Machine Learning

2024 (ML-2024)

Lecture 2. Linear models

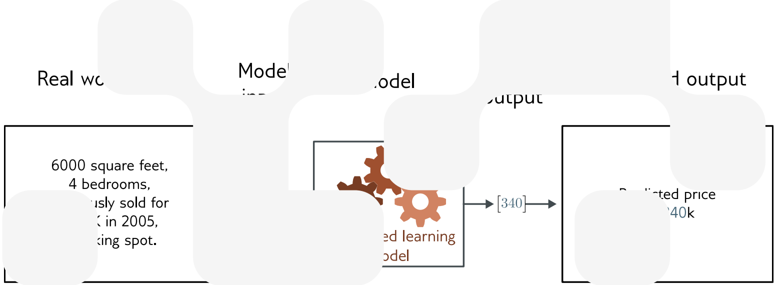
by Alexei Valerievich Kornaev, Dr. habil. in Eng. Sc.,
Researcher at the RC for AI, Assoc. Prof. of the Robotics and CV Master’s Program, Innopolis University
Researcher at the RC for AI, National RC for Oncology n.a. NN Blohin
Professor at the Dept. of Mechatronics, Mechanics, and Robotics, Orel State University

Agenda

I. Linear Regression and its Generalization

II. Logistic Regressioin and its Generalization

III. Setting of the models



All models are wrong, but some are useful.
/George Box/

Books

Handbook on Machine Learning by M. Artemyev et al., Yandex, 2022 (in Russian)
Understanding Deep Learning by Simon J.D. Prince, 2024
Practical Deep Learning / FastAI book by Jeremy Howard
Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016.

Online platforms, courses, resources

Sirius online courses on ML (in Russian)
Stepik online courses (in Russian)
Hugging Face online courses
Coursera is unavailable so far

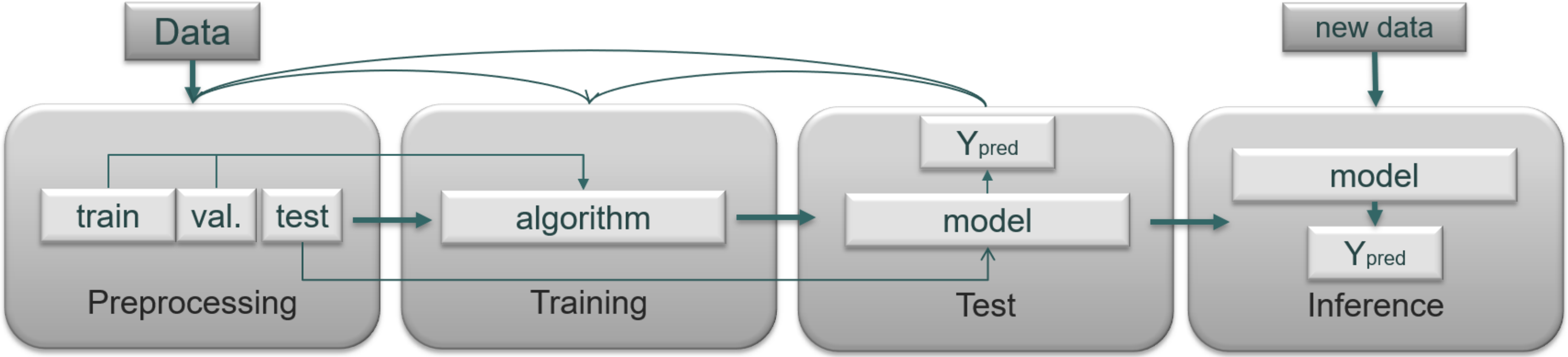
MIT Introduction to Deep Learning, MIT, 2024
Lecture Hall of the Faculty of Applied Mathematics and Informatics (in Russian)
Fast AI, courses, software, book by Jeremy Howard
Deep Learning, course by Semyon Kozlov (in Russian), 2019

