Entreposage et fouille de données (STA211)
Neural Networks and Deep Learning

Nicolas Thome
Conservatoire Nationnal des Arts et Métiers (CNAM)
Laboratoire CEDRIC - équipe MSDMA

Outline

1 Context

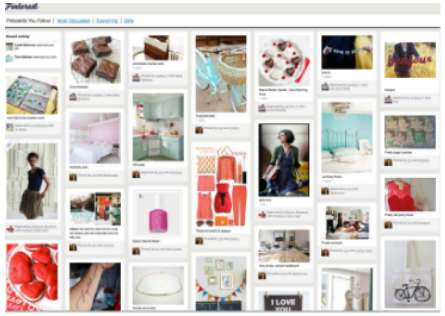
2 Neural Nets

Context

Big Data

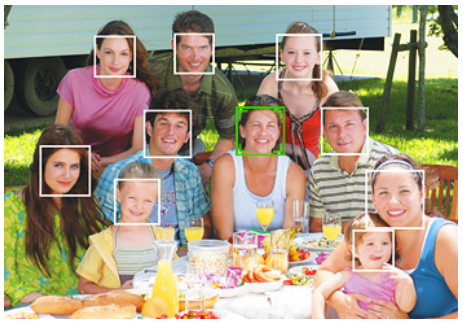
- Superabundance of data: images, videos, audio, text, use traces, *etc*

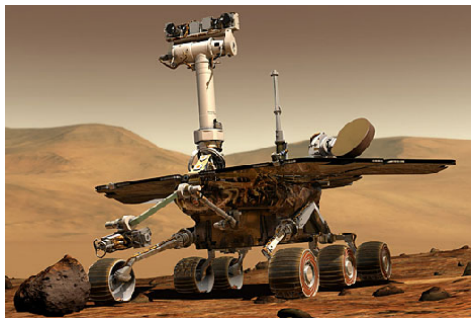
BBC: 2.4M videos


Facebook: 350B images
1B each day


100M monitoring cameras

- Obvious need to access, search, or classify these data: **Recognition**
- Huge number of applications: mobile visual search, robotics, autonomous driving, augmented reality, medical imaging *etc*
- Leading track in major ML/CV conferences during the last decade





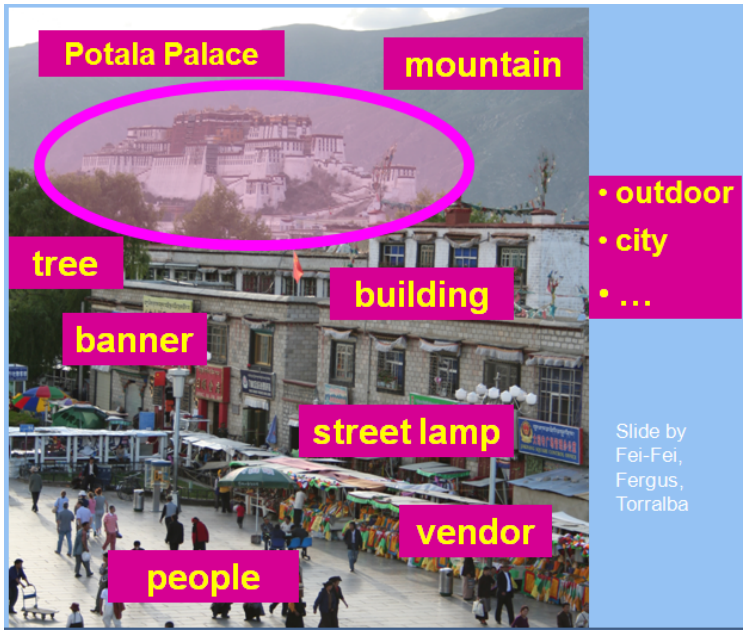
Recognition and classification

- Classification : assign a given data to a given set of pre-defined classes
- Recognition much more general than classification, *e.g.*
  - Ranking for document indexing
  - Localization, segmentation for image understanding
  - Sequence prediction for text, speech, audio, *etc*
- Many tasks can be cast as classification problems
  $\Rightarrow$ **importance of classification**

Focus on Visual Recognition: Perceiving Visual World

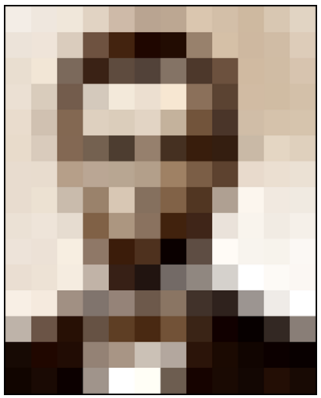
- Visual Recognition: archetype of low-level signal understanding
- Supposed to be a master class problem in the early 80's
- Certainly the most impacted topic by deep learning

- Scene categorization
- Object localization
- Context & Attribute recognition
- Rough 3D layout, depth ordering
- Rich description of scene, *e.g.* sentences



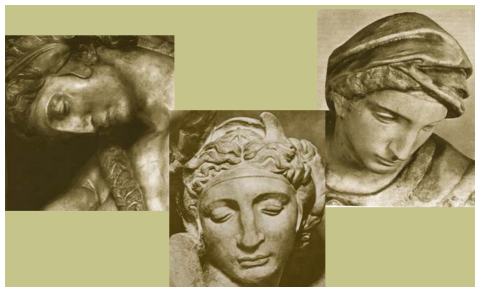
Recognition of low-level signals

Challenge: filling the semantic gap



What we perceive *vs*
What a computer sees

243	239	240	225	206	185	188	218	211	206	216	225
242	239	218	110	67	31	34	152	213	206	208	221
243	242	123	58	94	82	132	77	108	208	208	215
235	217	115	212	243	236	247	139	91	209	208	211
233	208	131	222	219	226	196	114	74	208	213	214
232	217	131	116	77	150	69	56	52	201	228	223
232	232	182	186	184	179	159	123	93	232	235	235
232	236	201	154	216	133	129	81	175	252	241	240
235	238	230	128	172	138	65	63	234	249	241	245
237	236	247	143	59	78	10	94	255	248	247	251
234	237	245	193	55	33	115	144	213	255	253	251
248	245	161	128	149	109	138	65	47	156	239	255
190	107	39	102	94	73	114	58	17	7	51	137
23	32	33	148	168	203	179	43	27	17	12	8
17	26	12	160	255	255	109	22	26	19	35	24

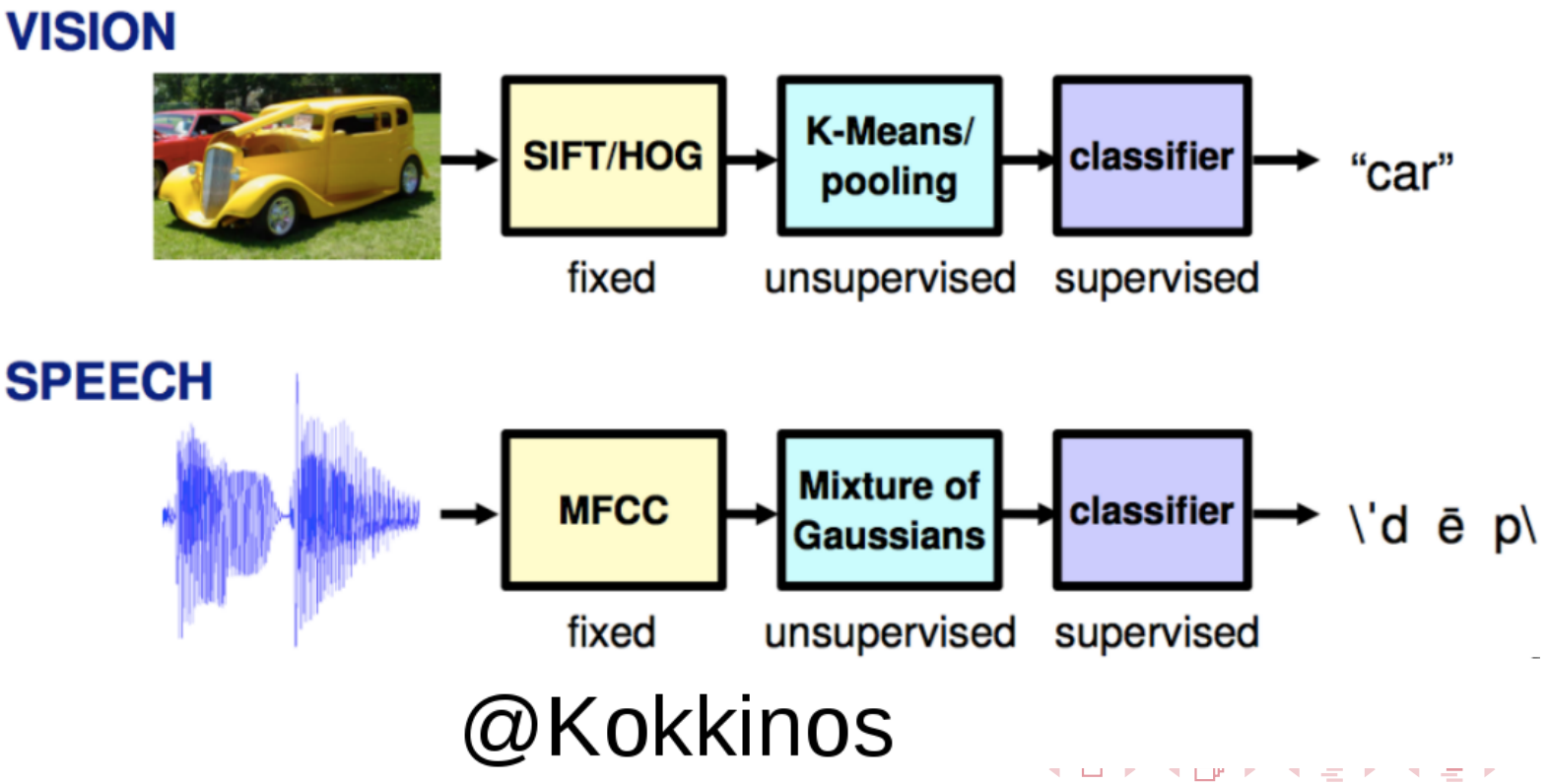


- Illumination variations
- View-point variations
- Deformable objects
- intra-class variance
- *etc*

