Insurance analytics

Neural networks

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

May 27, 2019

Acknowledgement

- Some of the figures in this presentation are taken from
  - Michael A. Nielsen (2015). Neural networks and deep learning. Determination Press.
  - all the beautiful work by prof. Taylor Arnold, in particular Chapter 8 of the book *A Computational Approach to Statistical Learning*.

Today's mission

- Today's mission:
  - de-mystify neural networks
  - sketch different types of neural networks and their applications
  - a discussion of specific considerations to keep in mind when using these predictive modeling techniques with frequency/severity data.

A famous challenge

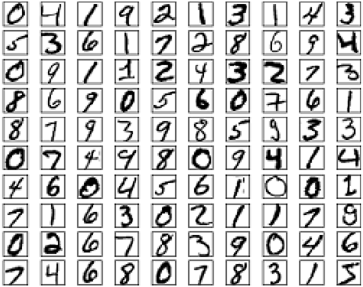
Recognizing handwritten digits

- Consider this

- Humans have a primary visual cortex (V_1) with 140M neurons, with 10s billions connections.
- Next to V_1, also V_2, V_3, V_4 and V_5, doing complex image processing.
- A supercomputer in our head!

A famous challenge

Recognizing handwritten digits

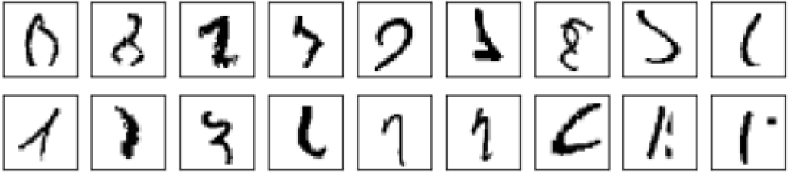


The goal: infer rules for recognizing handwritten digits.

The famous MNIST (Modified National Institute of Standards and Technology) database with handwritten digits commonly used for training various image processing systems.

A famous challenge

Recognizing handwritten digits



The winner: well-designed neural nets classify 9 979 out of 10 000 images correctly (in 2013).

Types of neural networks

What's in a name?

- Different types of neural networks and their applications:
  - ANN: Artificial Neural Network

    for regression and classification problems, with vectors as input data
  - CNN: Convolutional Neural Network

    for image processing, image/face/... recognition, with images as input data
  - RNN: Recurrent Neural Network

    for sequential data such as text or time series.

Types of neural networks

A bit of history

- 1943: McCulloch and Pitts (1943) with a computational model for neural networks based on maths and algorithms. Paved the way for neural network research in artificial intelligence.
- 1958: perceptron algorithm for pattern recognition by Rosenblatt.
- 1969: discovery of two fundamental problems with the computational machines that processed neural networks (e.g. computer power) by Minsky and Papert.
- 1986: backpropagation algorithm published in Nature.
- 1990 - 2010: not much happening.
- 2010 - ... : AI boom, driven by (deep) neural networks. Many factors play a role: power of CNNs and RNNs recognized, more computer power, higher demand due to more and more complex data.

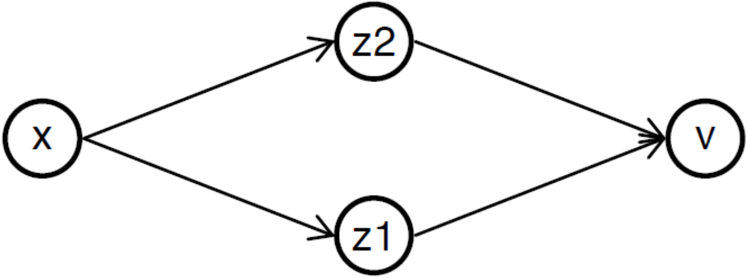
Artificial neural networks

A simple neural network

- De-mystify artificial neural networks (ANNs):
  - a collection of inter-woven linear models
  - extending linear approaches to detect non-linear interactions in high-dimensional data.

Artificial neural networks

A simple neural network



The goal: predict a scalar response y from scalar input x.

Artificial neural networks

A simple neural network

- Some terminology:
  - x is the input layer
  - v is output layer
  - middle layers are hidden layers
  - four neurons: x, z_1, z_2 and v.

Artificial neural networks

A simple neural network

- First, apply two independent linear models:

$$
\begin{aligned}
z_1 &= b_1 + x \cdot w_1 \\
z_2 &= b_2 + x \cdot w_2,
\end{aligned}
$$

using four parameters: two intercepts and two slopes.