3Blue1Brown, Animated Math
PyTorch Tutorial by Patrick Loeber, 2020

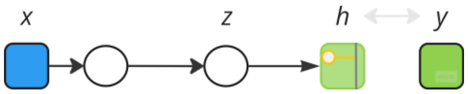
#someLinks

Read here: https://arxiv.org/, https://scholar.google.ru/
Collect the references here: https://mendeley.com/
Draw here: https://miro.com/app/dashboard/
Write the text here: https://www.overleaf.com/project
Write the code here: https://colab.research.google.com/
Collect the code here: https://github.com/
Find the journal here: https://journalfinder.elsevier.com/
Find the conference here: https://portal.core.edu.au/conf-ranks/?search=A

Flowchart for an ML model design



Linear Regression



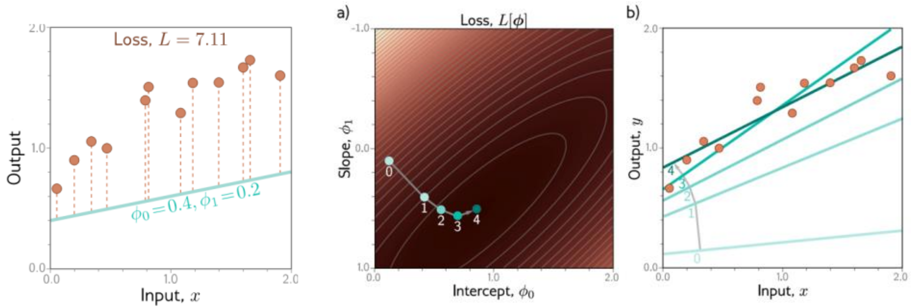
Model predicts output *h* given input *x*

$h = \phi_0 + \phi_1 x$

$$\boldsymbol{x} = \begin{bmatrix} x^{(1)} \\ \dots \\ x^{(m)} \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(m)} \end{bmatrix}, \boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix}.$$

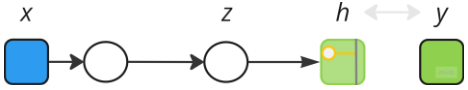
Consider a model $\boldsymbol{f} = [x^{(i)}, \boldsymbol{\phi}]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

$$\boxed{L(\boldsymbol{\phi}) = \frac{1}{2m}\sum_{i=1}^{m}\left(h^{(i)} - y^{(i)}\right)^2 \Rightarrow \min.}$$



Supervised learning intuition: S. J. Prince. Understanding Deep Learning. MIT Press, 2023. URL http://udlbook.com.

Linear Regression



Model predicts output *h* given input *x*

Consider a model $\boldsymbol{f} = [x^{(i)}, \boldsymbol{\phi}]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

add

$$\boldsymbol{x} = \begin{bmatrix} x^{(1)} \\ \dots \\ x^{(m)} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & x^{(1)} \\ \dots & \dots \\ 1 & x^{(m)} \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(m)} \end{bmatrix}, \boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix}.$$

$$h = x\phi = \begin{bmatrix} \phi_0 + \phi_1 x^{(1)} \\ \vdots \\ \phi_0 + \phi_1 x^{(m)} \end{bmatrix}$$

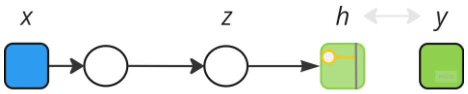
$$\boxed{L(\boldsymbol{\phi}) = \frac{1}{2m}\sum_{i=1}^{m}\left(h^{(i)} - y^{(i)}\right)^2 \Rightarrow \min.}$$

$$\boxed{\nabla L = \left[\frac{\partial L}{\partial \phi_j}\right] = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h} - \boldsymbol{y}).}$$

Training algorithm.