⇒ **How to design "good" intermediate representation ?**

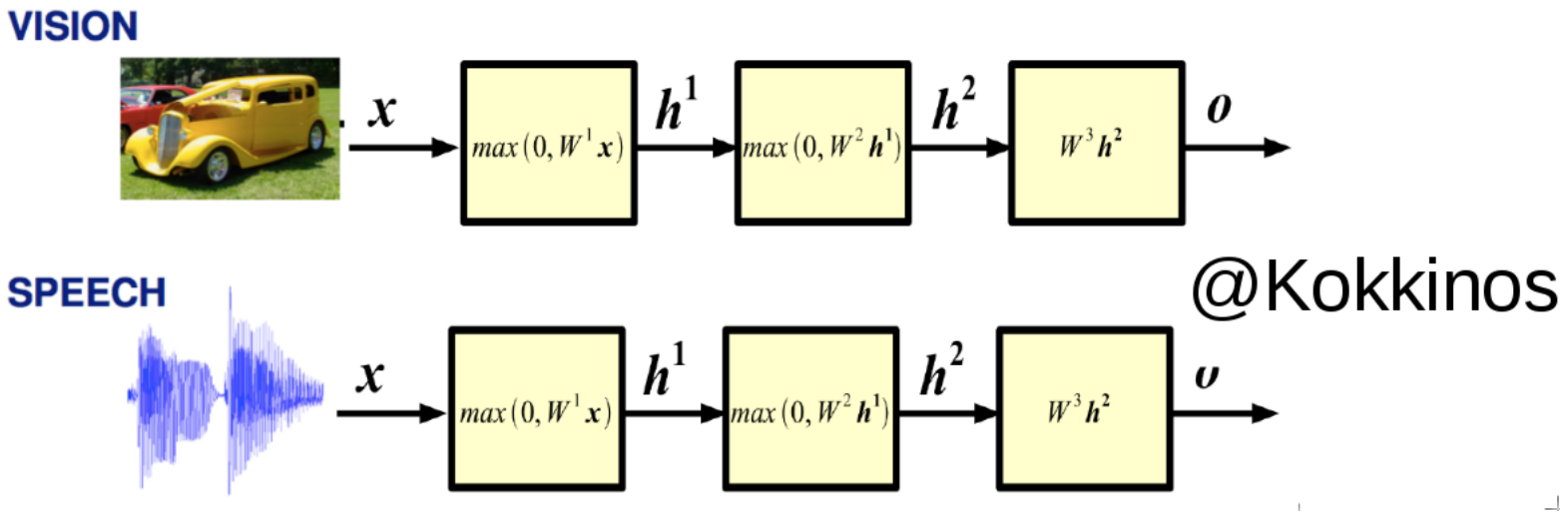
Deep Learning (DL) & Recognition of low-level signals

- DL: breakthrough for the recognition of low-level signal data
- Before DL: handcrafted intermediate representations for each task
  - ⊖ Needs expertise (PhD level) in each field
  - ⊖ Weak level of semantics in the representation



Deep Learning (DL) & Recognition of low-level signals

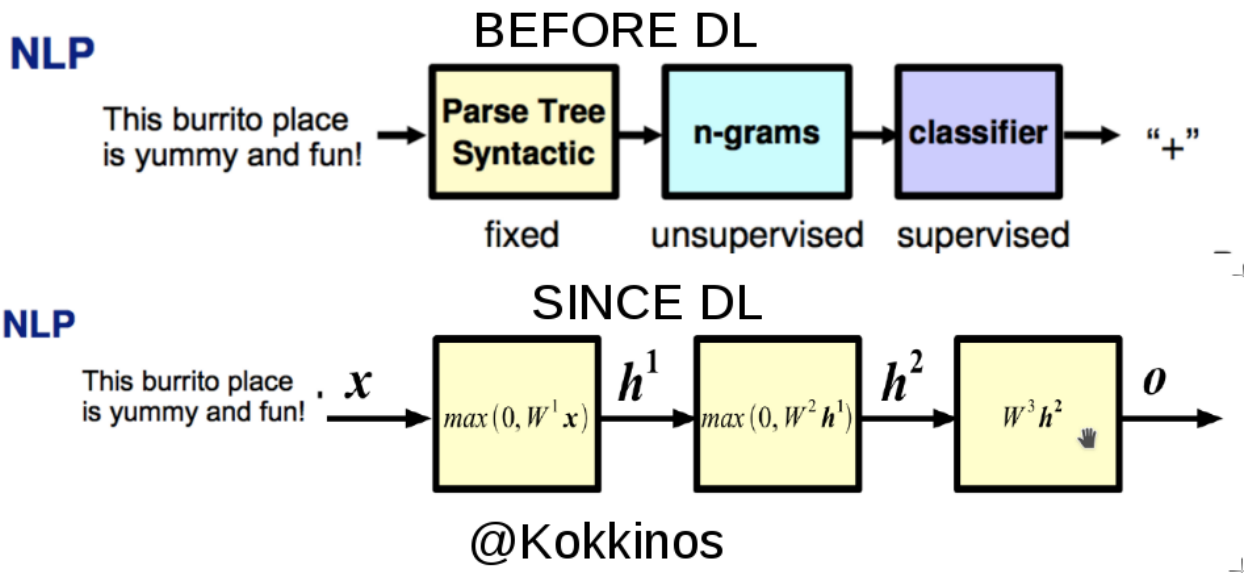
- DL: breakthrough for the recognition of low-level signal data
- Since DL: automatically **learning intermediate representations**
  - ⊕ Outstanding experimental performances >> handcrafted features
  - ⊕ Able to learn high level intermediate representations
  - ⊕ Common learning methodology ⇒ field independent, no expertise

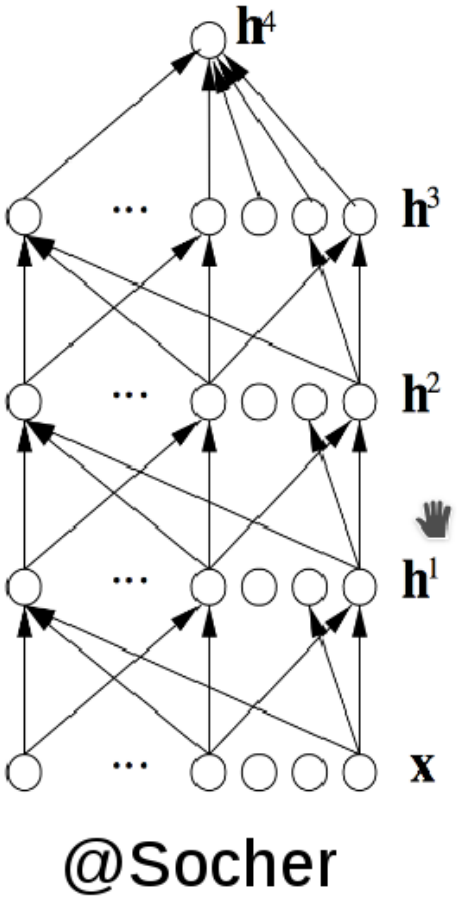


Deep Learning (DL) & Representation Learning

- DL: breakthrough for representation learning
  - Automatically learning intermediate levels of representation

- Ex: Natural language Processing (NLP)





Outline

1 Context

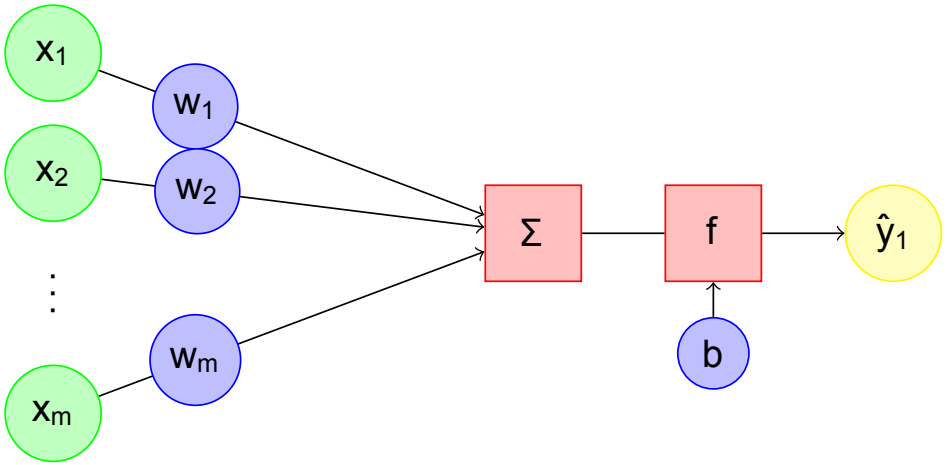
2 Neural Nets

The origins

History

- **1943:** The formal neuron [MP43]
- **1958:** First perceptron [Ros58]
- **1974:** Backpropagation algorithm [Wer74]
- **1980:** First deep feedforward network [Fuk80]
- **1989:** First convolutional neural network: LeNet-5 [LBD+89]