- Next, construct another linear model with the z_j as inputs:

$$
\hat{y} := v \quad = \quad b_3 + z_1 \cdot u_1 + z_2 \cdot u_2.
$$

Artificial neural networks

A simple neural network

- Putting it all together:

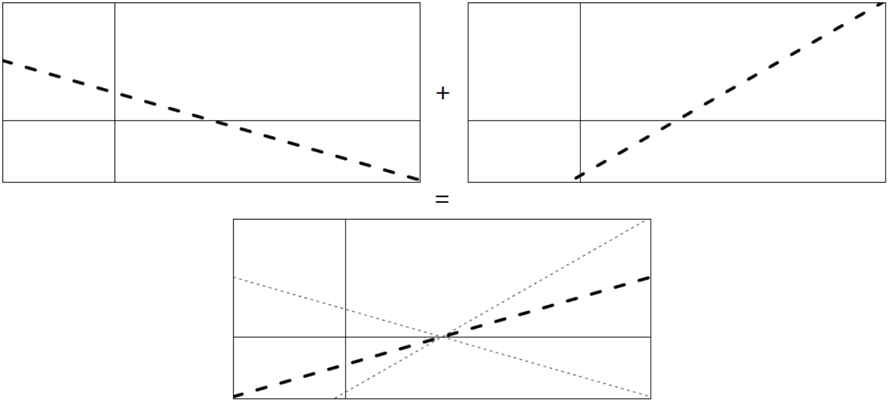
$$
\begin{aligned}
v &= b_3 + z_1 \cdot u_1 + z_2 \cdot u_2 \\
&= b_3 + (b_1 + x \cdot w_1) \cdot u_1 + (b_2 + x \cdot w_2) \cdot u_2 \\
&= (b_3 + u_1 \cdot b_1 + u_2 \cdot b_2) + (w_1 \cdot u_1 + w_2 \cdot u_2) \cdot x \\
&= (\text{intercept}) + (\text{slope}) \cdot x.
\end{aligned}
$$

- Model is over-parametrized, with infinitely many ways to describe same model.

- Essentially, still a linear model!

Artificial neural networks

A simple neural network



Artificial neural networks

Activation function

- Capture non-linear relationships between x and v by replacing

$$v = b_3 + z_1 \cdot u_1 + z_2 \cdot u_2.$$

with

$$\begin{aligned} v &= b_3 + \sigma(z_1) \cdot u_1 + \sigma(z_2) \cdot u_2 \\ &= b_3 + \sigma(b_1 + x \cdot w_1) \cdot u_1 + \sigma(b_2 + x \cdot w_2) \cdot u_2, \end{aligned}$$

where $\sigma(.)$ is an activation function, a mapping from $\mathbb{R}$ to $\mathbb{R}$.

Artificial neural networks

Activation function

- Capture non-linear relationships between x and v:

$$v = b_3 + \sigma(z_1) \cdot u_1 + \sigma(z_2) \cdot u_2,$$

  where $\sigma(.)$ is an activation function, a mapping from $\mathbb{R}$ to $\mathbb{R}$.

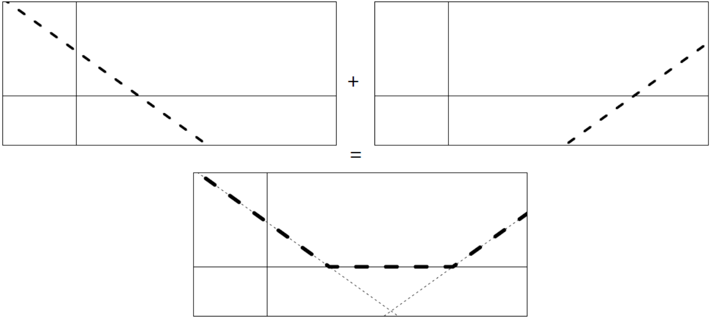
- For example, the rectified linear unit (ReLU) activation function:

$$\text{ReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

  Adding an activation function greatly increases the set of possible relations between x and v!

Artificial neural networks

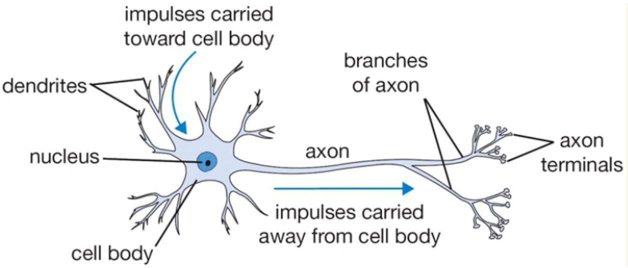
Activation function

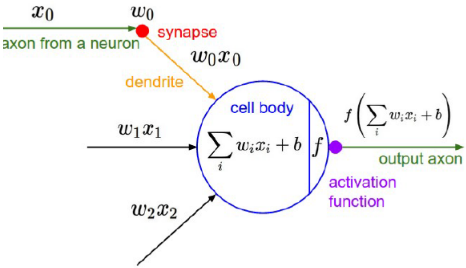


More activation functions: sigmoid, hyperbolic tan, leaky rectified linear unit, maxout.