1. Initialize the *weights* $\boldsymbol{\phi}$ with a random seed
2. Calculate the *hypothesis* matrix $\boldsymbol{h} = \boldsymbol{x\phi}$ and the *loss gradient*: $\nabla L = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h} - \boldsymbol{y})$
3. For the given ϕ_j components (annotated with idex 'prev', ϕ_j^{prev}) calculate the newer ones ϕ_j^{next} moving towards the direction, which is opposite to the loss gradient vector, with steps which are proportional to the *learning rate* α:
$\boldsymbol{\phi}^{next} = \boldsymbol{\phi}^{prev} - \alpha\nabla L$, or in the scalar form $\phi_0^{\text{next}} = \phi_0^{prev} - \alpha\frac{\partial L}{\partial \phi_0}$, $\phi_1^{\text{next}} = \phi_1^{prev} - \alpha\frac{\partial L}{\partial \phi_1}$
4. Repeat pp. 2-3 until the minimum of the loss function L is reached, based on the condition of small changes in its value over several neighboring iterations or based on the condition of reaching the maximum number of iterations : $L^{next} - L^{prev} < \delta$, #iter. > max # of iter
5. Save the trained model (model weights): $\boldsymbol{\phi}$.

Linear Regression. Generalization (multiple var., polynomial)



Model predicts output h given input x

Consider a model $\boldsymbol{f} = \left[x^{(i)}, \boldsymbol{\phi}\right]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

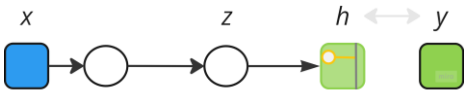
$$L(\boldsymbol{\phi}) = \frac{1}{2m}\sum_{i=1}^{m}\left(h^{(i)} - y^{(i)}\right)^2 \Rightarrow \min.$$

$$h = \phi_0 + \phi_1 x_1 + \ldots + \phi_n x_n$$

$$\boldsymbol{x} = \begin{bmatrix} 1 & x_1^{(1)} & \ldots & x_n^{(1)} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(m)} & \ldots & x_n^{(m)} \end{bmatrix}; \quad \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}; \quad \boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_n \end{bmatrix}; \quad \boldsymbol{h} = \boldsymbol{x}\boldsymbol{\phi}$$

Training algorithm.
1. Initialize $\boldsymbol{\phi}$
2. Calculate $\boldsymbol{h} = \boldsymbol{x\phi}$ and $\nabla L = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h} - \boldsymbol{y})$
3. Update $\boldsymbol{\phi}^{next} = \boldsymbol{\phi}^{prev} - \alpha \nabla L$
4. Repeat pp. 2-3 $L^{next} - L^{prev} < \delta$, #iter. > max # of iter
5. Save the trained model (model weights): $\boldsymbol{\phi}$.

Linear Regression. Generalization (multiple var., polynomial)



Model predicts output *h* given input *x*

Consider a model $\boldsymbol{f} = [x^{(i)}, \boldsymbol{\phi}]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

$$L(\boldsymbol{\phi}) = \frac{1}{2m}\sum_{i=1}^{m}\left(h^{(i)} - y^{(i)}\right)^2 \Rightarrow \min.$$

$$h = \varphi_0 + \varphi_1 \underbrace{x}_{x_1} + \varphi_2 \underbrace{x^2}_{x_2} + \dots + \varphi_n \underbrace{x^n}_{x_n}$$

$h = \varphi_0 + \varphi_1 x_1 + \dots + \varphi_n x_n$ → check the previous task

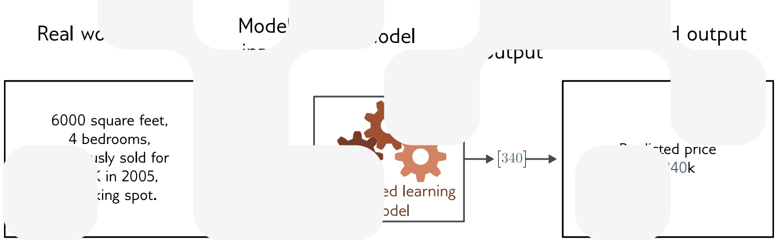
Training algorithm.
1. Initialize $\boldsymbol{\phi}$
2. Calculate $\boldsymbol{h} = \boldsymbol{x\phi}$ and $\nabla L = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h} - \boldsymbol{y})$
3. Update $\boldsymbol{\phi}^{next} = \boldsymbol{\phi}^{prev} - \alpha\nabla L$
4. Repeat pp. 2-3 $L^{next} - L^{prev} < \delta$, #iter. > max # of iter
5. Save the trained model (model weights): $\boldsymbol{\phi}$.