The formal neuron, basis of the neural networks



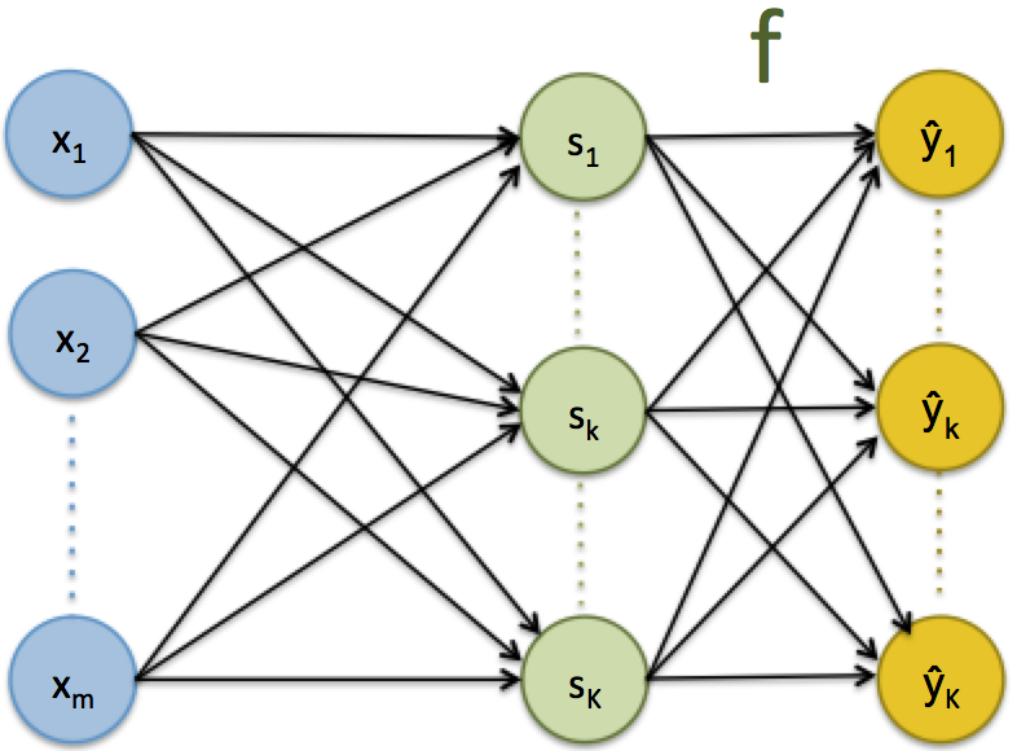
x_i: inputs
w_i, b: weights
f: activation function
y: output of the neuron

$$y = f(w^\top x + b)$$

Figure: The formal neuron – Credits: R. Herault

Neural Networks

- Stacking several formal neurons ⇒ **Perceptron**



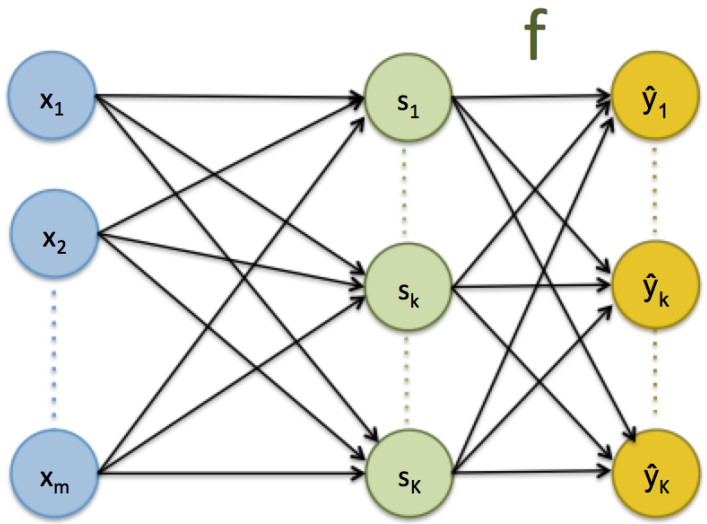
Perceptron and Multi-Class Classification

- **Soft-max Activation**:

$$\hat{y}_k = f(s_k) = \frac{e^{s_k}}{\sum_{k'=1}^{K} e^{s_{k'}}}$$

- **Probabilistic interpretation for multi-class classification**:
  - Each output neuron $\Leftrightarrow$ class
  - $\hat{y}_k \sim P(k/\mathbf{x}, \mathbf{w})$

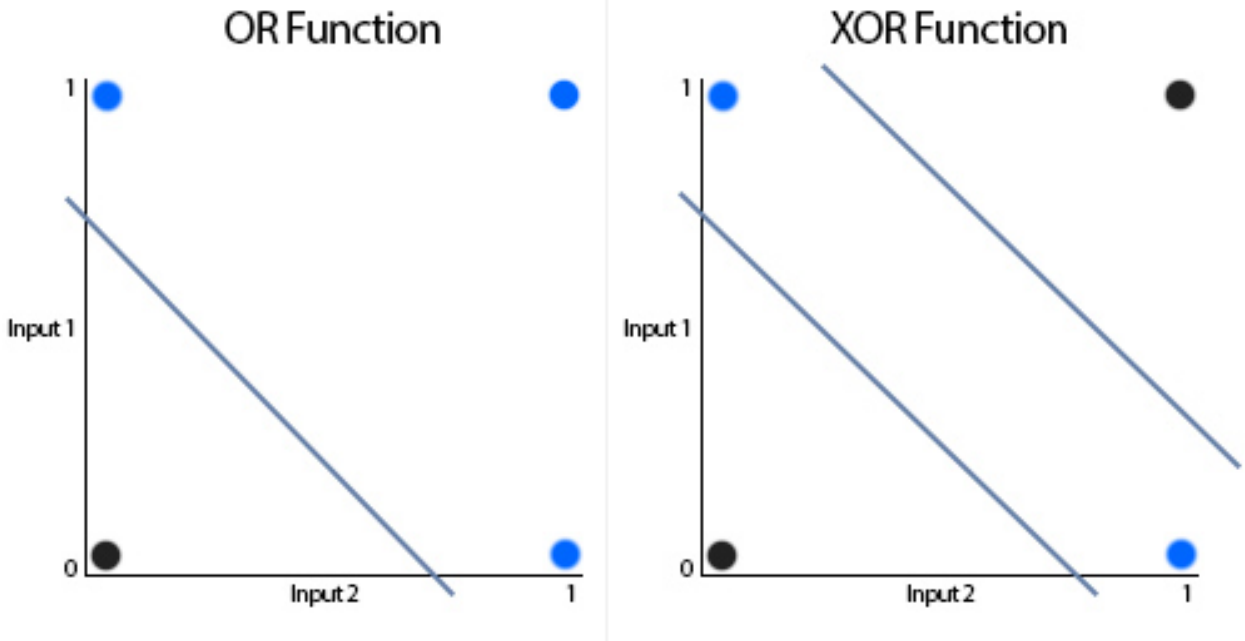
$\Rightarrow$ **Logistic Regression (LR) Model !**



Beyond Linear Prediction

X-OR Problem

- Logistic Regression (LR): NN with 1 input layer & 1 output layer
- LR: limited to linear decision boundaries
- **X-OR:** NOT 1 and 2 OR NOT 2 AND 1
  - **X-OR: Non linear decision function**



The Multi-Layer Perceptron (MLP)

