Machine Learning: short introduction (trees and NN)

Sergey Korpachev[1, 2] on behalf of Dépôt

[1]Moscow Institute of Physics and Technology
[2]Lebedev Physical Institute of the Russian Academy of Sciences

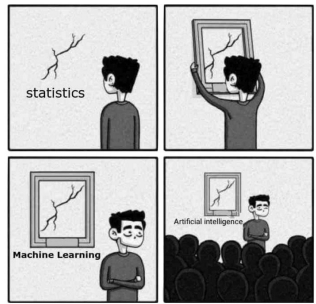
Outline

- Machine learning (ML) - what is it?
- Data
- ML pipeline
- Linear model
- Decision tree
- Neural network

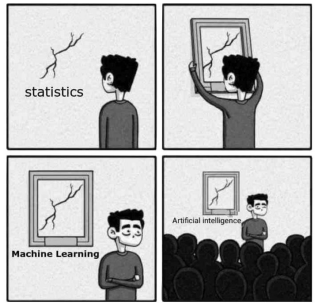
Machine learning

Machine learning

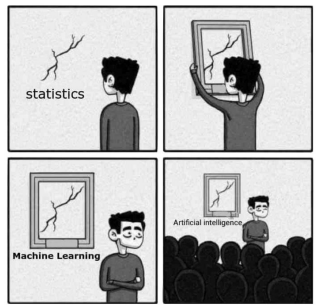
what is it?

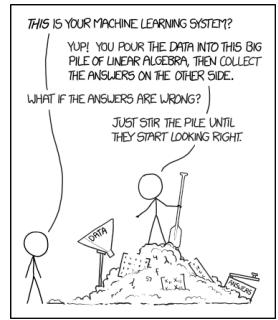


Machine learning — what is it?



Machine learning — what is it?





- Compare your understanding of machine learning now and one (for example) year from now. This will be your understanding.

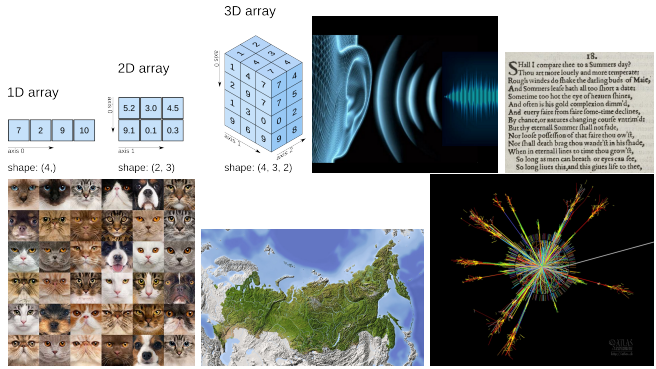
Data

Data

what is data?

Anything can be data:

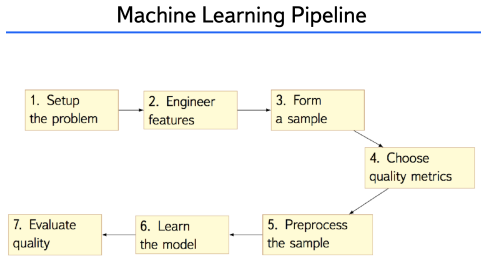
1. Numbers
2. Text
3. Images
4. Sound
5. Geomap
6. Particle collisions
7. Knowledge
8. You name it



ML pipeline

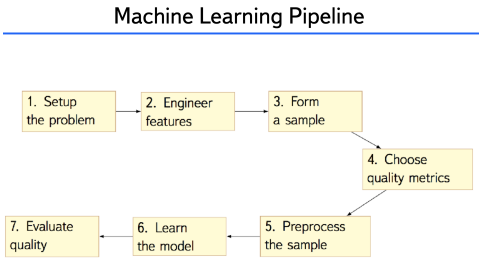
ML pipeline

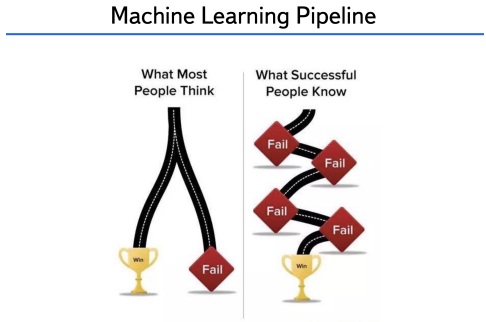
what is ML pipeline?



ML pipeline

what is ML pipeline?





Linear model

Linear model

regression

- $Y = \mathbb{R}$
- N objects with K real features: $D = \mathbb{R}^K$
- $a(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^{K} f_j(\boldsymbol{x}) \cdot \theta_j$
- Extend and reassign:
  $[1, f_1(\boldsymbol{x}), \ldots, f_K(\boldsymbol{x})] \equiv \boldsymbol{x}$
  $[\theta_0, \ldots, \theta_K] \equiv \boldsymbol{\theta}$
- Then $a(\boldsymbol{x}, \boldsymbol{\theta}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$

- Minimization problem:
  - $\mathcal{L}(\boldsymbol{\theta}) = \big(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle - y\big)^2$
  - $Q(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \big(\langle \boldsymbol{x_i}, \boldsymbol{\theta} \rangle - y_i\big)^2$
- Then we need to minimize $Q(\boldsymbol{\theta})$ by varying $\boldsymbol{\theta}$:
  - $\hat{a} = \arg\min_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta})$

Linear model

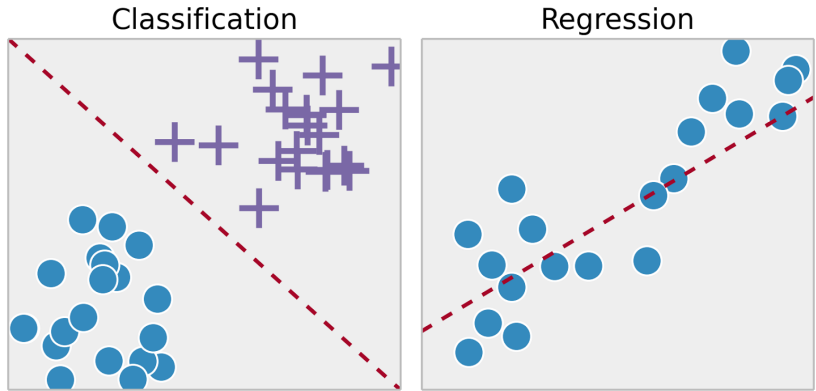
classification

- $Y = \{-1, +1\}$
- N objects with K real features: $D = \mathbb{R}^K$
- $a(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^{K} f_j(\boldsymbol{x}) \cdot \theta_j$
- Extend and reassign:
  $[1, f_1(\boldsymbol{x}), \ldots, f_K(\boldsymbol{x})] \equiv [1, x_1, \ldots, x_K] \equiv \boldsymbol{x}$
  $[\theta_0, \ldots, \theta_K] \equiv \boldsymbol{\theta}$
- Then $a(\boldsymbol{x}, \boldsymbol{\theta}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$

- Minimization problem:
  - $\mathcal{L}(\boldsymbol{\theta}) = \left[\text{sign}\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \neq y\right]$
  - $Q(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \left[\text{sign}\langle \boldsymbol{x_i}, \boldsymbol{\theta} \rangle \neq y_i\right]$
- Then we need to minimize $Q(\boldsymbol{\theta})$ by varying $\boldsymbol{\theta}$:
  - $\hat{a} = \arg\min_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta})$