Agenda

I. Linear Regression and its Generalization

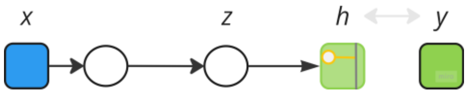
II. Logistic Regressioin and its Generalization

III. Setting of the models



All models are wrong, but some are useful.
/George Box/

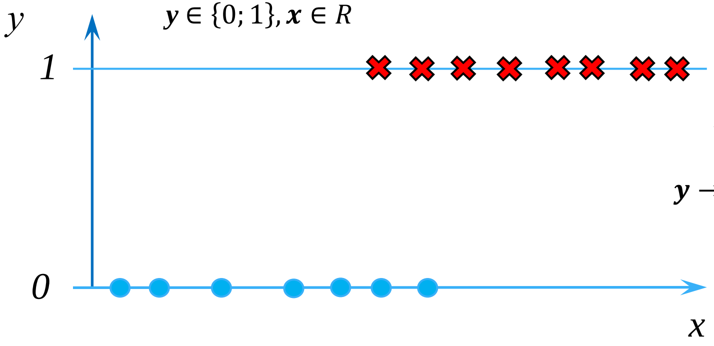
Logistic Regression



Model predicts output h given input x

Consider a model $\boldsymbol{f} = [x^{(i)}, \boldsymbol{\phi}]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

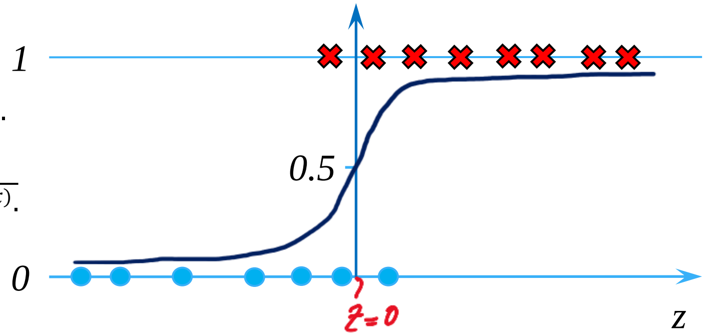
$$L(\boldsymbol{\phi}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\ln(h^{(i)}) + (1 - y^{(i)})(\ln(1 - h^{(i)})\right) \Rightarrow \min.$$



transform

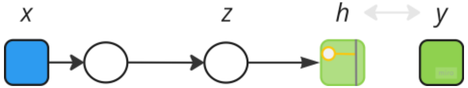
$\boldsymbol{x} \rightarrow \boldsymbol{z}$: $z = \phi_0 + \phi_1 x$.

$\boldsymbol{y} \rightarrow \boldsymbol{h}$: $h(z) = \dfrac{1}{1 + e^{-z(x)}}$.



$$\boldsymbol{x} = \begin{bmatrix} x^{(1)} \\ \dots \\ x^{(m)} \end{bmatrix}; \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(m)} \end{bmatrix}; \rightarrow \quad \boldsymbol{x} = \begin{bmatrix} 1 & x^{(1)} \\ \dots & \dots \\ 1 & x^{(m)} \end{bmatrix}; \boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix}; \boldsymbol{z} = \boldsymbol{x\phi}; \quad \rightarrow \quad h(z) = \frac{1}{1+e^{-z(x)}}, \text{ or } \boldsymbol{h} = \sigma(\boldsymbol{z}).$$