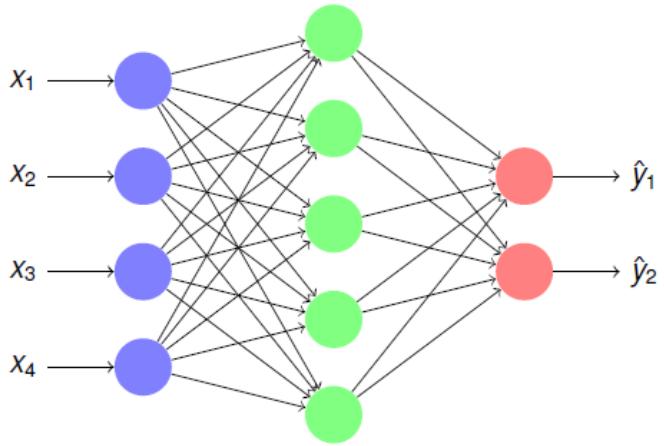


Figure: Perceptron with 1 hidden layer – Credits: R. Herault

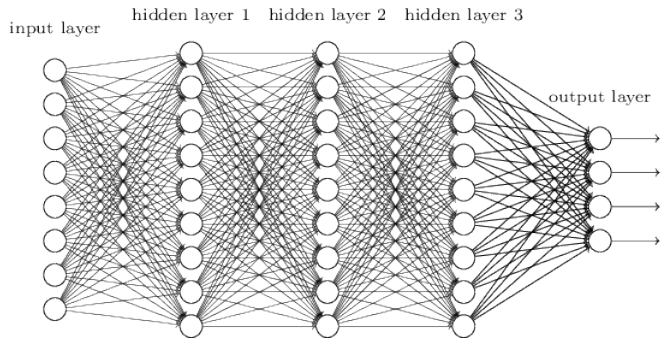
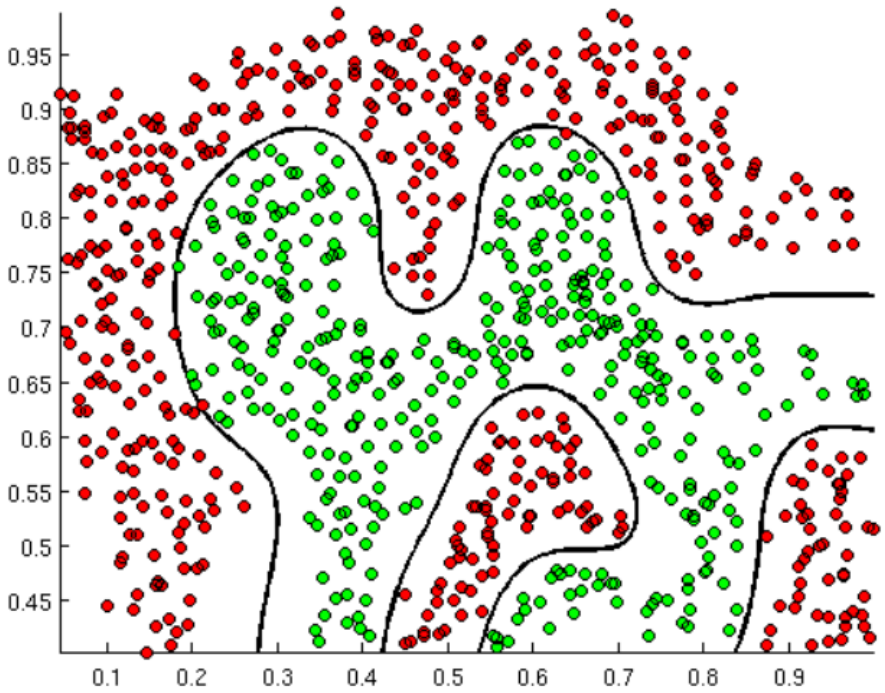


Figure: Stacking more layers, toward "deep learning" – Credits: M. Nielsen

- Basis of the "deep learning" field
- Principle: Stacking layers of neural networks to allow more complex and rich functions
- With a hidden layer, **can approximate any function** given enough hidden units [Cyb89]
- Can be seen as **different levels of abstraction** from low-level features to the high-level ones

The Multi-Layer Perceptron (MLP)

- **Neural network with one single hidden layer ⇒ universal approximator [Cyb89]**
  - Ex for classification: any decision boundaries can be expressed

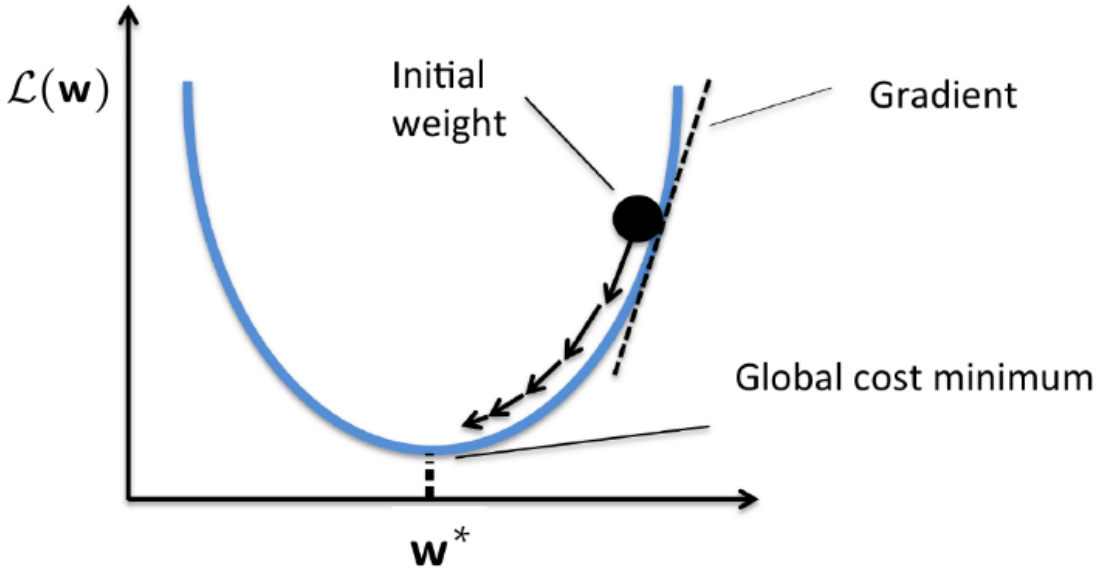


Training Multi-Layer Perceptron (MLP)

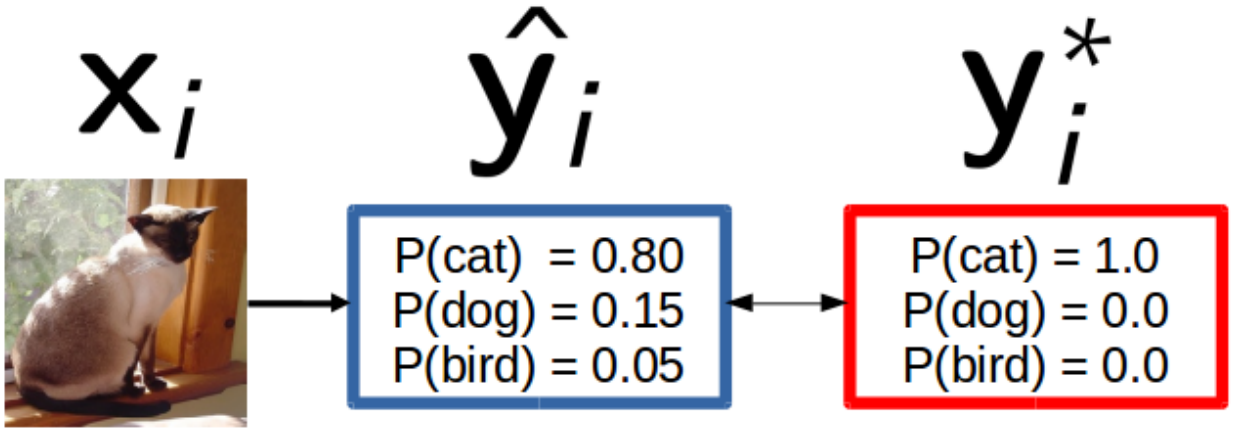
- Input $\mathbf{x}$, output $\mathbf{y}$
- A parametrized model $\mathbf{x} \Rightarrow \mathbf{y}$: $f_{\mathbf{w}}(\mathbf{x}_i) = \hat{\mathbf{y}}_i$
- Supervised context: training set $\mathcal{A} = \{(\mathbf{x}_i, \mathbf{y}_i^*)\}_{i \in \{1,2,\ldots,N\}}$
  - A loss function $\ell(\hat{\mathbf{y}}_i, \mathbf{y}_i^*)$ for each annotated pair $(\mathbf{x}_i, \mathbf{y}_i^*)$
- Assumptions: parameters $\mathbf{w} \in \mathbb{R}^d$ continuous, $\mathcal{L}$ differentiable
- Gradient $\nabla_{\mathbf{w}} = \frac{\partial \mathcal{L}}{\partial \mathbf{w}}$: steepest direction to decrease loss $\mathcal{L}$

MLP Training

- Gradient descent algorithm:
  - Initialyze parameters $\mathbf{w}$
  - Update: $\boxed{\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{w}}}$
  - Until convergence, *e.g.* $\|\nabla_{\mathbf{w}}\|^2 \approx 0$



MLP Training: loss function



- **Input $\mathbf{x}_i$, ground truth output supervision $\mathbf{y}_i^*$**
- One hot-encoding for $\mathbf{y}_i^*$:

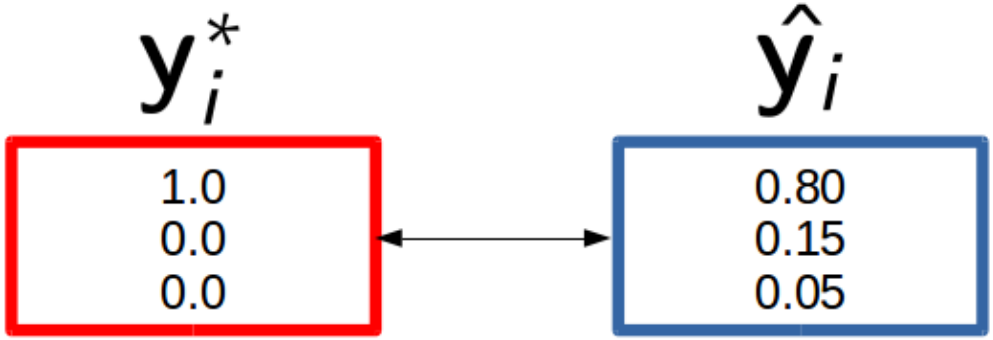
$$y_{c,i}^* = \begin{cases} 1 & \text{if c is the groud truth class for } \mathbf{x}_i \\ 0 & \text{otherwise} \end{cases}$$