Linear model

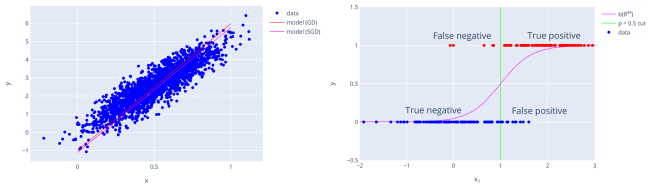
classification and regression



Decision tree

Decision tree

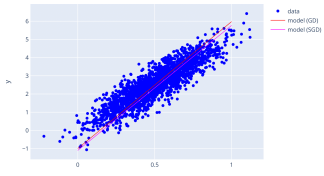
motivation

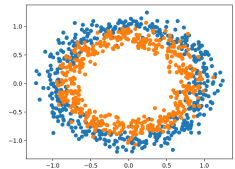


- In the previous part we explored a family of linear models:
  - ➔ nicely describe linear correlations
  - ➔ fast and easy to train
  - ➔ performing well with large number of samples/features

Decision tree

motivation

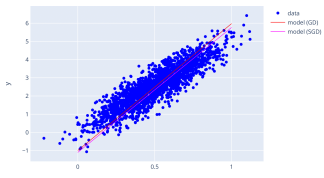


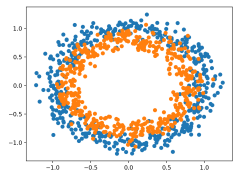


- In the previous part we explored a family of linear models:
  - ➔ nicely describe linear correlations
  - ➔ fast and easy to train
  - ➔ performing well with large number of samples/features
- However, they are poor in modelling non-linearities
- Additional heuristics might be applied (e.g. polynomial features) but are ad-hoc

Decision tree

motivation





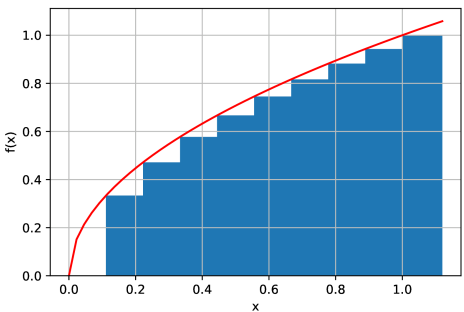
- In the previous part we explored a family of linear models:
  - ➔ nicely describe linear correlations
  - ➔ fast and easy to train
  - ➔ performing well with large number of samples/features
- However, they are poor in modelling non-linearities
- Additional heuristics might be applied (e.g. polynomial features) but are ad-hoc

Is there a more general way to describe non-linearities?

Decision tree

motivation

- What is the simplest way to approximate a function?
- Using **piecewise linear function**

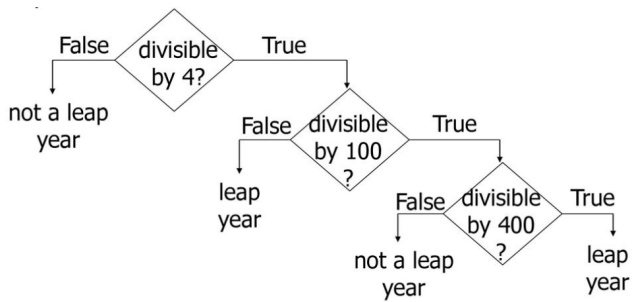


$$f(x) = \begin{cases} 0, & x < 0.12 \\ 0.35, & 0.12 \leq x < 0.22 \\ 0.47, & 0.22 \leq x < 0.36 \\ \dots \end{cases}$$

Decision tree

intuition

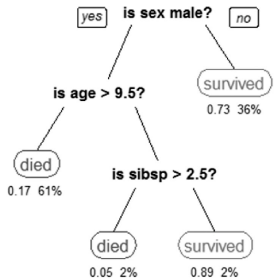
- What is the simplest way to categorize things?
- Ask **yes/no questions**



Decision tree

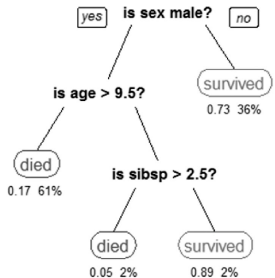
intuition

- What is the simplest way to categorize things?
- Ask **yes/no questions**



Decision tree — intuition

- What is the simplest way to categorize things?
- Ask **yes/no questions**



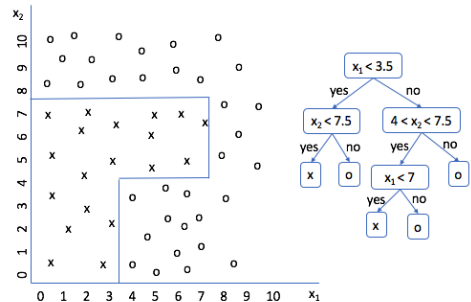
Can we mathematically formulate this?

Decision tree — algorithm

Algorithm 1: Decision tree

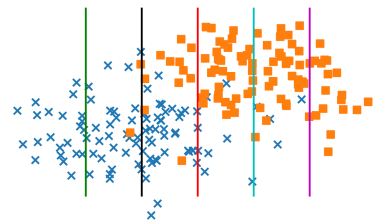
1: Initialize: hyperparameters, leaf set = $\{X_{\text{train}}\}$

2: **while** not stopping criteria **do**
3: leaf to split = Choose(leaf set)
4: left leaf, right leaf = Split(leaf to split)
5: Add left leaf, right leaf to leaf set
6: Remove leaf to split from leaf set
7: **end while**



Decision tree — split criteria

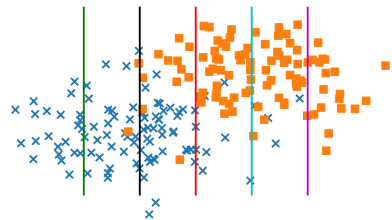
- Suppose there is a classification problem
- So we want to separate one class from the other by constructing **decision boundary**
- And using decision tree algorithm described earlier
- To do that we need to start splitting on features
- Which **split** is the best?



Decision tree — split criteria

- Suppose there is a classification problem
- So we want to separate one class from the other by constructing **decision boundary**
- And using decision tree algorithm described earlier
- To do that we need to start splitting on features
- Which **split** is the best?

→ We need some criteria



Decision tree

split criteria

more about IG

- Suppose we have n features $x^{(1)}, x^{(2)}, ..., x^{(n)}$ in our dataset X
- If we want to split on some feature $x^{(j)}$ at threshold t then we propose to maximize **Information Gain** (IG):

$$\mathrm{IG}(\mathrm{X}, \mathrm{a}) = \mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X} \mid \mathrm{a}) \to \max_{\mathrm{a}}$$

- In case of a binary tree with L(R) elements in the left (right) split we've got:

$$\mathrm{H}(j, t) = \frac{|\mathrm{L}|}{|\mathrm{X}|}\mathrm{H}(\mathrm{L}) + \frac{|\mathrm{R}|}{|\mathrm{X}|}\mathrm{H}(\mathrm{R}) \to \min_{j,t}$$

Decision tree

split criteria

more about IG

- Suppose we have n features $x^{(1)}, x^{(2)}, ..., x^{(n)}$ in our dataset X
- If we want to split on some feature $x^{(j)}$ at threshold t then we propose to maximize **Information Gain** (IG):

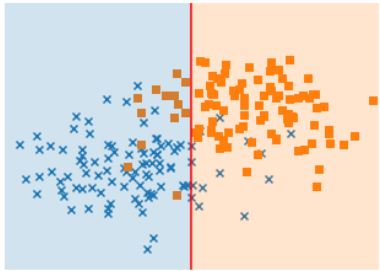
$$\mathrm{IG}(\mathrm{X}, \mathrm{a}) = \mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X} \mid \mathrm{a}) \to \max_{\mathrm{a}}$$

- In case of a binary tree with L(R) elements in the left (right) split we've got:

$$\mathrm{H}(j, t) = \frac{|\mathrm{L}|}{|\mathrm{X}|}\mathrm{H}(\mathrm{L}) + \frac{|\mathrm{R}|}{|\mathrm{X}|}\mathrm{H}(\mathrm{R}) \to \min_{j,t}$$

How to choose H(X)?