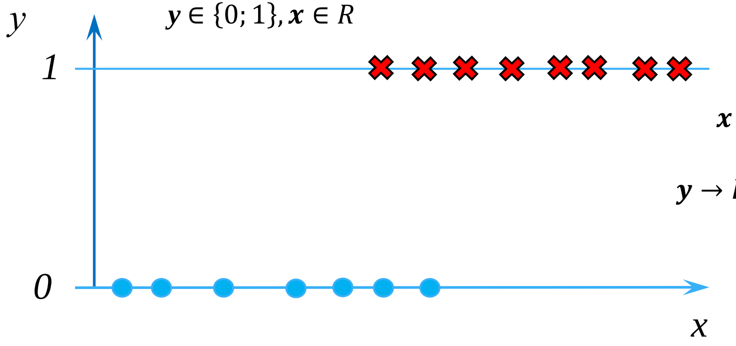
Logistic Regression



Model predicts output *h* given input *x*

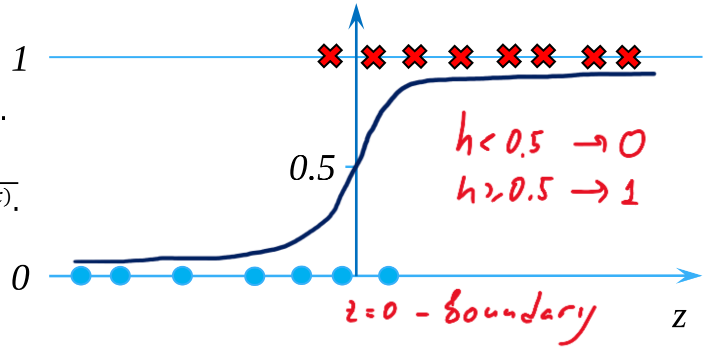
Consider a model $\boldsymbol{f} = [x^{(i)}, \boldsymbol{\phi}]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

$$\boxed{L(\boldsymbol{\phi}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\ln(h^{(i)}) + (1-y^{(i)})(\ln(1-h^{(i)})\right) \Rightarrow \min.}$$



$\boldsymbol{x} \to \boldsymbol{z}$: $z = \phi_0 + \phi_1 x$.

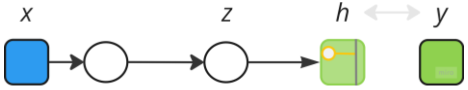
$\boldsymbol{y} \to \boldsymbol{h}$: $h(z) = \dfrac{1}{1+e^{-z(x)}}$.



Training algorithm.

1. Initialize $\boldsymbol{\phi}$
2. Calculate $\boldsymbol{z} = \boldsymbol{x\phi}$, $\boldsymbol{h} = \sigma(\boldsymbol{z})$, then $\nabla L = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h}-\boldsymbol{y})$
3. Update $\boldsymbol{\phi}^{next} = \boldsymbol{\phi}^{prev} - \alpha\nabla L$
4. Repeat pp. 2-3 $L^{next} - L^{prev} < \delta$, #iter. > max # of iter
5. Save the trained model (model weights): $\boldsymbol{\phi}$.

Logistic Regression. Generalization (multiple var., polynomial)



Model predicts output *h* given input *x*

Consider a model $\boldsymbol{f} = [x^{(i)}, \boldsymbol{\phi}]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

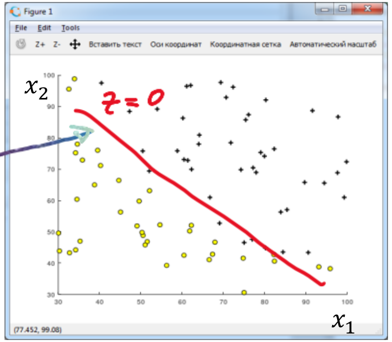
$$z = \phi_0 + \phi_1 x_1 + \phi_2 x_2 , \quad h = \sigma(z).$$

$$\boxed{L(\boldsymbol{\phi}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\ln(h^{(i)}) + (1-y^{(i)})(\ln(1-h^{(i)})\right) \Rightarrow \min.}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ \dots & \dots \\ x_1^{(m)} & x_2^{(m)} \end{bmatrix}; \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(m)} \end{bmatrix}; \quad \rightarrow \quad \boldsymbol{x} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ \dots & \dots & \dots \\ 1 & x_1^{(m)} & x_2^{(m)} \end{bmatrix}; \boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \end{bmatrix}; \boldsymbol{z} = \boldsymbol{x}\boldsymbol{\phi}; \quad \rightarrow \quad h(z) = \frac{1}{1+e^{-z(x)}}, \text{ or } \boldsymbol{h} = \sigma(\boldsymbol{z}).$$