MLP Training

- Loss function: multi-class Cross-Entropy (CE) ℓ_{CE}
- ℓ_{CE}: Kullback-Leiber divergence between $\mathbf{y}_i^*$ and $\hat{\mathbf{y}}_i$

$$\ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*) = KL(\mathbf{y}_i^*, \hat{\mathbf{y}}_i) = -\sum_{c=1}^{K} y_{c,i}^* log(\hat{y}_{c,i}) = -log(\hat{y}_{c^*,i})$$

- KL asymmetric: $KL(\hat{\mathbf{y}}_i, \mathbf{y}_i^*) \neq KL(\mathbf{y}_i^*, \hat{\mathbf{y}}_i)$

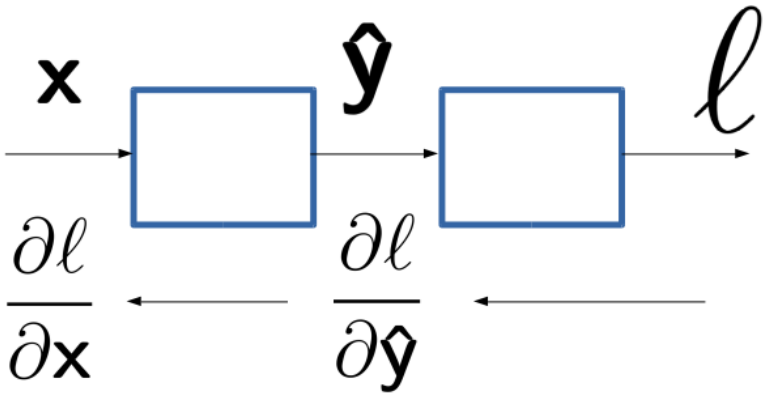


$$KL(\mathbf{y}_i^*, \hat{\mathbf{y}}_i) = -log(\hat{y}_{c^*,i}) = -log(0.8) \approx 0.22$$

MLP Training: Backpropagation

- $\mathcal{L}_{CE}(\mathbf{W}, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^{N} \ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*) = -\frac{1}{N} \sum_{i=1}^{N} log(\hat{y}_{c^*,i})$
- ℓ_{CE} smooth convex upper bound of $\ell_{0/1}$
  $\Rightarrow$ **gradient descent optimization**
- Gradient descent: $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \frac{\partial \mathcal{L}_{CE}}{\partial \mathbf{W}}$
  $(\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - \eta \frac{\partial \mathcal{L}_{CE}}{\partial \mathbf{b}})$
- Computing $\frac{\partial \mathcal{L}_{CE}}{\partial \mathbf{W}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_{CE}}{\partial \mathbf{W}}$?
  $\Rightarrow$ **Backpropagation of gradient error!**
  $\Rightarrow$ **Key Property:** chain rule $\frac{\partial \mathbf{x}}{\partial z} = \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial z}$

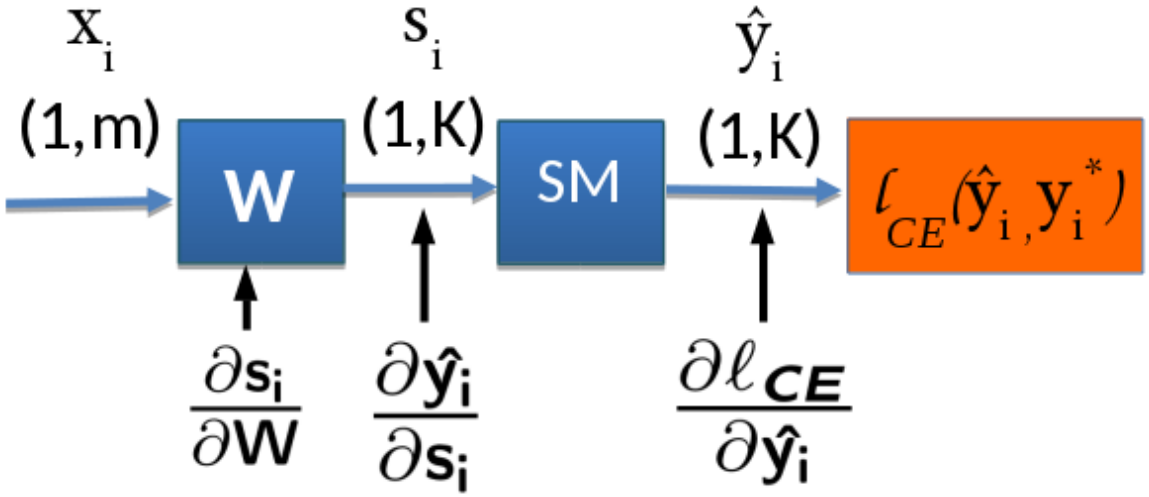
Chain Rule



$$\frac{\partial \ell}{\partial x} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial x}$$

- Logistic regression:

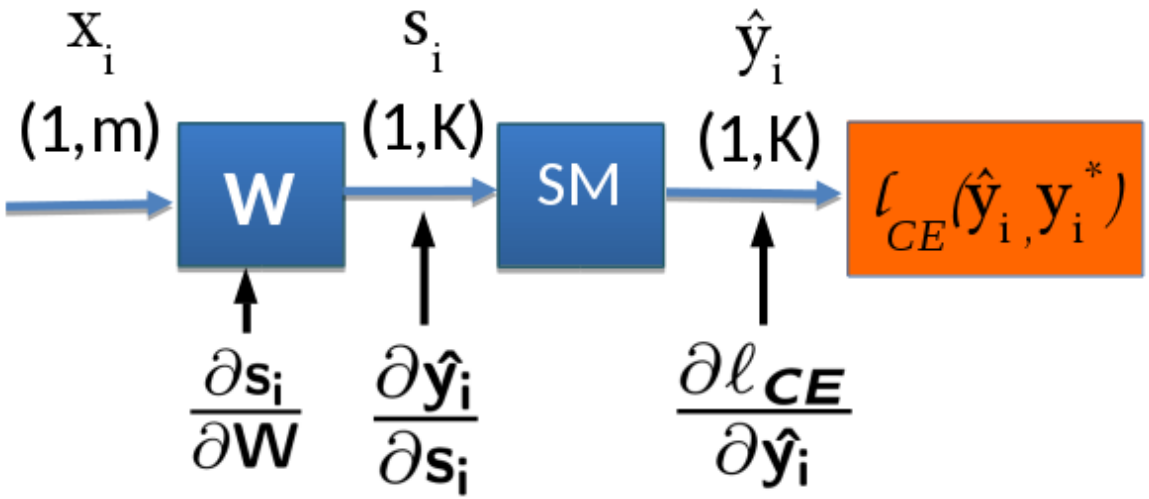
$$\frac{\partial \ell_{CE}}{\partial W} = \frac{\partial \ell_{CE}}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial s_i} \frac{\partial s_i}{\partial W}$$



Logistic Regression Training: Backpropagation

$\frac{\partial \ell_{CE}}{\partial \mathbf{W}} = \frac{\partial \ell_{CE}}{\partial \hat{\mathbf{y}}_i} \frac{\partial \hat{\mathbf{y}}_i}{\partial \mathbf{s}_i} \frac{\partial \mathbf{s}_i}{\partial \mathbf{W}}$, $\ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*) = -log(\hat{y}_{c^*,i}) \Rightarrow$ Update for 1 example:

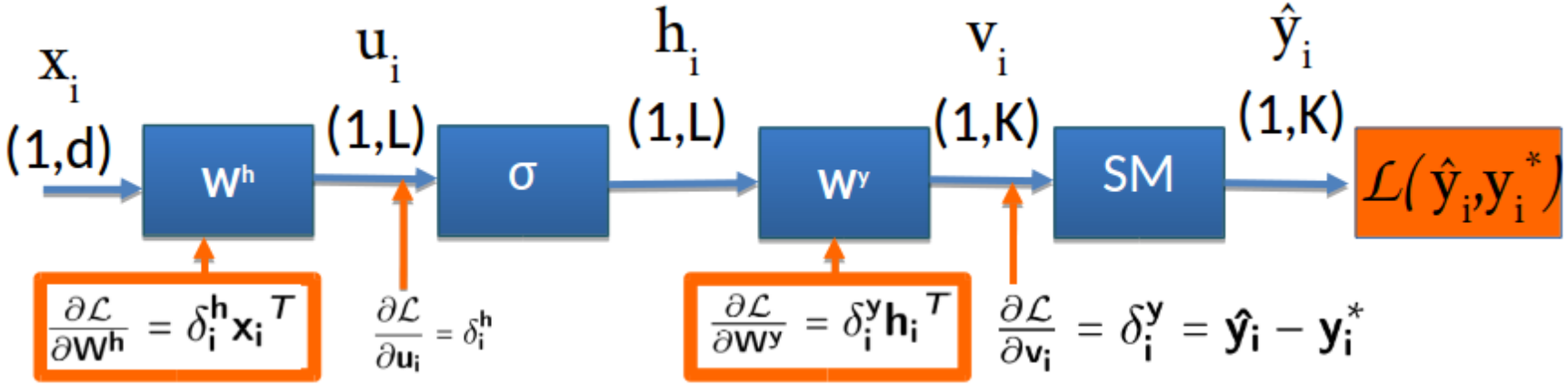
- $\frac{\partial \ell_{CE}}{\partial \hat{\mathbf{y}}_i} = \frac{-1}{\hat{y}_{c^*,i}} = \frac{-1}{\hat{\mathbf{y}}_i} \odot \delta_{\mathbf{c},\mathbf{c}^*}$
- $\frac{\partial \ell_{CE}}{\partial \mathbf{s}_i} = \hat{\mathbf{y}}_i - \mathbf{y}_i^* = \delta_i^{\mathbf{y}}$
- $\frac{\partial \ell_{CE}}{\partial \mathbf{W}} = \mathbf{x}_i{}^T \delta_i^{\mathbf{y}}$