Decision tree — split criteria



The ”best” split is a central one because it introduces **purity** in the best way. And we have some functions to measure the purity!

Decision tree — split criteria

- Define H(R) in a leaf as a goodness of approximation with the ″best″ constant c

$$\mathrm{H}(\mathrm{R}) = \min_{c \in \mathbb{Y}} \frac{1}{|N|} \sum_{(x_i, y_i) \in \mathrm{R}} \mathcal{L}(y_i, c)$$

Decision tree — split criteria

- Define H(R) in a leaf as a goodness of approximation with the "best" constant c

$$\mathrm{H(R)} = \min_{c \in \mathbb{Y}} \frac{1}{|N|} \sum_{(x_i, y_i) \in \mathrm{R}} \mathcal{L}(y_i, c)$$

- or c_k with $\sum_k c_k = 1$, if tree predicts class probabilities

Decision tree — split criteria

- Define H(R) in a leaf as a goodness of approximation with the "best" constant c

$$\mathrm{H(R)} = \min_{c \in \mathbb{Y}} \frac{1}{|N|} \sum_{(x_i, y_i) \in \mathrm{R}} \mathcal{L}(y_i, c)$$

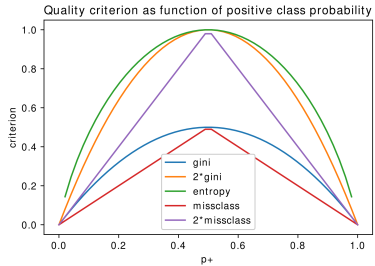
- or c_k with $\sum_k c_k = 1$, if tree predicts class probabilities
- Define a fraction of class k in a leaf as p_k

Decision tree — split criteria

- Define H(R) in a leaf as a goodness of approximation with the "best" constant c

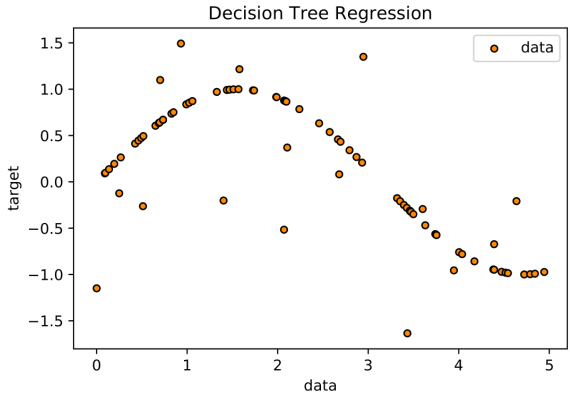
$$\mathrm{H(R)} = \min_{c \in \mathbb{Y}} \frac{1}{|N|} \sum_{(x_i, y_i) \in \mathrm{R}} \mathcal{L}(y_i, c)$$

- or c_k with $\sum_k c_k = 1$, if tree predicts class probabilities
- Define a fraction of class k in a leaf as p_k
- Then one can show:
  - **Misclassification criteria**
    $\mathcal{L}(y, c) = [y \neq c] \Rightarrow \mathrm{H(R)} = 1 - \max_k \{p_k\}$
  - **Gini Impurity**
    $\mathcal{L}(y, c) = \sum_k (c_k - [y = k])^2 \Rightarrow \mathrm{H(R)} = 1 - \sum_k (p_k)^2$
  - **Entropy criteria**
    $\mathcal{L}(y, c) = -\sum_k [y = k] \log c_k \Rightarrow \mathrm{H(R)} = -\sum_k p_k \log_2 p_k$



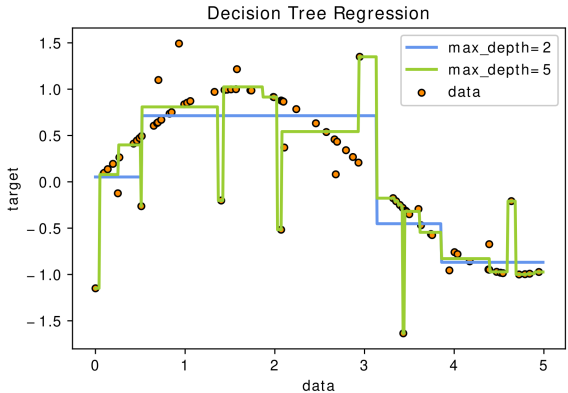
Lower diversity (higher purity)
$\Rightarrow$ **lower impurity** criterion H(R)

Decision tree — regression



Can we use decision tree approach for regression problem?

Decision tree — regression



Can we use decision tree approach for regression problem? Yes

Decision tree — regression

- How should we modify decision criteria?

$$\mathrm{H}(j, t) = \frac{|\mathrm{L}|}{|\mathrm{Q}|}\mathrm{H}(\mathrm{L}) + \frac{|\mathrm{R}|}{|\mathrm{Q}|}\mathrm{H}(\mathrm{R})$$

→ Mean squared error: $\mathcal{L}(y, c) = (y - c)^2 \Rightarrow \mathrm{H}(\mathrm{R}) = \min_c \frac{1}{|\mathrm{R}|} \sum_{(x_i, y_i) \in \mathrm{R}} (y_i - c)^2$

→ Mean absolute error: $\mathcal{L}(y, c) = |y - c| \Rightarrow \mathrm{H}(\mathrm{R}) = \min_c \frac{1}{|\mathrm{R}|} \sum_{(x_i, y_i) \in \mathrm{R}} |y_i - c|$

Decision tree — regression

- How should we modify decision criteria?

$$\mathrm{H}(j,t) = \frac{|\mathrm{L}|}{|\mathrm{Q}|}\mathrm{H}(\mathrm{L}) + \frac{|\mathrm{R}|}{|\mathrm{Q}|}\mathrm{H}(\mathrm{R})$$

→ Mean squared error: $\mathcal{L}(y,c) = (y-c)^2 \Rightarrow \mathrm{H}(\mathrm{R}) = \min_c \frac{1}{|\mathrm{R}|}\sum_{(x_i,y_i)\in \mathrm{R}} (y_i - c)^2$

→ Mean absolute error: $\mathcal{L}(y,c) = |y-c| \Rightarrow \mathrm{H}(\mathrm{R}) = \min_c \frac{1}{|\mathrm{R}|}\sum_{(x_i,y_i)\in \mathrm{R}} |y_i - c|$

- One can further show that:

$$c^* = \frac{1}{|\mathrm{R}|}\sum_{y_i \in \mathrm{R}} y_i$$

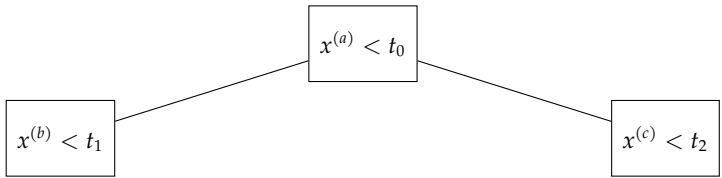
$$c^* = \mathrm{median}(y_i \in \mathrm{R})$$

What about tree construction?