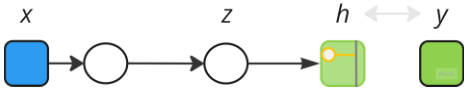
boundary : $0 = \phi_0 + \phi_1 x_1 + \phi_2 x_2$

$$x_2 = -\frac{\phi_0}{\phi_2} - \frac{\phi_1}{\phi_2} x_1$$



Training algorithm.
1. Initialize $\boldsymbol{\phi}$
2. Calculate $\boldsymbol{z} = \boldsymbol{x}\boldsymbol{\phi}$, $\boldsymbol{h} = \sigma(\boldsymbol{z})$, then $\nabla L = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h} - \boldsymbol{y})$
3. Update $\boldsymbol{\phi}^{next} = \boldsymbol{\phi}^{prev} - \alpha\nabla L$
4. Repeat pp. 2-3 $L^{next} - L^{prev} < \delta$, #iter. > max # of iter
5. Save the trained model (model weights): $\boldsymbol{\phi}$.

Logistic Regression. Generalization (multiple var., polynomial)



Model predicts output h given input x

Consider a model $\boldsymbol{f} = \left[x^{(i)}, \boldsymbol{\phi}\right]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

$$\boxed{L(\boldsymbol{\phi}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\ln(\,h^{(i)}) + (1-y^{(i)})(\ln(\,1-h^{(i)})\right) \Rightarrow \min.}$$

$h = \phi_0 + \phi_1 x_1 + \ldots + \phi_n x_n$

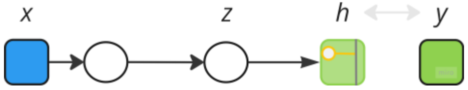
Gray scale picture of "Nine"

$$\boldsymbol{x} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & & x_n^{(1)} \\ \ldots & \ldots & \ldots & \ldots \\ x_1^{(m)} & x_2^{(m)} & & x_n^{(m)} \end{bmatrix}; \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \ldots \\ y^{(m)} \end{bmatrix}; \quad \rightarrow \boldsymbol{x} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & & x_n^{(1)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_1^{(m)} & x_2^{(m)} & & x_n^{(m)} \end{bmatrix}; \boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \ldots \\ \phi_n \end{bmatrix}; \boldsymbol{z} = \boldsymbol{x\phi};\ \boldsymbol{h} = \sigma(\boldsymbol{z}).$$

Training algorithm.

1. Initialize $\boldsymbol{\phi}$
2. Calculate $\boldsymbol{z} = \boldsymbol{x\phi}$, $\boldsymbol{h} = \sigma(\boldsymbol{z})$, then $\nabla L = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h} - \boldsymbol{y})$
3. Update $\boldsymbol{\phi}^{next} = \boldsymbol{\phi}^{prev} - \alpha\nabla L$
4. Repeat pp. 2-3 $L^{next} - L^{prev} < \delta$, #iter. > max # of iter
5. Save the trained model (model weights): $\boldsymbol{\phi}$.

Logistic Regression. Generalization (multiple var., polynomial)



Model predicts output *h* given input *x*

Consider a model $\boldsymbol{f} = [x^{(i)}, \boldsymbol{\phi}]$ parameterized with weights $\boldsymbol{\phi}$ that maps each i-th input sample $x^{(i)}$ into the output $z^{(i)}$ which then transforms into the hypothesis $h^{(i)}$ that should be close to the label $y^{(i)}$.

$$L(\boldsymbol{\phi}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\ln(h^{(i)}) + (1-y^{(i)})(\ln(1-h^{(i)})\right) \Rightarrow \min.$$

$z = \phi_0 + \phi_1 x + \phi_2 x^2$

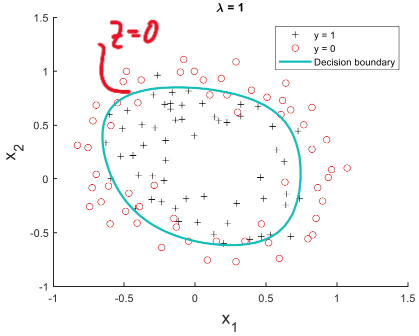
$x \to x_1$, $x^2 \to x_2$

$$\boldsymbol{x} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ \dots & \dots \\ x_1^{(m)} & x_2^{(m)} \end{bmatrix}; \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(m)} \end{bmatrix}; \quad \to$$

$z = \phi_0 + \phi_1 x_1 + \phi_2 x_2$ - Check the previous task.