Perceptron Training: Backpropagation

- Perceptron *vs* Logistic Regression: adding hidden layer (sigmoid)
- **Goal:** Train parameters $\mathbf{W}^y$ and $\mathbf{W}^h$ (+bias) with Backpropagation

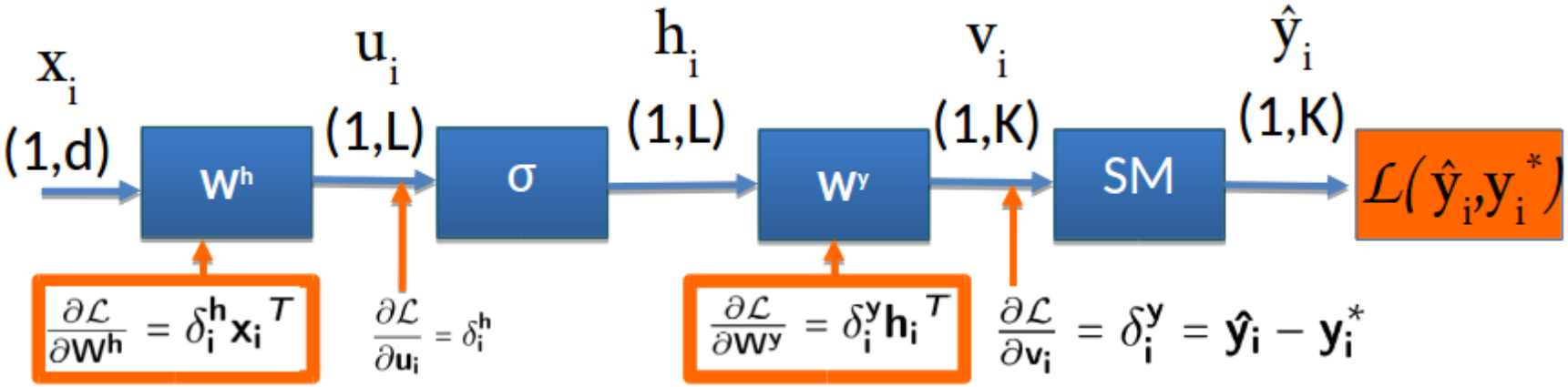
$\Rightarrow$ computing $\boxed{\frac{\partial \mathcal{L}_{CE}}{\partial \mathbf{W}^y} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_{CE}}{\partial \mathbf{W}^y}}$ and $\boxed{\frac{\partial \mathcal{L}_{CE}}{\partial \mathbf{W}^h} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_{CE}}{\partial \mathbf{W}^h}}$



- Last hidden layer ~ Logistic Regression
- First hidden layer: $\frac{\partial \ell_{CE}}{\partial \mathbf{W}^h} = \mathbf{x_i}^T \frac{\partial \ell_{CE}}{\partial \mathbf{u_i}} \Rightarrow$ **computing** $\frac{\partial \ell_{CE}}{\partial \mathbf{u_i}} = \delta_i^h$

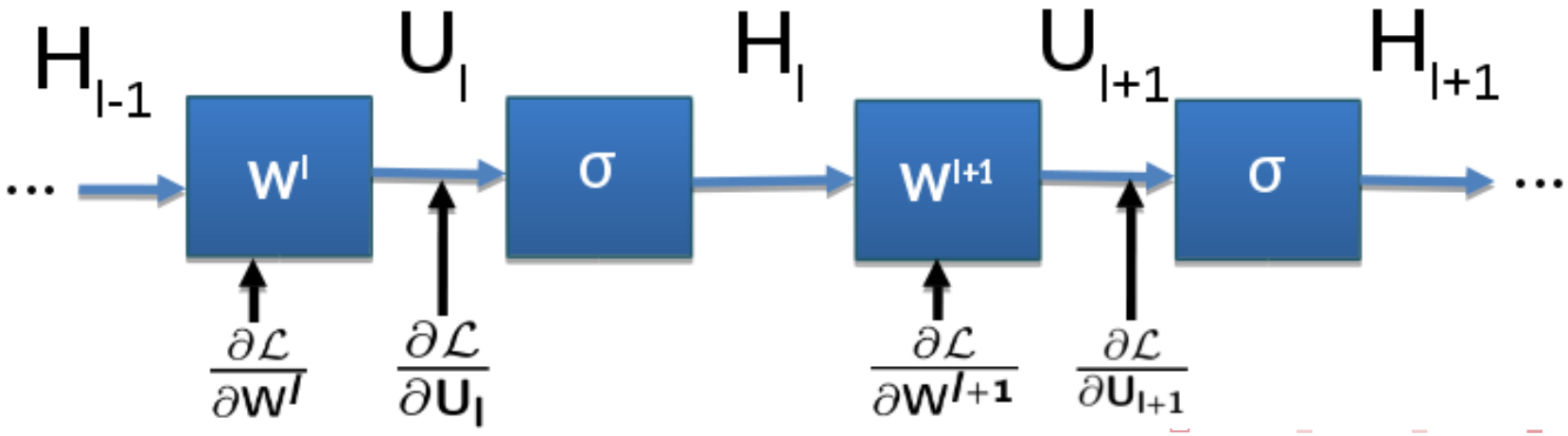
Perceptron Training: Backpropagation

- Computing $\frac{\partial \ell_{CE}}{\partial \mathbf{u_i}} = \delta_{\mathbf{i}}^{\mathbf{h}} \Rightarrow$ use chain rule: $\frac{\partial \ell_{CE}}{\partial \mathbf{u_i}} = \frac{\partial \ell_{CE}}{\partial \mathbf{v_i}} \frac{\partial \mathbf{v_i}}{\partial \mathbf{h_i}} \frac{\partial \mathbf{h_i}}{\partial \mathbf{u_i}}$
- ... Leading to:
  $\frac{\partial \ell_{CE}}{\partial \mathbf{u_i}} = \delta_{\mathbf{i}}^{\mathbf{h}} = \delta_{\mathbf{i}}^{\mathbf{y}\,T} \mathbf{W^y} \odot \sigma'(\mathbf{h_i}) = \delta_{\mathbf{i}}^{\mathbf{y}\,T} \mathbf{W^y} \odot (\mathbf{h_i} \odot (1 - \mathbf{h_i}))$



Deep Neural Network Training: Backpropagation

- Multi-Layer Perceptron (MLP): adding more hidden layers
- Backpropagation update ~ Perceptron: **assuming** $\frac{\partial \mathcal{L}}{\partial \mathbf{U_{l+1}}} = \Delta^{l+1}$ **known**
  - $\boxed{\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{l+1}} = \mathbf{H_l}^T \Delta^{l+1}}$
  - $\boxed{\text{Computing } \frac{\partial \mathcal{L}}{\partial \mathbf{U_l}} = \Delta^l}$ $(= \Delta^{l+1\,T} \mathbf{W^{l+1}} \odot \mathbf{H_l} \odot (1 - \mathbf{H_l})$ sigmoid$)$
  - $\boxed{\frac{\partial \mathcal{L}}{\partial \mathbf{W}^l} = \mathbf{H_{l-1}}^T \Delta^{h_l}}$



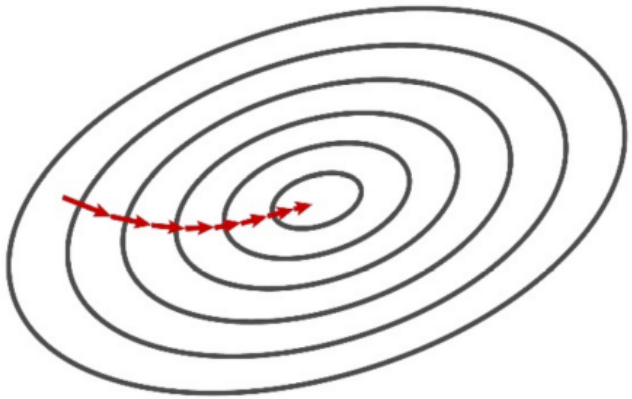
Neural Network Training: Optimization Issues