$$x^{(a)} < t_0$$

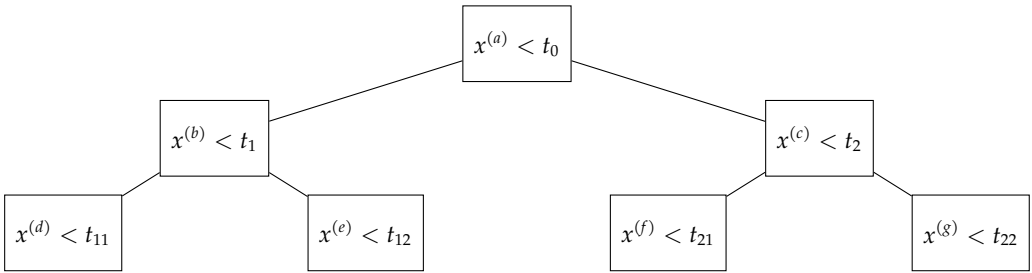
Decision tree — growing

What about tree construction? Do it recursively by **greedy** algorithm



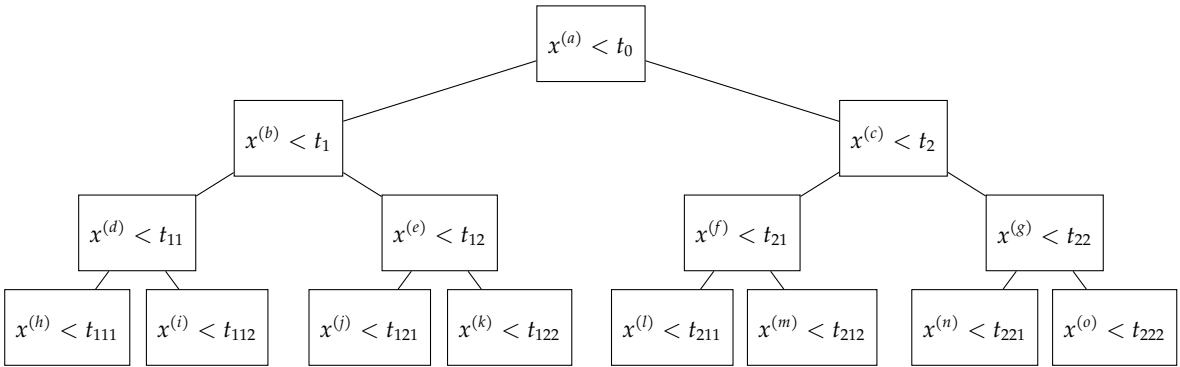
Decision tree — growing

What about tree construction? Do it recursively by **greedy** algorithm



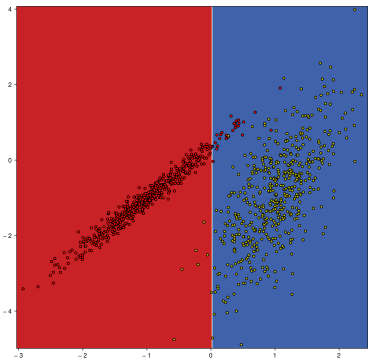
Decision tree — growing

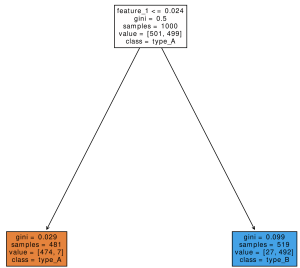
What about tree construction? Do it recursively by **greedy** algorithm



Decision tree — growing

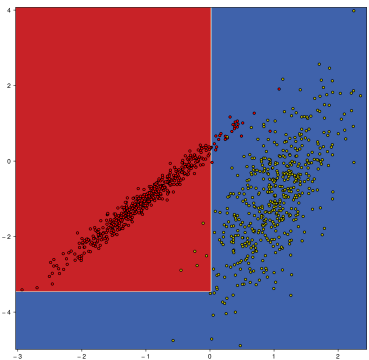
check out also these visuals

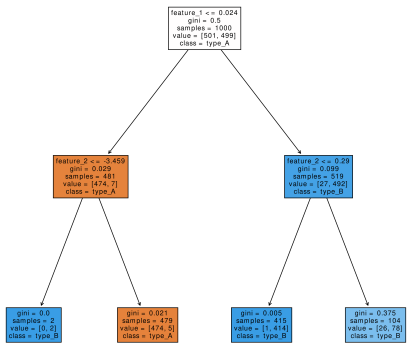




Decision tree — growing

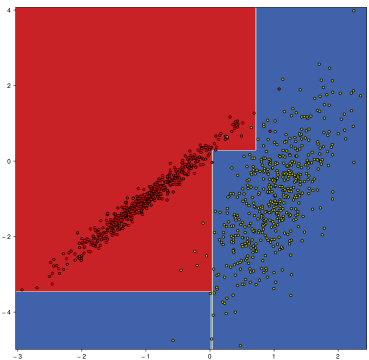
check out also these visuals

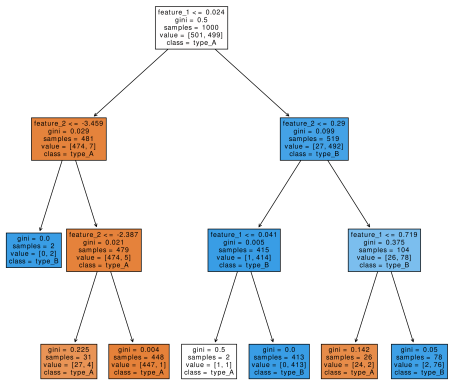




Decision tree — growing

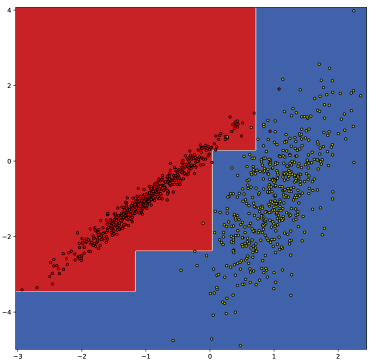
check out also these visuals

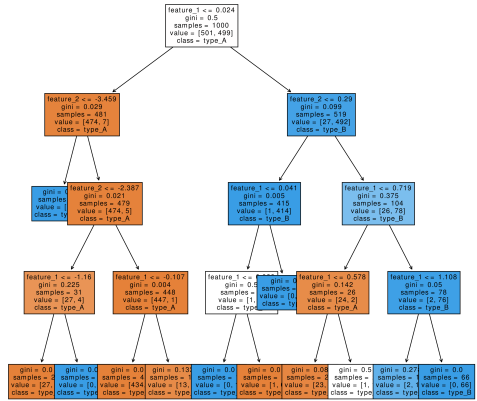




Decision tree — growing

check out also these visuals

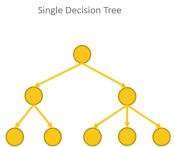


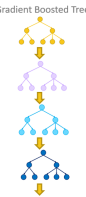


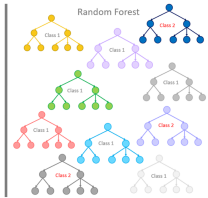
Random forest and gradient boosting

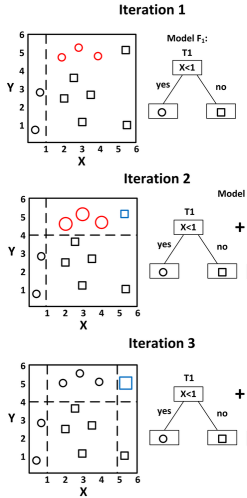
Random forest and gradient boosting

what's next after decision trees?





