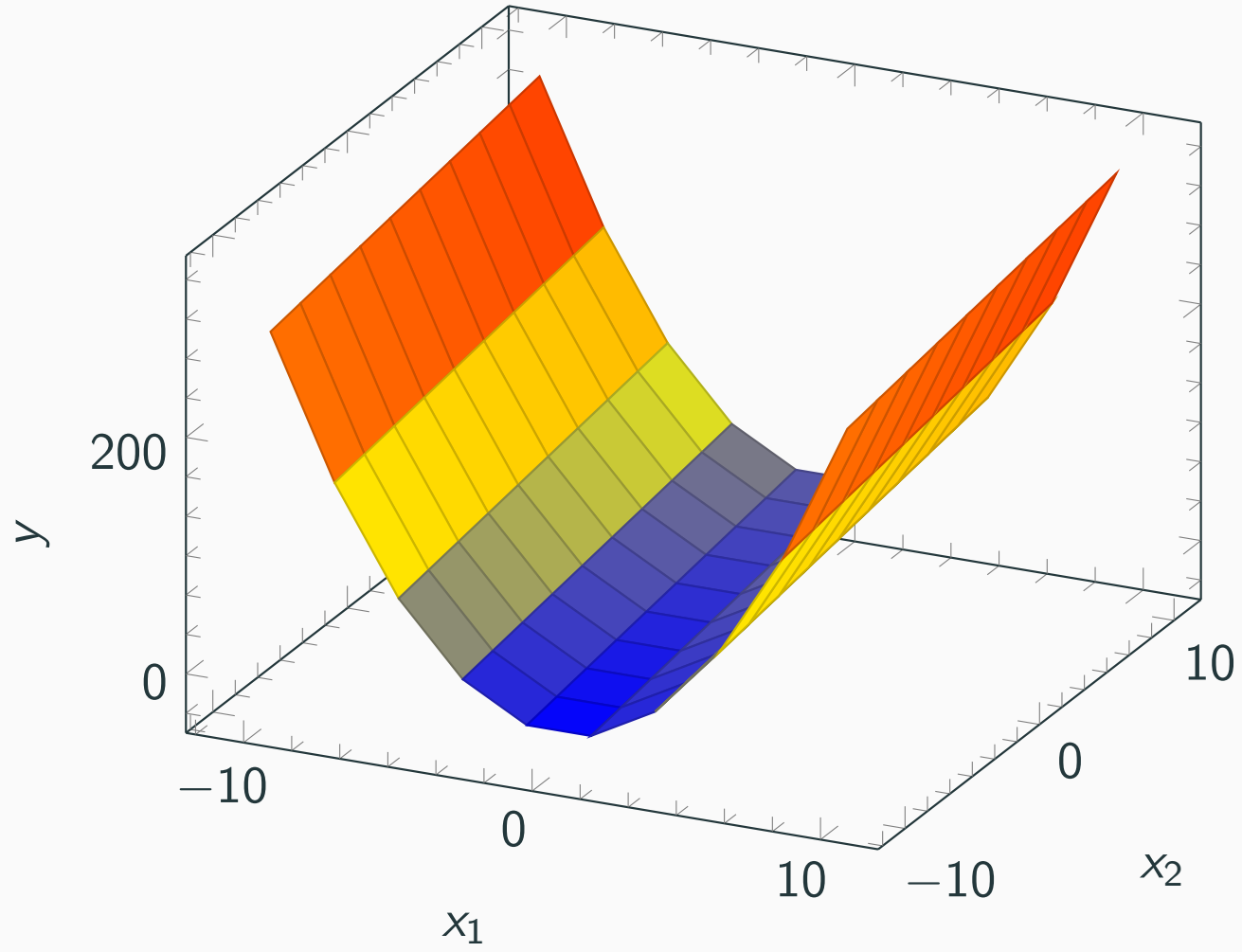
The α parameter is between 0 and 1, where closer to 1 returns the result that is closer to the one given by ridge regression, 0 for lasso.

What should be the values λ be?

It's not easy to find the best value of λ, but we can use cross-validation method to test at least a few values of λ and compare the results.