- Classification loss over training set ($\mathbf{w}$):

$$\mathcal{L}_{CE}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*) = -\frac{1}{N}\sum_{i=1}^{N} log(\hat{y}_{c^*,i})$$

- Gradient descent optimization:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\frac{\partial\mathcal{L}_{CE}}{\partial\mathbf{w}}\left(\mathbf{w}^{(t)}\right) = \mathbf{w}^{(t)} - \eta\nabla,$$



- Gradient $\nabla_{\mathbf{w}}^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\ell_{CE}(\hat{\mathbf{y}}_i,\mathbf{y}_i^*)}{\partial\mathbf{w}}\left(\mathbf{w}^{(t)}\right)$

  linearly scales wrt:
  - $\mathbf{w}$ dimension
  - Training set size

$\Rightarrow$ **Too slow even for moderate dimensionality & dataset size!**

Stochastic Gradient Descent

- **Solution:** approximate $\nabla_{\mathbf{w}}^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*)}{\partial \mathbf{w}} \left(\mathbf{w}^{(t)}\right)$ with subset

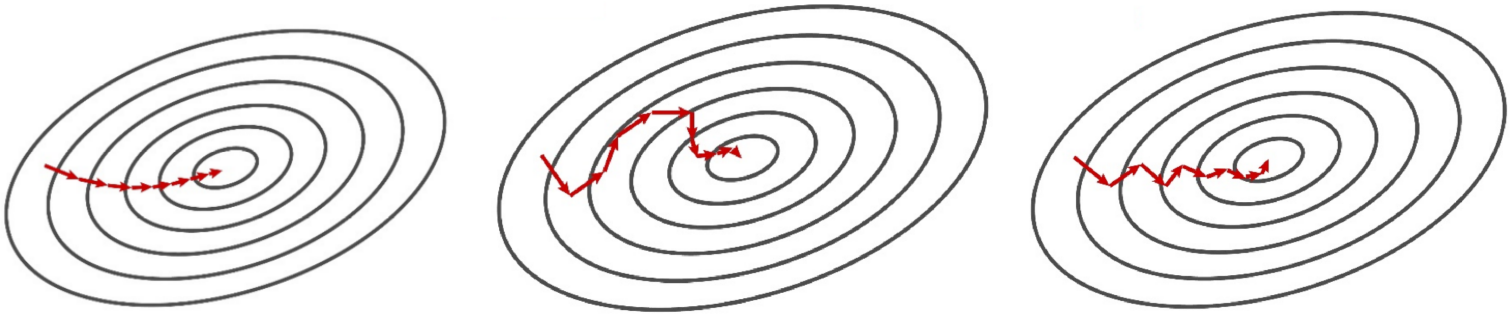
  $\Rightarrow$ **Stochastic Gradient Descent (SGD)**

  - Use a single example (online):

$$\nabla_{\mathbf{w}}^{(t)} \approx \frac{\partial \ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*)}{\partial \mathbf{w}} \left(\mathbf{w}^{(t)}\right)$$

  - Mini-batch: use $B < N$ examples:

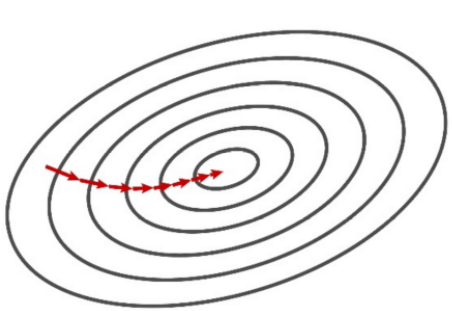
$$\nabla_{\mathbf{w}}^{(t)} \approx \frac{1}{B} \sum_{i=1}^{B} \frac{\partial \ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*)}{\partial \mathbf{w}} \left(\mathbf{w}^{(t)}\right)$$



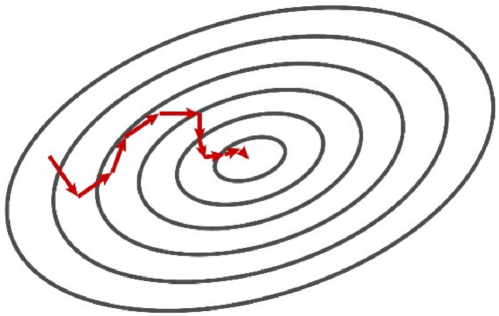
Full gradient | SGD (online) | SGD (mini-batch)

Stochastic Gradient Descent

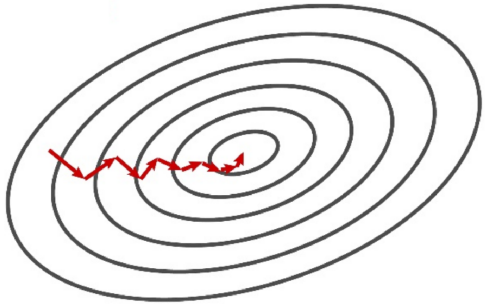
- **SGD: approximation of the true Gradient $\nabla_{\mathbf{w}}$!**
  - Noisy gradient can lead to bad direction, increase loss
  - **BUT:** much more parameter updates: online $\times N$, mini-batch $\times \frac{N}{B}$
  - **Faster convergence**, at the core of Deep Learning for large scale datasets



Full gradient



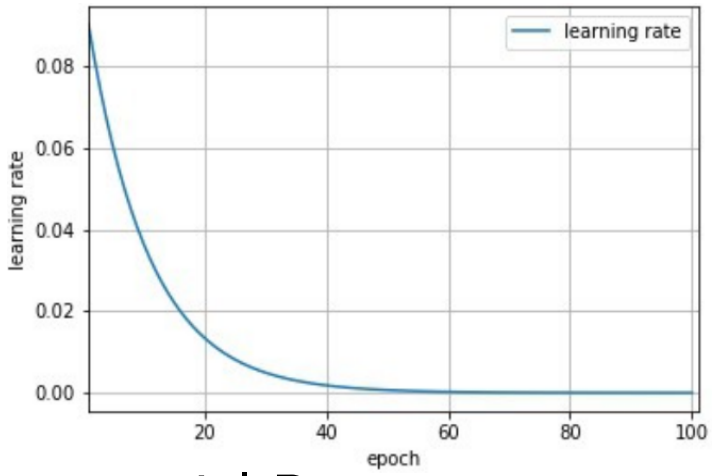
SGD (online)



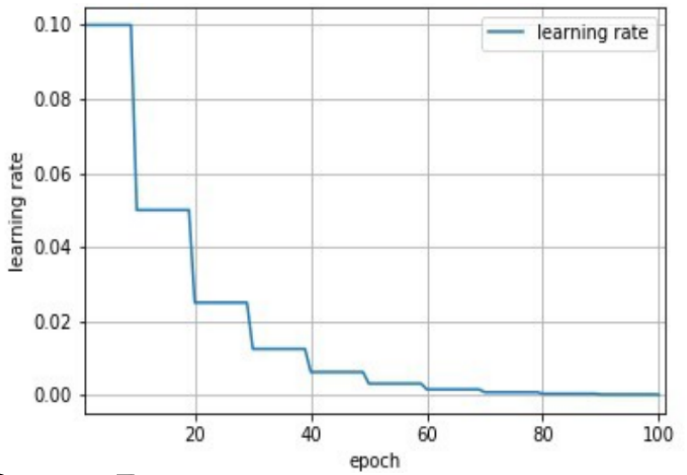
SGD (mini-batch)

Optimization: Learning Rate Decay

- Gradient descent optimization: $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla_{\mathbf{w}}^{(t)}$
- η setup ? $\Rightarrow$ open question
- Learning Rate Decay: decrease η during training progress
  - Inverse (time-based) decay: $\eta_t = \frac{\eta_0}{1+r \cdot t}$, r decay rate
  - Exponential decay: $\eta_t = \eta_0 \cdot e^{-\lambda t}$
  - Step Decay $\eta_t = \eta_0 \cdot r^{\frac{t}{t_u}}$ ...