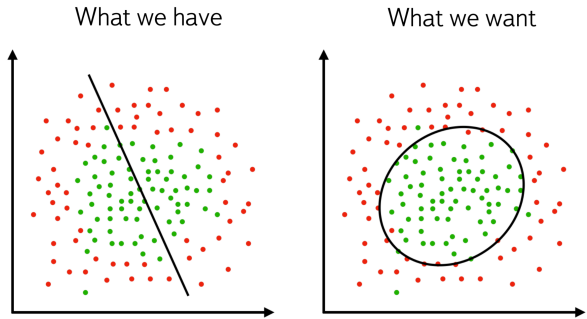




Modelling nonlinearities

Modelling nonlinearities

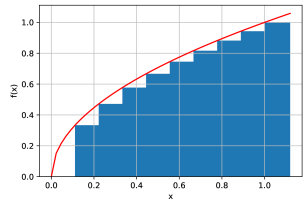
linear models



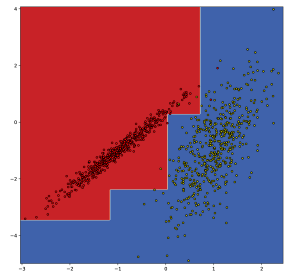
Linear models can't simply describe complex nonlinear data

Modelling nonlinearities

trees

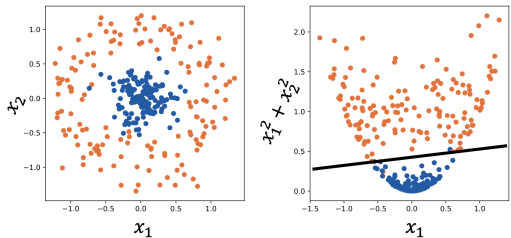


- (Ensembles of) Trees were designed to approximate nonlinearities and are **pretty good** in it + they are **fast and interpretable**
- But they are just ”brute-force” algorithms – **don’t infer symmetries** in data by design
- **Ad-hoc, cut-based and piecewise approximations** of data at hand + not differentiable and smooth



Modelling nonlinearities

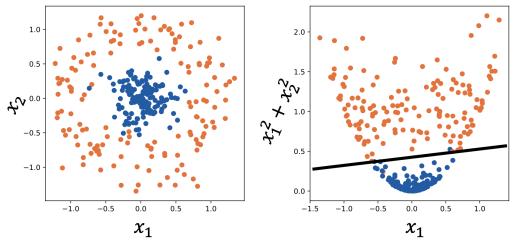
feature engineering



- But sometimes we know *a priori* that there are transformations simplifying the problem $\rightarrow$ even linear model can do the job
- However, this **feature engineering** is non-trivial, requires domain knowledge and is time-consuming

Modelling nonlinearities

feature engineering



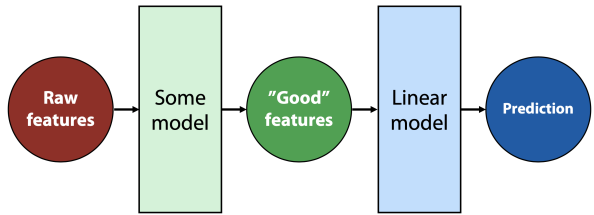
- But sometimes we know *a priori* that there are transformations simplifying the problem $\rightarrow$ even linear model can do the job
- However, this **feature engineering** is non-trivial, requires domain knowledge and is time-consuming

What if we design a model which could **automatically** feature-engineer itself?

Neural Network

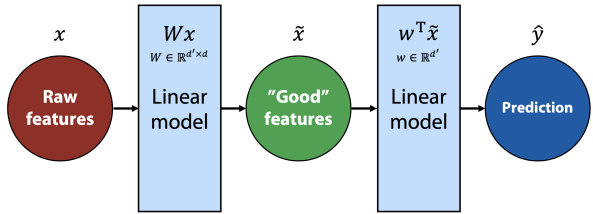
Neural Network — automating FE

NN illustrations from ML in HEP 2020



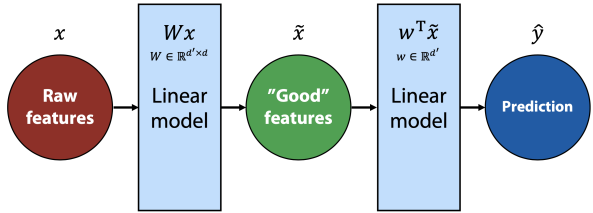
- Let's use a **simple linear model** to solve our supervised problem
- Add a block to a linear model which will automatically **generate new features** for it
- Two blocks would work together as a **single model** $\Rightarrow$ their parameters are updated simultaneously
- And **automatically**, by e.g. gradient descent (given their differentiability)

Neural Network — automating FE



- Would a linear model work as a feature generating model?

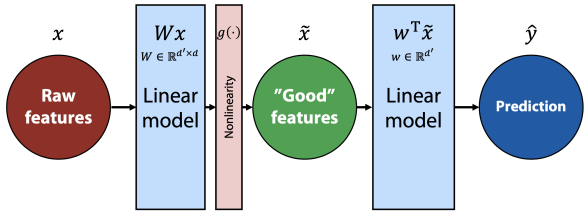
Neural Network — automating FE



- Would a linear model work as a feature generating model? No
- $\hat{y} = w^T\tilde{x} = w^T(Wx) = (w^TW)x = w'^Tx \Rightarrow$ it is still a linear model

→ Input feature space has not changed, only the model weights

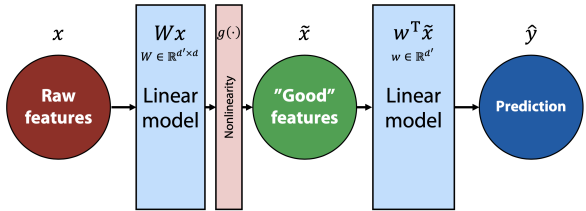
Neural Network — automating FE



- Let's then introduce **nonlinearity** to our model

➔ $\hat{y} = w^T \tilde{x} = w^T g(Wx)$,
where $g(\cdot)$ – some nonlinear scalar function (applied elementwise)

Neural Network — automating FE

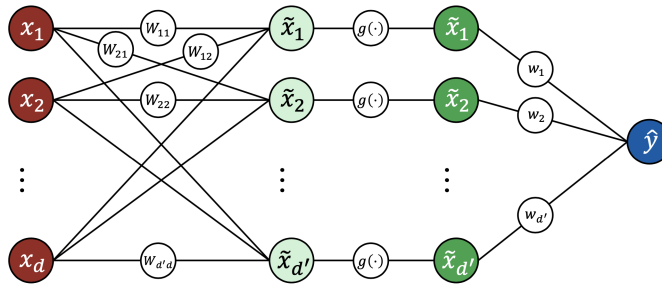


- Let's then introduce **nonlinearity** to our model

➔ $\hat{y} = w^T \tilde{x} = w^T g(Wx)$,
where $g(\cdot)$ – some nonlinear scalar function (applied elementwise)

➔ This is the simplest example of a **neural network**

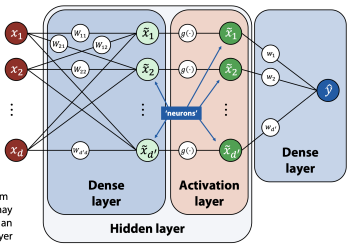
Neural Network — architecture



Essentially, NN is just a **composite function** that maps a set of X to a set of Y

$$\hat{y} = w^T \tilde{x} = w^T g(Wx)$$

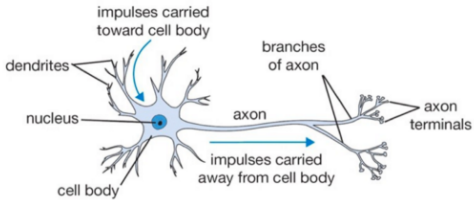
Feed-forward network:

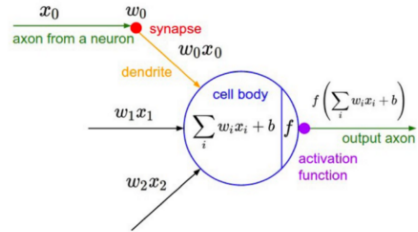


Note: the term "activation" may also stand for an output of a layer

- Brown nodes $x_1, x_2, ..., x_d$ – features from an **input layer**
- Green nodes $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_{d'}$– **neurons** from a **hidden layer**
- Blue node $\hat{y}$ – neuron from an **output layer**
- Straight lines (edges) between neurons – **weights** w_{ij}
- $g(\cdot)$ – nonlinear **activation** function, e.g. sigmoid $\sigma(\cdot)$
- Important: each neuron has additional **bias** b associated to it and added to other inputs (not illustrated)

Neural Network — human brain





Training

Training

how to train?