Artificial neural networks

Artificial vs biological





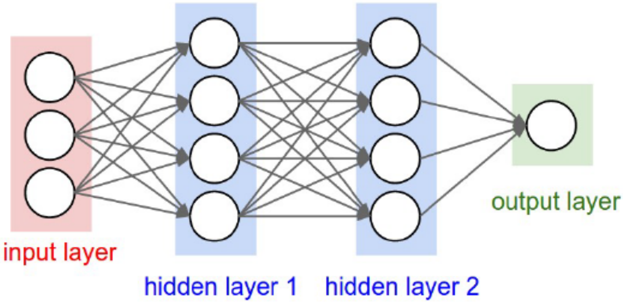
Artificial neural networks

The architecture

- Artificial Neural networks (NNs):
  - a collection of neurons
  - organized into an ordered set of layers
  - directed connections pass signals between neurons in adjacent layers
  - to train: update parameters describing the connections by minimizing loss function over training data
  - to predict: pass $\boldsymbol{x}_i$ to first layer, output of final layer is $\hat{y}_i$.
- The network is *dense* if each neuron in a layer receives an input from all the neurons present in the previous layer; *densely connected*.

Artificial neural networks

The architecture



A basic neural network. Source : http://blog.christianperone.com

This is a feedforward neural network - no loops!

Artificial neural networks

Terminology

- Using the NN language:
  - intercept called *the bias*
  - slopes called *weights*
  - L layers in total, with input layer denoted as layer 0
  - use a (from *activation*) to denote the output of a given layer.
- Let's start with a single layer network (called an artificial neuron).

Artificial neural networks

Terminology

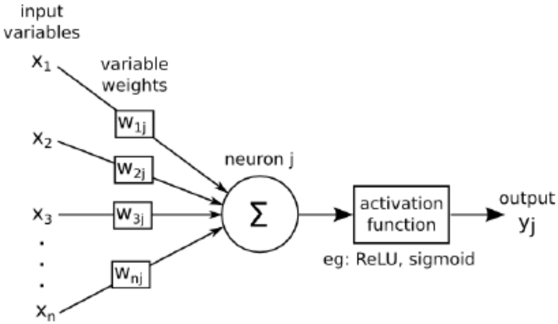
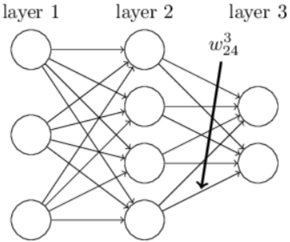


Figure 1: source: andrewjames turner.co.uk

Artificial neural networks

Terminology

- Use w^l_{jk} to denote the weight for the connection:
  - from neuron k in layer $(l-1)$
  - to neuron j in layer l.

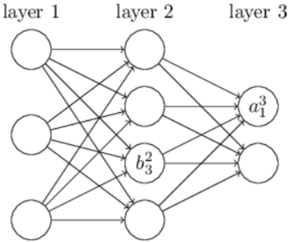


w^l_{jk} is the weight from the k^{th} neuron in the $(l-1)^{\text{th}}$ layer to the j^{th} neuron in the l^{th} layer

Artificial neural networks

Terminology

- Use
  - b_j^l for the bias of neuron j in layer l
  - a_j^l for the activation of neuron j in layer l.



Artificial neural networks

Terminology

- d-dimensional inputs

$$a^0 = x,$$

- weight $w^1_{j,k}$ on kth neuron in layer 0 within jth neuron in layer 1

$$w^1_j = (w^1_{j,1}, \ldots, w^1_{j,d})$$

- with bias term b^1_j

$$\begin{aligned} z^1_j &= w^1_{j,1} \cdot x_1 + \ldots + w^1_{j,d} \cdot x_d + b^1_j \\ &= < w^1_j, x > + b^1_j, \end{aligned}$$

- apply activation function

$$a^1_j = \sigma(z^1_j),$$

the output of neuron j in layer 1.

Artificial neural networks

Terminology

- With L layers in total: (in vectorized form)

$$\begin{aligned} z^l &= < w^l, a^{l-1} > + b^l \text{ (the weighted input)} \\ a^l &= \sigma(z^l), \end{aligned}$$

for l from 1 up to (and including) L.

- Finally,

$$a^L = \hat{y} = \tilde{\sigma}(z^L),$$

with examples of output activations: identity (regression), sigmoid (binary classification), softmax (muti-class output).