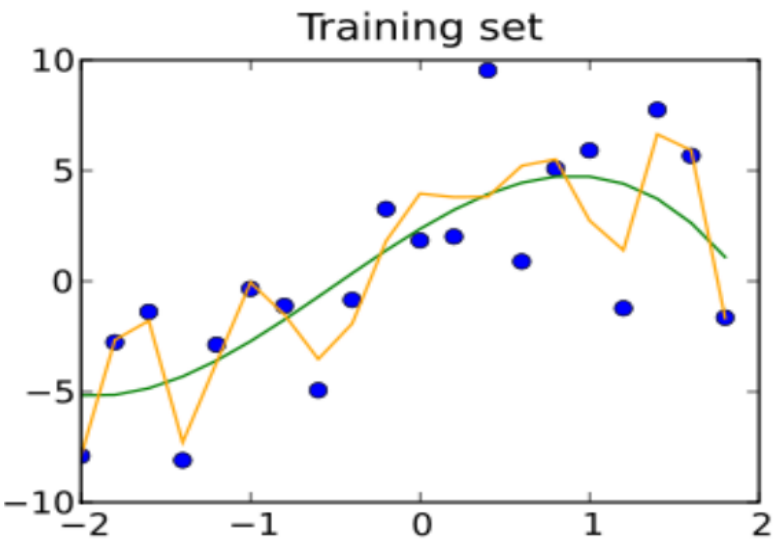
Exponential Decay ($\eta_0 = 0.1$, $\lambda = 0.1$s)

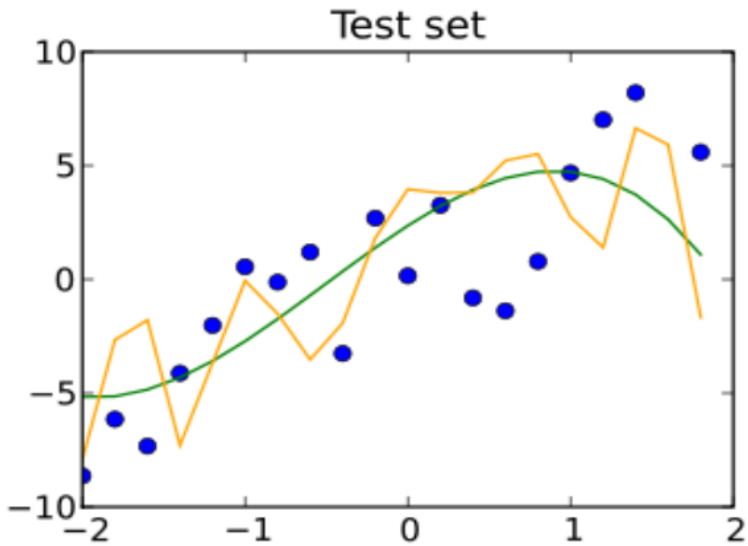


Step Decay ($\eta_0 = 0.1$, $r = 0.5$, $t_u = 10$)

Generalization and Overfitting

- **Learning:** minimizing classification loss $\mathcal{L}_{CE}$ over training set
  - Training set: sample representing data *vs* labels distributions
  - **Ultimate goal:** train a prediction function with low prediction error on the **true (unknown) data distribution**





$\mathcal{L}_{train} = 4$, $\mathcal{L}_{train} = 9$

$\mathcal{L}_{test} = 15$, $\mathcal{L}_{test} = 13$

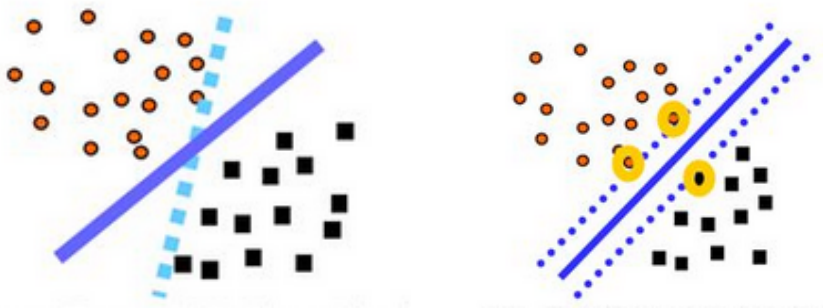
$\Rightarrow$ **Optimization $\neq$ Machine Learning!**
$\Rightarrow$ **Generalization / Overfitting!**

Regularization

- **Regularization:** improving generalization, *i.e.* test ($\neq$ *train*) performances
- Structural regularization: add ***Prior*** $R(\mathbf{w})$ in training objective:

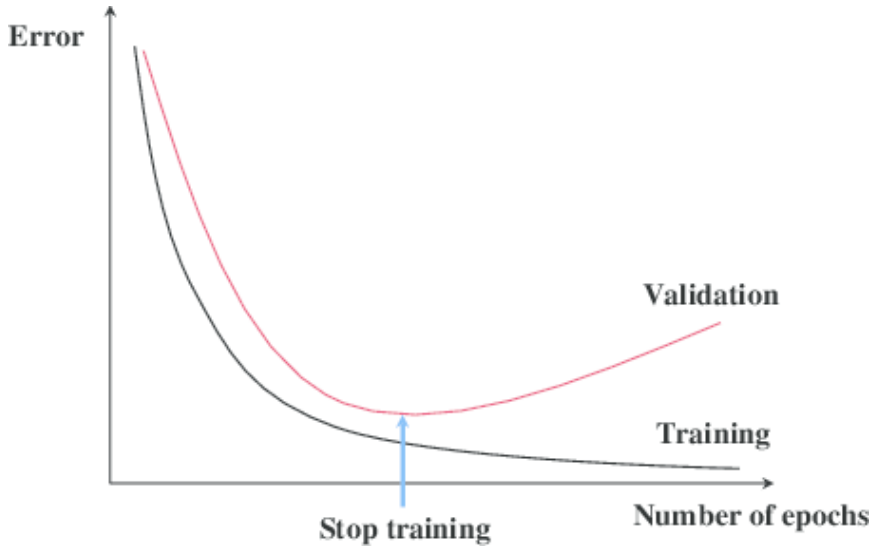
$$\mathcal{L}(\mathbf{w}) = \mathcal{L}_{CE}(\mathbf{w}) + \alpha R(\mathbf{w})$$

- L^2 **regularization: weight decay**, $R(\mathbf{w}) = \|\mathbf{w}\|^2$
  - Commonly used in neural networks
  - Theoretical justifications, generalization bounds (SVM)
- Other possible $R(\mathbf{w})$: L^1 regularization, dropout, *etc*



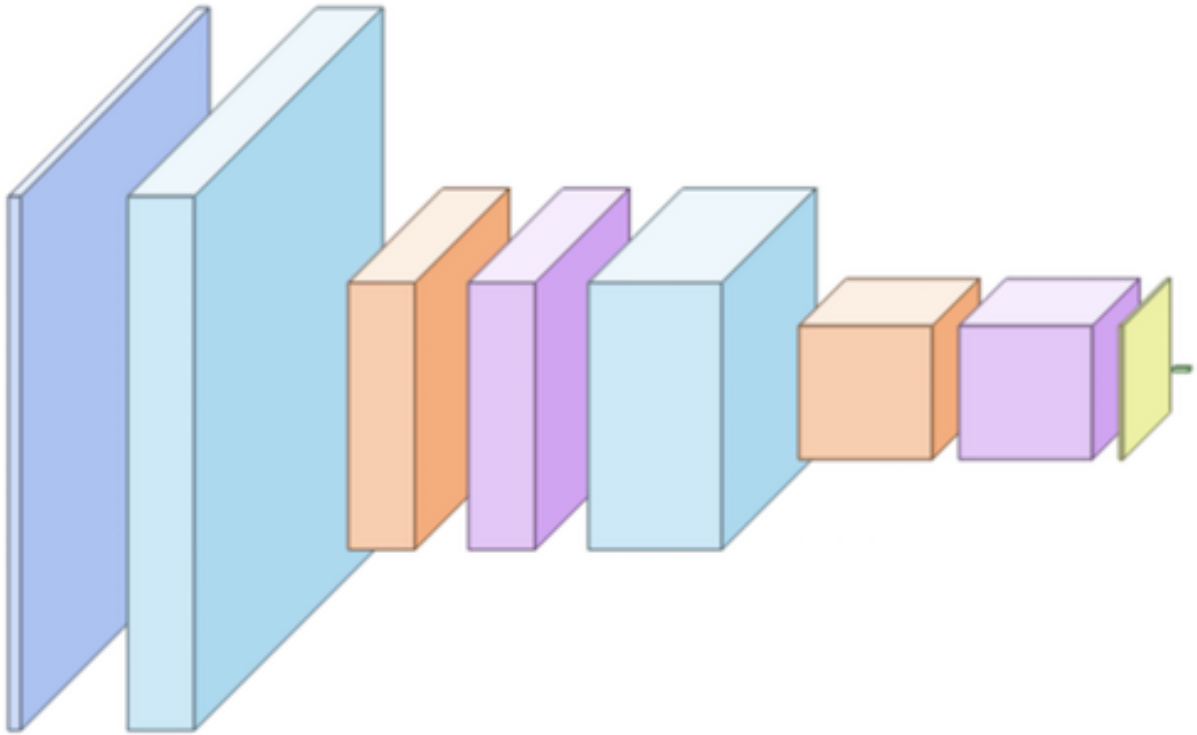
Regularization and hyper-parameters

- **Neural networks:** hyper-parameters to tune:
  - **Training parameters:** learning rate, weight decay, learning rate decay, # epochs, *etc*
  - **Architectural parameters:** number of layers, number neurones, non-linearity type, *etc*
- **Hyper-parameters tuning:** ⇒ improve generalization: estimate performances on a validation set



Neural networks: Conclusion

- Training issues at several levels: optimization, generalization, cross-validation
- **Limits of fully connected layers and Convolutional Neural Nets ? $\Rightarrow$ next course!**



References I


George Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of control, signals and systems **2** (1989), no. 4, 303–314.


Kunihiko Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological cybernetics **36** (1980), no. 4, 193–202.


Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, *Backpropagation applied to handwritten zip code recognition*, Neural computation **1** (1989), no. 4, 541–551.


Warren S McCulloch and Walter Pitts, *A logical calculus of the ideas immanent in nervous activity*, The bulletin of mathematical biophysics **5** (1943), no. 4, 115–133.


F. Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review **65** (1958), no. 6, 386–408.


P. Werbos, *Beyond regression: New tools for prediction and analysis in the behavioral sciences*, Ph.D. thesis, Harvard University, 1974.