- Since NN fundamentally is a parametrized differentiable model we can optimise the loss function and use **gradient descent** to train it:

$$\boldsymbol{\omega}_{k+1} \leftarrow \boldsymbol{\omega}_k - \eta \cdot \nabla Q(\boldsymbol{\omega}_k)$$

- But gradients are hard to derive analytically – writing down all the derivatives is tough and tedious (especially for large NN)
- Note that NN is just a **composite model** $\Rightarrow$ can use **chain rule** for differentiating it

Training — chain rule

- Computing derivative of a "base" function is simple ⇒ decompose composite function into a set of base ones and differentiate them one by one
- Let's recall the rule:

$$\frac{\partial f\big(t(x)\big)}{\partial x} = \frac{\partial f(t)}{\partial t} \cdot \frac{\partial t(x)}{\partial x}$$

Exercise: can you compute derivative of sigmoid function by decomposing it into base functions?

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Training — chain rule

- Computing derivative of a "base" function is simple ⇒ decompose composite function into a set of base ones and differentiate them one by one
- Let's recall the rule:

$$\frac{\partial f\big(t(x)\big)}{\partial x} = \frac{\partial f(t)}{\partial t} \cdot \frac{\partial t(x)}{\partial x}$$

Exercise: can you compute derivative of sigmoid function by decomposing it into base functions?

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\sigma(x) = f(g(h(p(x)))), \quad f(z) = \frac{1}{z}, \quad g(z) = 1 + z, \quad h(z) = e^z, \quad p(z) = -z$$

Training — chain rule

- Computing derivative of a "base" function is simple ⇒ decompose composite function into a set of base ones and differentiate them one by one
- Let's recall the rule:

$$\frac{\partial f\big(t(x)\big)}{\partial x} = \frac{\partial f(t)}{\partial t} \cdot \frac{\partial t(x)}{\partial x}$$

Exercise: can you compute derivative of sigmoid function by decomposing it into base functions?

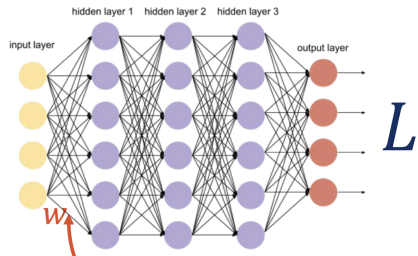
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\sigma(x) = f(g(h(p(x)))), \quad f(z) = \frac{1}{z}, \quad g(z) = 1 + z, \quad h(z) = e^z, \quad p(z) = -z$$

$$\sigma'(x) = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial h} \cdot \frac{\partial h}{\partial p} \cdot \frac{\partial p}{\partial x} = -\frac{1}{g^2} \cdot 1 \cdot e^p \cdot (-1) = \ldots = \sigma(x) \cdot \big(1 - \sigma(x)\big)$$

Training — backpropagation

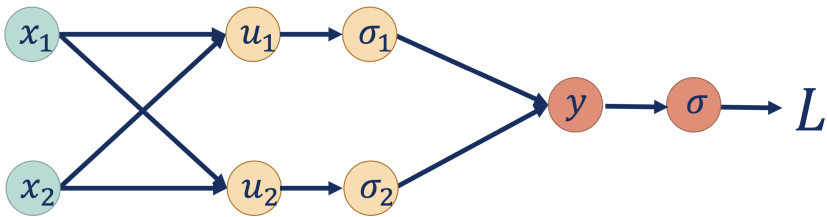
backprop illustrations from DMIA



- So how to find partial derivatives of loss function with respect to some weight?

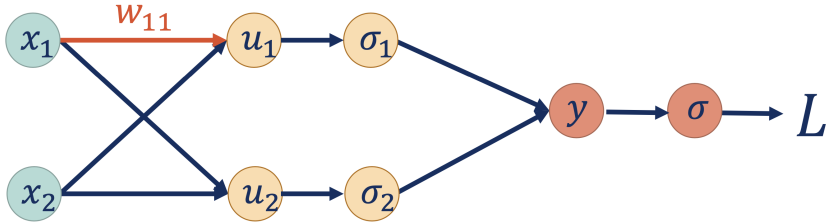
Training — backpropagation

backprop illustrations from DMIA



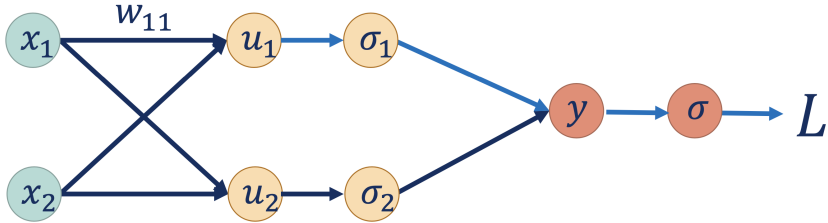
- Let's consider a simplified neural network
- Represent it in the form of a **computational graph**

Training — backpropagation



$$w_{11}^{(t+1)} = w_{11}^{(t)} - \gamma_t \frac{\partial L}{\partial w_{11}}$$

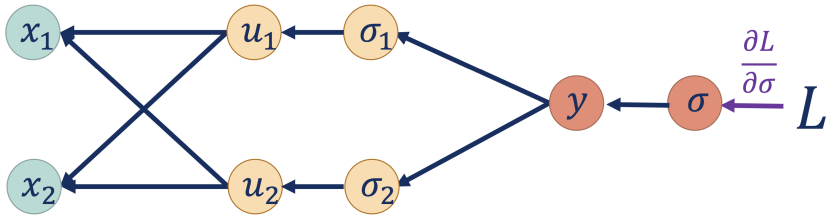
Training — backpropagation



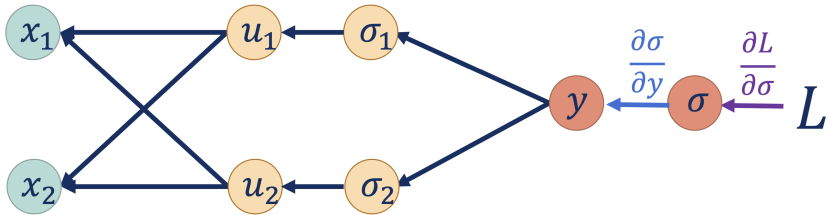
$$w_{11}^{(t+1)} = w_{11}^{(t)} - \gamma_t \frac{\partial L}{\partial w_{11}}$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial L}{\partial u_1}\frac{\partial u_1}{\partial w_{11}} = \frac{\partial L}{\partial u_1} x_1$$