Stochastic Gradient Descent

In many libraries the way how the weights are found using the stochastic gradient descent.



SGD – idea

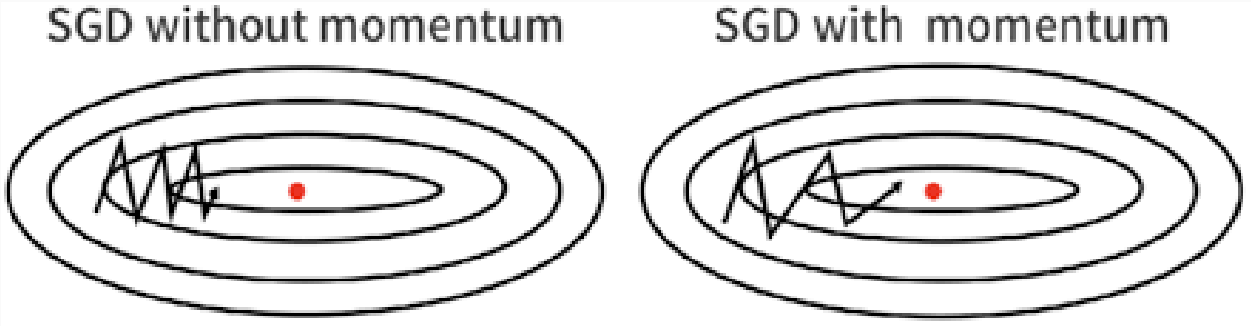


Figure 2: SGD, source: `https://paperswithcode.com/method/sgd`

SGD – idea

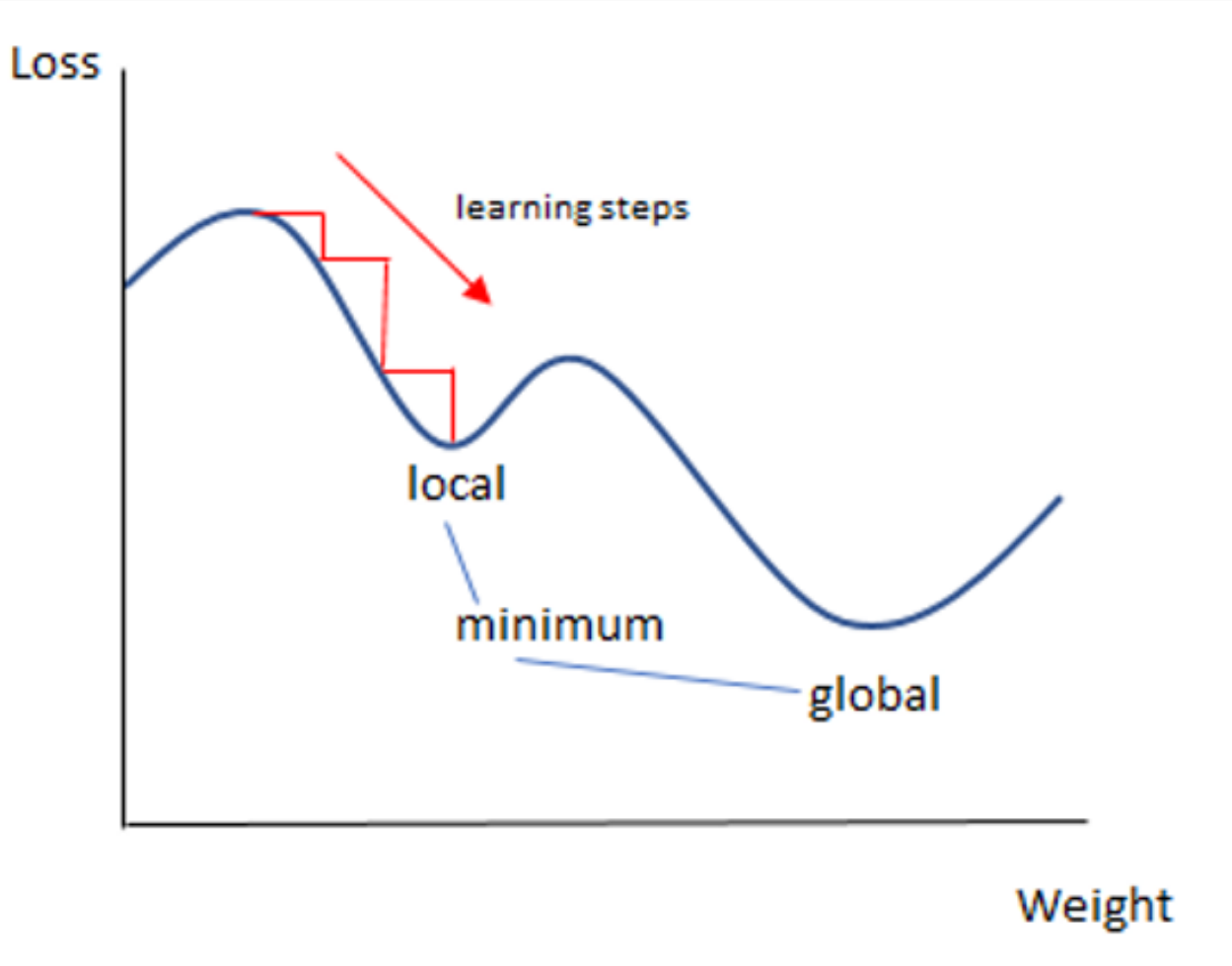


Figure 3: SGD with momentum, source:
`https://paperswithcode.com/method/sgd-with-momentum`

Stochastic Gradient Descent for linear regression

Cost function:

$$J(w, b) = \sum_{i=1}^{n} (y - haty), \tag{10}$$

where $\hat{y}$ is our prediction.

New weights:

$$w_j = w_i - r(\frac{\partial J}{\partial w}), \tag{11}$$

where r is the learning rate. The partial derivatives are defined as:

$$\frac{\partial J}{\partial w} = \sum_{i=2}^{n} -2x_i(y - \hat{y}) \tag{12}$$

. For w_0 it is similar:

$$w_{0j} = w_{0i} - r(\frac{\partial J}{\partial w_0}), \tag{13}$$

.

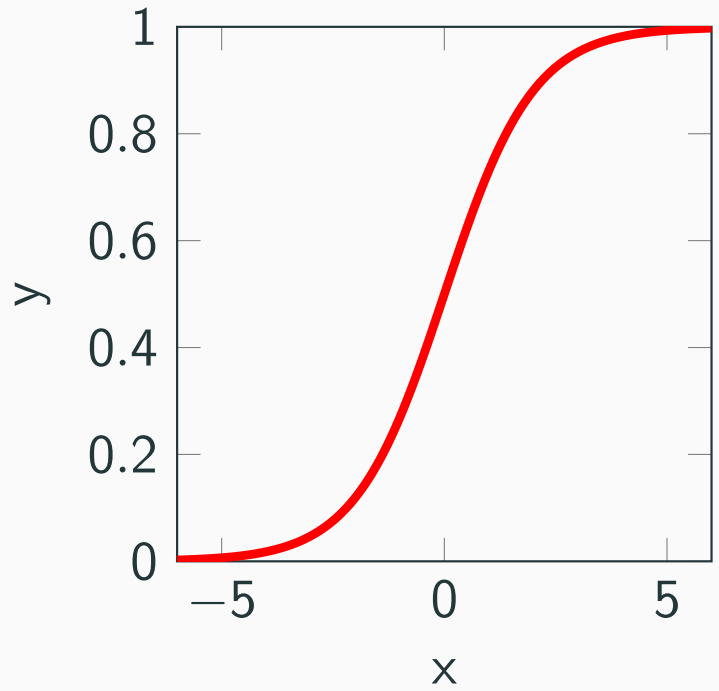
Logistic regression

Logistic regression

A different type of regression is logistic regression. Logistic regression is based on a logistic function. It is very useful especially when we do calculations based on probability theory, because logistic function gives values from 0 to 1, so it easily correspond with probability. It can be calculated as follows:

$$\hat{y}_{log} = \frac{e^{a+bx_i}}{1+e^{a+bx_i}} = \frac{1}{1+e^{-(a+bx_i)}}. \tag{14}$$

Logistic function



Logistic regression

As in linear regression we need to find parameters for each problem separately. In logistic regression it is a vector of parameters that is called weights:

$$w = [b, a]. \tag{15}$$

Logistic regression – weights

In logistic regression the goal is to find those two parameters that gives the best representation of data of a given problem. It would be best if we could find such parameters that for each x_i set into equation:

$$\hat{y}_{log} = \frac{e^{a+bx_i}}{1+e^{a+bx_i}} = \frac{1}{1+e^{-(a+bx_i)}}. \tag{16}$$

we get the proper y_i.