More general:

$z = \phi_0 + \phi_1 x_1 + \dots + \phi_n x_1^n + \phi_{n+1} x_2 + \dots \phi_k x_2^n$



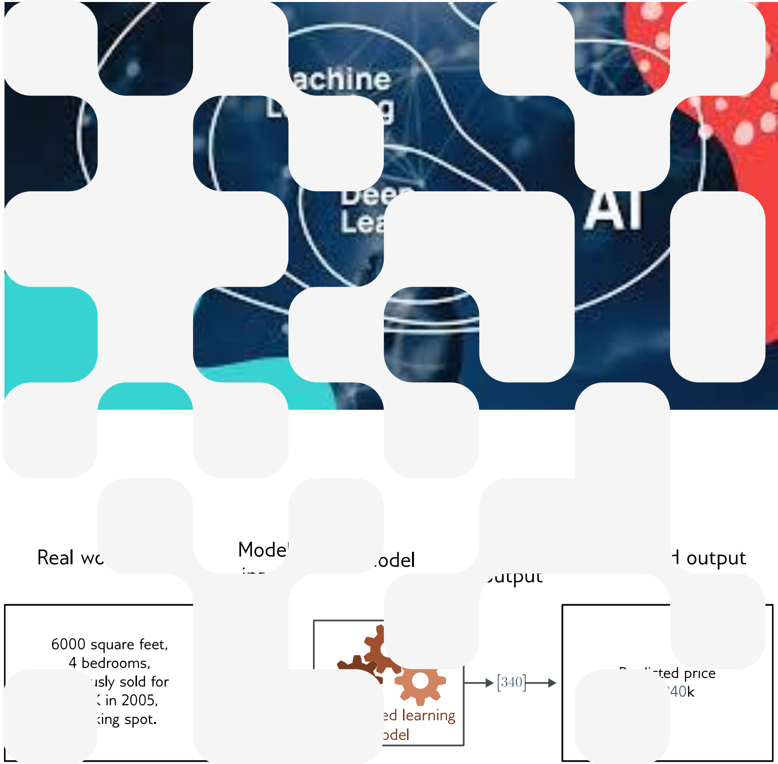
Training algorithm.
1. Initialize $\boldsymbol{\phi}$
2. Calculate $\boldsymbol{z} = \boldsymbol{x\phi}$, $\boldsymbol{h} = \sigma(\boldsymbol{z})$, then $\nabla L = \frac{1}{m}\boldsymbol{x}^T(\boldsymbol{h} - \boldsymbol{y})$
3. Update $\boldsymbol{\phi}^{next} = \boldsymbol{\phi}^{prev} - \alpha \nabla L$
4. Repeat pp. 2-3 $L^{next} - L^{prev} < \delta$, #iter. > max # of iter
5. Save the trained model (model weights): $\boldsymbol{\phi}$.

Agenda

I. Linear Regression and its Generalization

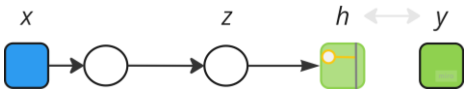
II. Logistic Regressioin and its Generalization