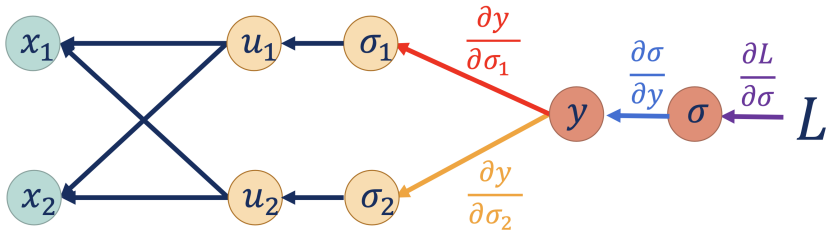
Training — backpropagation



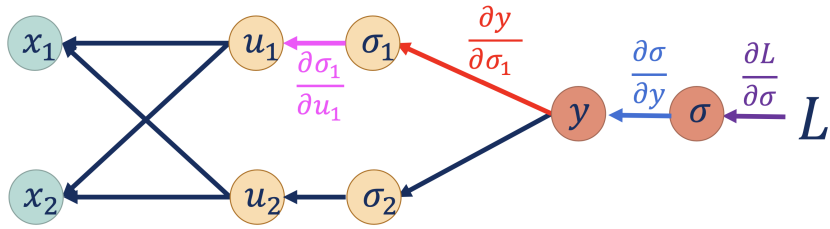
Training — backpropagation



Training — backpropagation

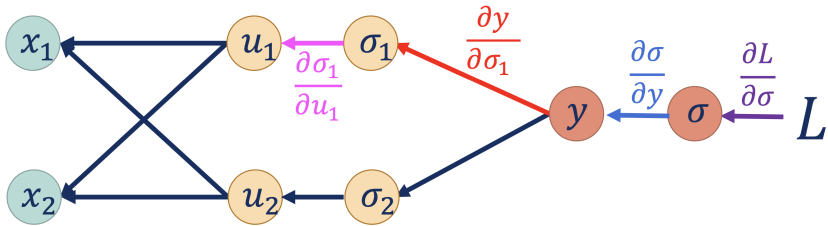


Training — backpropagation



$$\frac{\partial L}{\partial u_1} = \frac{\partial L}{\partial \sigma}\frac{\partial \sigma}{\partial y}\frac{\partial y}{\partial \sigma_1}\frac{\partial \sigma_1}{\partial u_1}$$

Training — backpropagation



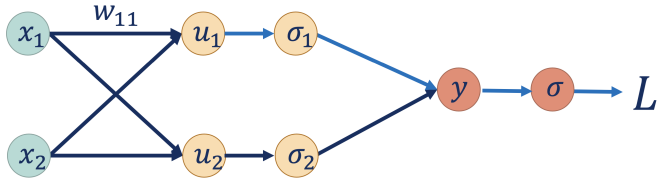
$$\frac{\partial L}{\partial u_1} = \frac{\partial L}{\partial \sigma}\frac{\partial \sigma}{\partial y}\frac{\partial y}{\partial \sigma_1}\frac{\partial \sigma_1}{\partial u_1}$$

- this procedure is called **backpropagation**
- its idea is to **collect derivatives** at each step in the computational graph w/o recalculating them every single time for every weight

Training

wrap it up

train NN in your browser!

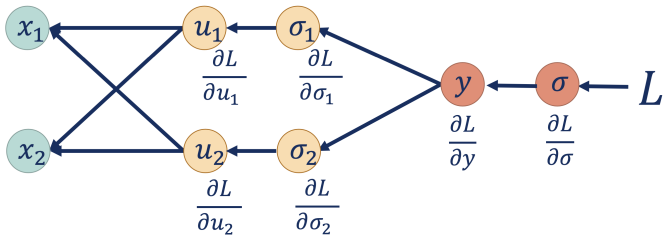


1. make **forward pass** through NN to calculate the output of each neuron and value of loss function

Training

wrap it up

train NN in your browser!

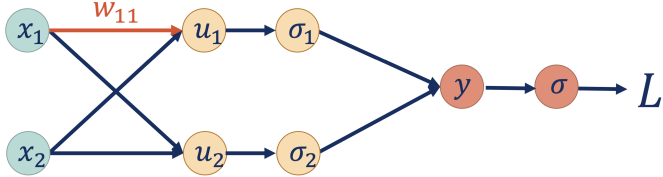


1. make **forward pass** through NN to calculate the output of each neuron and value of loss function
2. with **backward pass** go back through NN to calculate gradients for all weights

Training

wrap it up

train NN in your browser!



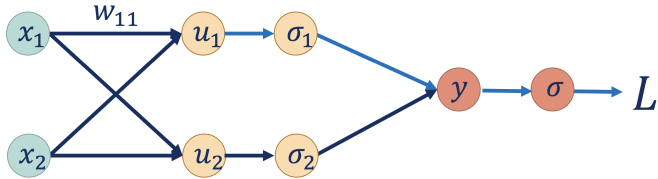
$$w_{11}^{(t+1)} = w_{11}^{(t)} - \gamma_t \frac{\partial L}{\partial w_{11}}$$

1. make **forward pass** through NN to calculate the output of each neuron and value of loss function
2. with **backward pass** go back through NN to calculate gradients for all weights
3. update weights with their gradients

Training

wrap it up

train NN in your browser!



1. make **forward pass** through NN to calculate the output of each neuron and value of loss function
2. with **backward pass** go back through NN to calculate gradients for all weights
3. update weights with their gradients
4. repeat until convergence

*one iteration (**epoch**) = forward and backward pass

Summary

Summary

- What is ML?
- Data
- Pipeline
- Linear model
- Decision tree
- Neural network

In our course, you can learn more about machine learning, related machine learning software and all the necessary materials: Dépôt.

Backup slides

Supervised learning

Classification

- b , c, uds jet
- π, K, μ particle
- $t\bar{t}$ or QCD event
- select or reject trigger candidate
- ...

Regression

- energy resolution
- pile-up mitigation
- ...

Don't forget about **unsupervised learning**, **reinforcement learning**, **semi-supervised learning** and so on.

Backup slides

features in ML: not HEP

Supervised learning

Classification

- cat, dog or muffin
- relevant or spam
- disease or not
- good or bad
- ...

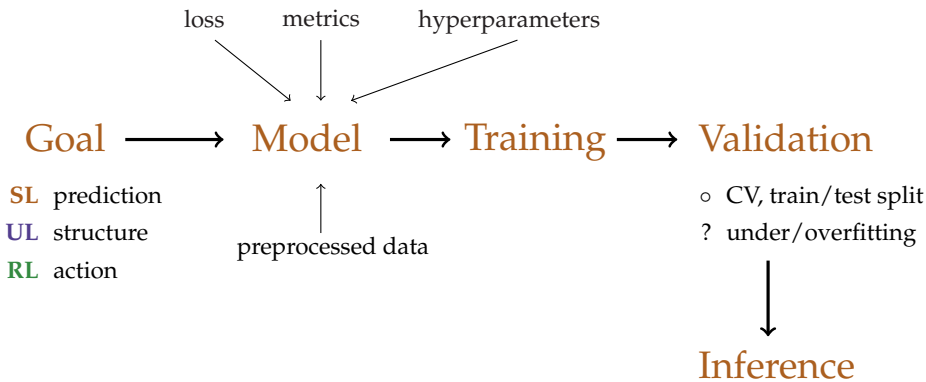
Regression

- rent price
- temperature
- annual profit
- driving time
- ...

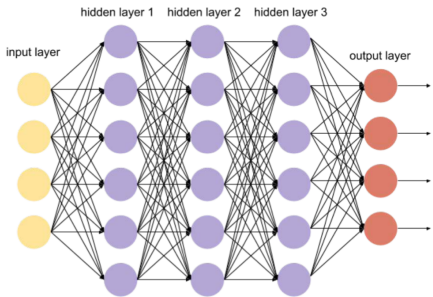
Don't forget about **unsupervised learning**, **reinforcement learning**, **semi-supervised learning** and so on.

Backup slides

what is ML pipeline?