Logistic regression – training

It can be done using one of the most known method of maximum likehood estimator which Newton-Raphson method is. It is a iterative method and needs more calculations to be done compared to linear regression. Weights are calculated in each iteration like following:

$$w_{k+1} = w_k + (X^T V X)^{-1} X^T (y - \hat{p}_i), \tag{17}$$

where k is the iteration number and V is a diagonal weight matrix. Diagonal weight matrix elements can be calculated like following:

$$v_{ii} = \hat{y}_{log_i}(1 - \hat{y}_{log_i}). \tag{18}$$

Logistic regression – training stop criterion

The loop can end because of two reasons. We can set a fixed number of iterations or we can set a value of weight difference between two iterations and end the loop when the change in each iteration is below that value. Usually it is set to 0.01 or lower.

Logistic regression – Example

We have six patients with a skin lesion. Three lesions are known not to be a cancer and other three are known to be a cancer.

$$y = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

We have two features that indicate if it's a cancer or not, so we need to estimate the value of three parameters and weights vector looks at the start like:

$$w = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}.$$

Logistic regression – Example

Let's construct our features vector. The first feature is the asymmetry of a lesion. It is a value from a range $0 - 2$ where 2 means a total asymmetry and 0 total symmetry. The second is the number of colors that are within the lesion. It is a value from a range $0 - 6$.

Logistic regression – Example

we need to multiply weights vector with features vector matrix. It means that it need to have three rows instead of two. We need to add one column at the beginning filled with 1. Let the X look like following:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 5 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 0 & 4 \\ 1 & 2 & 3 \end{bmatrix}.$$

Logistic regression – Example

We need to calculate $w^T x_i$ in the first place:

$$w^T x_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0.5 \end{bmatrix} = 0.$$

The next step it so calculate the diagonal elements of matrix V. Before that we need to calculate logistic regression value for each x_i:

$$y_{log_1} = \frac{e^0}{1 + e^0} = \frac{1}{2}.$$

Logistic regression – Example

The y_{log_i} values are the same for each x_i in the first iteration. The same with diagonal elements:

$$v_{11} = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}.$$

The weight matrix looks now like following:

$$V_0 = \begin{bmatrix} 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 \end{bmatrix}.$$

Logistic regression – Example

In the last step we need to calculate new weights values. It is a bigger computation of multiplied matrices and vectors, so we divide it into few parts to keep in clear and understandable.

Logistic regression – Example

In the first place let's calculate $X^T V_0$:

$$X^T V_0 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 5 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 0 & 4 \\ 1 & 2 & 3 \end{bmatrix}^T \cdot \begin{bmatrix} 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0.5 & 0.25 & 0.25 & 0 & 0.5 \\ 0 & 1.25 & 0.5 & 0.75 & 1 & 0.75 \end{bmatrix}$$

Logistic regression – Example

Once we are done with it we need to calculate the output vector and current p_0 which is a vector that consist of $p_0(i) = \frac{e^{w^T x_i}}{1+e^{w^T x_i}}$ i elements for each x_i:

$$y - p = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} - \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix} =$$
$$= \begin{bmatrix} -0.5 & 0.5 & -0.5 & -0.5 & 0.5 & 0.5 \end{bmatrix}$$