III. Setting of the models



All models are wrong, but some are useful.
/George Box/

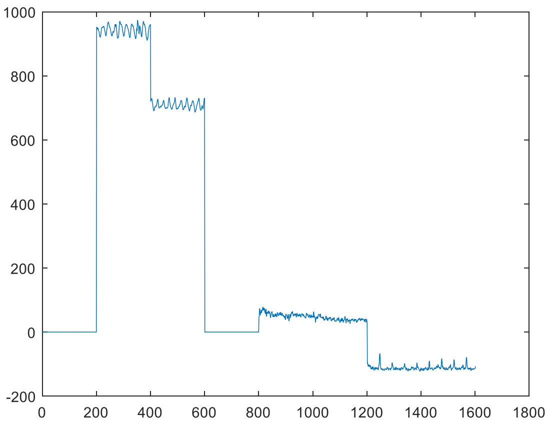
ML Settings

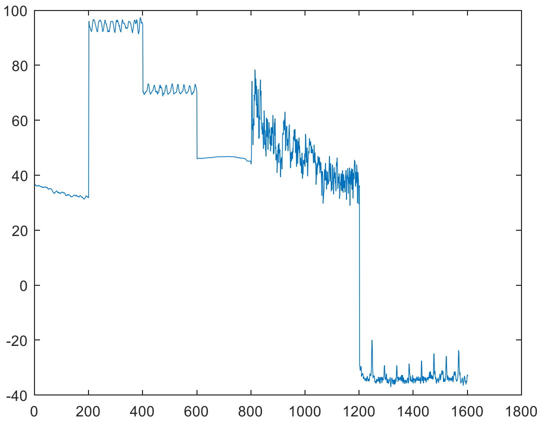


Model predicts output *h* given input *x*

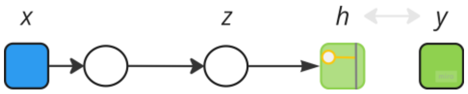
Model *parameters* are determined during the solution of the ML problem. For example, in regression problems, the parameters are the components of the matrix of weights $\boldsymbol{\phi}$. *Hyperparameters* are set by the user, usually not in a single way, and their values affect the values of the sought parameters.

1. Feature Scaling





ML Settings



Model predicts output *h* given input *x*

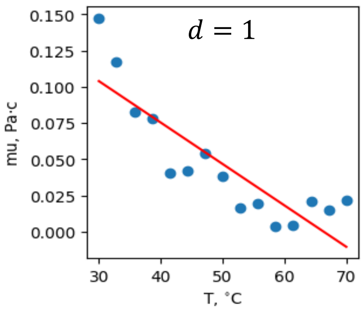
1. Feature Scaling
2. Learning Rate
3. Error and # of iterations
4. **Regularization (L2)**

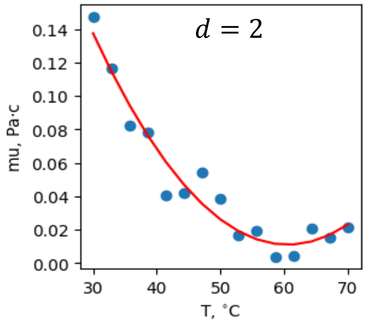
Model *parameters* are determined during the solution of the ML problem. For example, in regression problems, the parameters are the components of the matrix of weights $\boldsymbol{\phi}$. *Hyperparameters* are set by the user, usually not in a single way, and their values affect the values of the sought parameters.

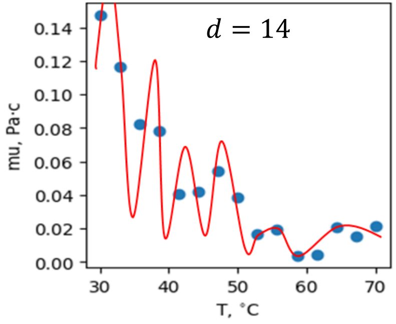
Training $\{(x_i, y_i)\}$	validation	test

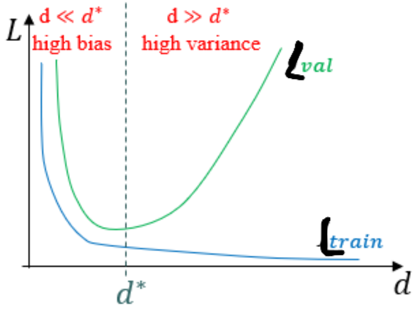
$$L = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h^{(i)} - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \phi_j^2\right] \Rightarrow \min.$$

$$h(x) = \theta_j x^j \quad ,(j = 0,..d)$$









Just think about it

1. How can the gradient descent method be improved to find global minima instead of local ones?
2. Can the discussed linear regression problems be solved analytically without using the gradient descent method?
3. Why is the use of high-degree polynomials generally not recommended when building regression models?

Thank you for your attention!

a.kornaev@innopolis.ru, @avkornaev

Notes