Neural network

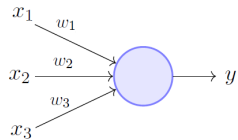


$$h_t = f(W h_{t-1} + b)$$

$$a(x) = f_3(W_3 f_2(W_2 f_1(W_1 f_0(W_0 x + b_0) + b_1) + b_2) + b_3)$$

Backup slides

calculating weights: slide 1



Perceptron Model (Minsky-Papert in 1969)

Source

$$Q\big(y_{model}, y_{target}\big) = Q\big(y(x, w, b), y_{target}\big)$$

$$out = y = \sigma\big(sum\big) = \sigma\Big(\sum_{i=1}^{3}(x_i \cdot w_i) + b\Big)$$

$$\frac{\partial Q}{\partial w_i} = \frac{\partial Q}{\partial y} \cdot \frac{\partial y}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial sum} \cdot \frac{\partial sum}{\partial w_i} = \frac{\partial Q}{\partial y} \cdot 1 \cdot \frac{\partial \sigma}{\partial sum} \cdot \frac{\partial sum}{\partial w_i}$$

$$\frac{\partial \sigma}{\partial sum} = \sigma(sum) \cdot \big(1 - \sigma(sum)\big)$$

Backup slides

calculating weights: slide 2

$$Q\big(y(x,w,b), y_{target}\big)$$
$$Q = \big(y_{target} - y(x_1, w_1, b)\big)^2$$
$$out = y = (x \cdot w) + b$$
$$\frac{\partial Q}{\partial w_i} = \frac{\partial Q}{\partial y} \cdot \frac{\partial y}{\partial w_i}$$

Input: x = 1, w = 0.1 and b = 1

$y_{target} = 2$ and learning rate: $\eta = 0.1$

Backup slides

calculating weights: slide 2

$$Q\big(y(x,w,b),y_{target}\big)$$
$$Q=\big(y_{target}-y(x_1,w_1,b)\big)^2$$
$$out=y=(x\cdot w)+b$$
$$\frac{\partial Q}{\partial w_i}=\frac{\partial Q}{\partial y}\cdot\frac{\partial y}{\partial w_i}$$

Input: x = 1, w = 0.1 and b = 1

$y_{target}=2$ and learning rate: $\eta=0.1$

How to find w_{new} and b_{new}?

$$\frac{\partial Q}{\partial y}=-2\cdot(y_{target}-y)$$
$$\frac{\partial Q}{\partial w}=\frac{\partial Q}{\partial y}\cdot\frac{\partial y}{\partial w}=-2\cdot x\cdot(y_{target}-y)$$
$$\frac{\partial Q}{\partial b}=\frac{\partial Q}{\partial y}\cdot\frac{\partial y}{\partial b}=-2\cdot(y_{target}-y)$$
$$[w_{new},b_{new}]\Rightarrow\theta_{i+1}=\theta_i-\eta\cdot\frac{\partial Q(\theta_i)}{\partial\theta_i}$$

Backup slides

calculating weights: slide 2

$$Q\big(y(x,w,b), y_{target}\big)$$
$$Q = \big(y_{target} - y(x_1, w_1, b)\big)^2$$
$$out = y = (x \cdot w) + b$$
$$\frac{\partial Q}{\partial w_i} = \frac{\partial Q}{\partial y} \cdot \frac{\partial y}{\partial w_i}$$

Input: x = 1, w = 0.1 and b = 1

$y_{target} = 2$ and learning rate: $\eta = 0.1$

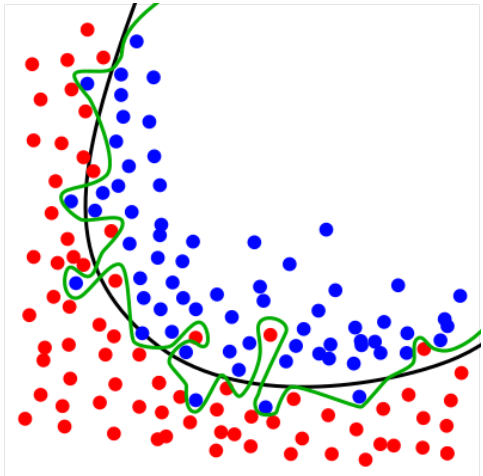
How to find w_{new} and b_{new}?

$$\frac{\partial Q}{\partial y} = -2 \cdot (y_{target} - y)$$
$$\frac{\partial Q}{\partial w} = \frac{\partial Q}{\partial y} \cdot \frac{\partial y}{\partial w} = -2 \cdot x \cdot (y_{target} - y)$$
$$\frac{\partial Q}{\partial b} = \frac{\partial Q}{\partial y} \cdot \frac{\partial y}{\partial b} = -2 \cdot (y_{target} - y)$$
$$[w_{new}, b_{new}] \Rightarrow \theta_{i+1} = \theta_i - \eta \cdot \frac{\partial Q(\theta_i)}{\partial \theta_i}$$

y	Q	$\frac{\partial Q}{\partial y}$	$\frac{\partial Q}{\partial w}$	$\frac{\partial Q}{\partial b}$	w_{new}	b_{new}
1.1	0.81	-1.8	-1.8	-1.8	0.28	1.18

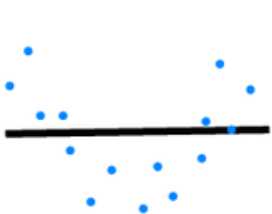
Backup slides

overtraining: slide 1



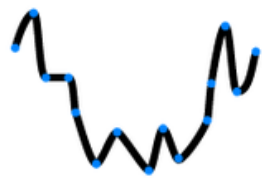
Backup slides

overtraining: slide 2



Underfitting

Desired

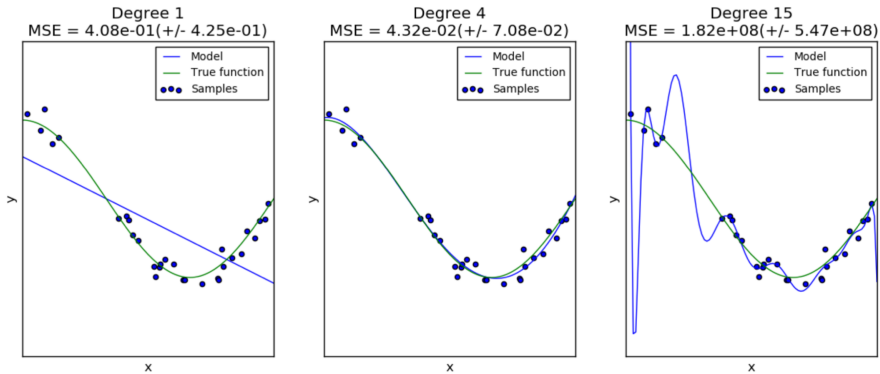


Overfitting

Backup slides

overtraining: slide 3

Overfitting



Backup slides

overtraining: slide 4

История про танки

Классификатор: есть танки на снимке или нет

Backup slides

ML challenges: slide 1 (A. Ustyuzhanin)

- Precise and fast particle tracking (single tracks, shower, jets)
- Particle identification
- Fast and accurate online data processing and filtering
- Anomaly detection (data quality monitoring, infrastructure monitoring)
- Detector design optimization (bayesian optimization, surrogate modelling)
- Data analysis (signal from background separation, ...)
- Simulation (speed-up simulation using generative models, simulator parameters optimization - tuning)
- ...

Backup slides

ML challenges: slide 2 (A. Ustyuzhanin)

Tracking system features

- Particle momentum
- Particle charge
- Track parameters
- Quality of track fit
- Number of track hits
- ...

RICH features

- Angle θ
- Quality of angle reconstruction
- Reconstructed particle type
- Reconstructed particle energy
- Light intensity
- ...

Calorimeter features

- Measured particle energy
- Shower parameters: length, width, ...
- Number of clusters in each layer Intensity of the clusters
- Intensity of the clusters
- Distance form track of the original particle
- ...

Muon detector features

- Muon track parameters
- Quality of track fit
- Number of active layers
- Distance between the track and the active layers
- Length of shower
- ...