Logistic regression – Example

The next step is to multiply inverted matrix $(X^T VX)^{-1}$ with X^T:

$$(X^T VX)^{-1} X^T = \begin{bmatrix} 2.91 & -0.32 & -0.68 \\ -0.32 & 1.37 & -0.37 \\ -0.68 & -0.37 & 0.37 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 5 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 0 & 4 \\ 1 & 2 & 3 \end{bmatrix}^T =$$

$$= \begin{bmatrix} 2.91 & -1.12 & 1.23 & 0.55 & 0.2 & 0.23 \\ -0.32 & 0.57 & 0.31 & -0.06 & -1.8 & 1.31 \\ -0.68 & 0.43 & -0.31 & 0.06 & 0.8 & -0.31 \end{bmatrix}$$

Logistic regression – Example

The last step in an iteration is to multiply two matrices that we have just calculated together:

$$(X^T VX)^{-1} X^T (y - p) = \begin{bmatrix} 2.91 & -1.12 & 1.23 & 0.55 & 0.2 & 0.23 \\ -0.32 & 0.57 & 0.31 & -0.06 & -1.8 & 1.31 \\ -0.68 & 0.43 & -0.31 & 0.06 & 0.8 & -0.31 \end{bmatrix} \cdot$$

$$\cdot \begin{bmatrix} -0.5 & 0.5 & -0.5 & -0.5 & 0.5 & 0.5 \end{bmatrix} =$$

$$= \begin{bmatrix} -2.69 & 0.077 & 0.92 \end{bmatrix}$$

Logistic regression – Example

As the previous weights vector was filled with zeros we the current weight vectors like following:

$$w = \begin{bmatrix} -2.69 & 0.077 & 0.92307692 \end{bmatrix}.$$

After three iteration we get the following weight vector:

$$w = \begin{bmatrix} -10.02052458 & 1.22700068 & 2.86618458 \end{bmatrix}.$$

Logistic regression – Example

We could also do some more iterations. Now check the logistic regression value for each x_i. It is shown in the table below.

$\mathbf{x_i}$	$-(\alpha + \beta_1 x_i + \beta_2 x_2)$	$y_{log}(x_i)$	y_i
$(0, 0)$	10.02	0.00004	0
$(2, 5)$	$-(-10.02 + 2.46 + 14.35) = 6.78$	0.99	1
$(1, 2)$	$-(-10.02 + 1.23 + 5.74) = -3.05$	0.045	0
$(1, 3)$	$-(-10.02 + 1.23 + 8.61) = -3.23$	0.038	0
$(0, 4)$	$-(-10.02 + 11.48) = 1.46$	0.811	1
$(2, 3)$	$-(10.02 + 2.46 + 8.61) = 2.51$	0.924	1

Table 3: Results after three iteration of example

Logistic regression – Example

The presented results indicate that logistic regression can be useful in some cases. Each x_i that is a cancer has a high value of logistic regression, close to 1. Benign lesions' logistic regression value is close to 0.

Logistic regression – Example

We can draw it as presented in figure below.

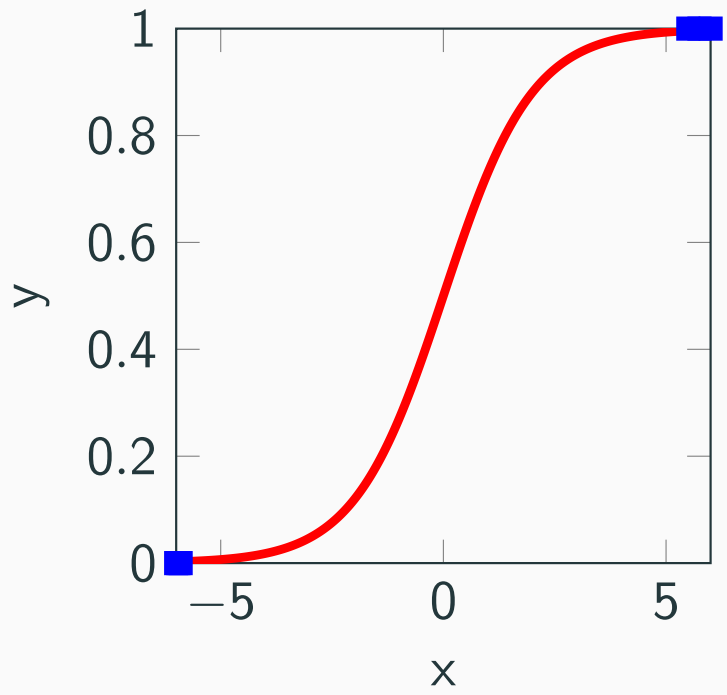
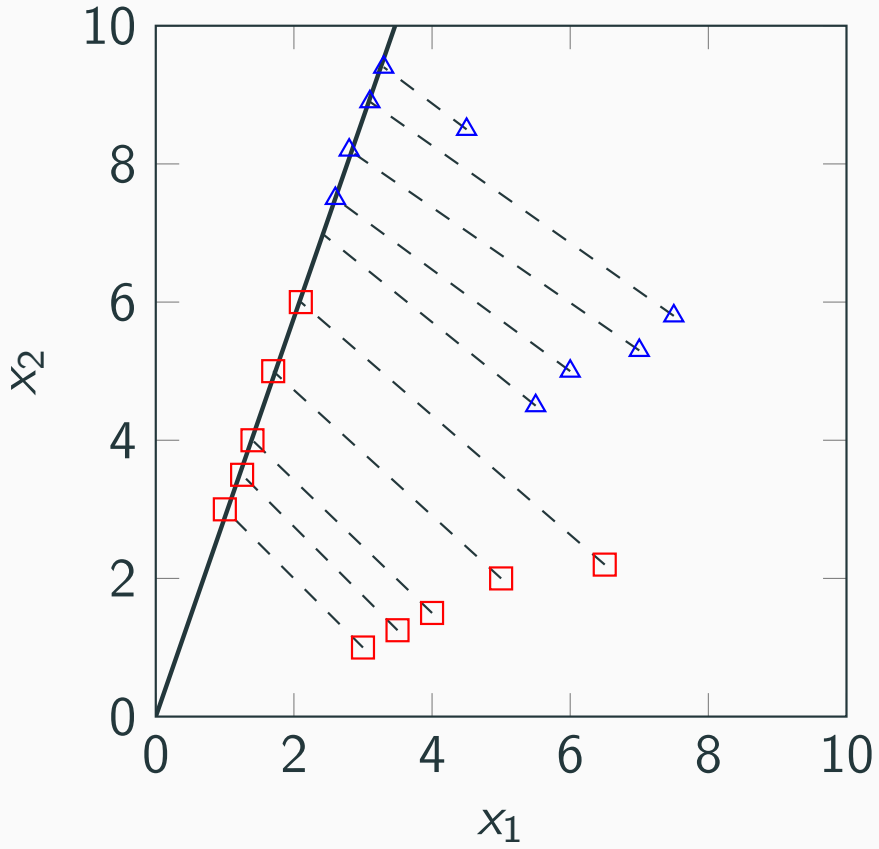


Figure 4: Logistic regression of skin lesion diagnosis example

Linear Discriminate Analysis

Fisher's linear classifier (LDA) – Idea

The idea of the Fisher's classifier is to move the data into a reduced dimension feature space and do the classification there.



LDA – the math behind

The whole process is to calculate the between-class variance with the class means (m_1, m_2), and the within-class variance (S_w, S_i). The means can be calculated as:

$$m_i = \frac{1}{n_i} \sum_{x \in i} x \tag{19}$$

The within-class variance can be calculated as:

$$S_i = \frac{1}{n_i - 1} \sum_{x \in i} (x - m_i)(x - m_i^T), \tag{20}$$

and

$$S_W = \frac{1}{n_{-1} + n_{+1} - 2} ((n_{-1} - 1) S_{-1} (n_{+1} - 1) S_{+1}). \tag{21}$$

Finally, the weight can be calculated as:

$$w = S_1^{-1} (m_{+1} - m_{-1}). \tag{22}$$

LDA – the math behind

The discriminant function can be written as

$$\hat{g}(x) = w_F^T x = (m_{+1} - m_{-1})^T S_W^{-1} x, \tag{23}$$

where:

$$w_0 = \frac{1}{2}(w^T m_{+1} + w^T m_{-1}). \tag{24}$$

This brings us to:

$$\hat{g}(x) = w^T x - w_0 \begin{cases} > 0, x \in 1, \\ < 0, x \in -1. \end{cases} \tag{25}$$

Questions?