Stochastic gradient descent

Basic idea - GD

- We want to find

$$\min_{w} f(w),$$

  then gradient descent updates as follows (cfr. lecture on *boosting methods*)

$$w_{\text{new}} \quad = \quad w_{\text{old}} - \eta \cdot \nabla_w f(w_{\text{old}}),$$

  with learning rate η. Move in the direction the function locally decreases the fastest!

- A good choice for NNs because faster second-order methods involve the Hessian and this is infeasible when having tons of parameters.

Stochastic gradient descent

From GD to SGD

- Let $\mathcal{L}(w; y_i, x_i)$ be the loss function and w the parameters. For example,

$$\begin{aligned}\mathcal{L}(w; y_i, x_i) &= \frac{1}{2n}\sum_i (\hat{y}_i(w) - y_i)^2 \\ &= \frac{1}{n}\sum_i \mathcal{L}_i(w; y_i, x_i).\end{aligned}$$

- With the gradient descent update rule

$$w_{\text{new}} = w_{\text{old}} - \eta \cdot \nabla_w \mathcal{L}(w),$$

where $\nabla_w \mathcal{L}(w) = \frac{1}{n}\sum_i \nabla_w \mathcal{L}_i$.

Stochastic gradient descent
Mini-batches

- Stochastic gradient descent picks randomly m training inputs $x_1, \ldots, x_m$, a mini-batch:

$$\frac{\sum_{j=1}^{m} \nabla_w \mathcal{L}_j}{m} \approx \frac{\sum_i \nabla_w \mathcal{L}_i}{n} = \nabla_w \mathcal{L}.$$

- Thus,

$$\nabla_w \mathcal{L} \approx \frac{1}{m} \sum_{j=1}^{m} \nabla \mathcal{L}_j(w).$$

- Estimate the gradient only on a small subset of the entire training set.

Stochastic gradient descent

- Partition input randomly into disjoint groups $M_1,\ M_2, \ldots, M_{n/m}$.
- Updates:

$$
\begin{aligned}
w_{k+1} &= w_k - \frac{\eta}{m} \sum_{i \in M_1} \nabla \mathcal{L}_i \\
w_{k+2} &= w_{k+1} - \frac{\eta}{m} \sum_{i \in M_2} \nabla \mathcal{L}_i \\
&\vdots \\
w_{k+n/m+1} &= w_{k+n/m} - \frac{\eta}{m} \sum_{i \in M_{n/m}} \nabla \mathcal{L}_i.
\end{aligned}
$$

Speed up the process of doing gradient descent. Going through the entire data set is called an epoch.

Backward propagation of errors

- Compute the gradient of the loss function wrt all trainable parameters:
  - tons of parameters
  - need for efficient algorithm to calculate gradient
  - generic algorithm usable for arbitrary number of layers and neurons in each layer.
- The strategy (Rumelhart et al., 1986, Nature)

  backwards propagation of errors or backpropagation

  uses chain rule for derivatives.

Backward propagation of errors

- With a loss function $\mathcal{L}$, e.g. squared error loss

$$\mathcal{L}(y, a^L) \;=\; \frac{1}{2n}(y - a^L)^2.$$

- Equations defining backpropagation: (recall: $z_j^l = \sum_k w_{jk}^l a_k^{l-1} + b_j^l$)

  starting point - gradient terms of b_j^l

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial b_j^l} &= \frac{\partial \mathcal{L}}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial b_j^l} \\ &= \frac{\partial \mathcal{L}}{\partial z_j^l} \cdot 1 = \frac{\partial \mathcal{L}}{\partial z_j^l}. \end{aligned}$$

  Thus, problem centers around derivatives wrt z^l!

Backward propagation of errors

- With a loss function $\mathcal{L}$, e.g. squared error loss

$$\mathcal{L}(y, a^L) = \frac{1}{2n}(y - a^L)^2.$$

- Equations defining backpropagation: (recall: $z_j^l = \sum_k w_{jk}^l a_k^{l-1} + b_j^l$)

  starting point - gradient terms of weights w_{jk}^l

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_{jk}^l} &= \frac{\partial \mathcal{L}}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial w_{jk}^l} \\ &= \frac{\partial \mathcal{L}}{\partial z_j^l} \cdot a_k^{l-1}.\end{aligned}$$

  Thus, problem centers around derivatives wrt z^l!

Backward propagation of errors

- The derivatives in layer L are straightforward to calculate (with $a_j^L = \sigma(z_j^L)$)

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial z_j^L} &= \sum_k \frac{\partial \mathcal{L}}{\partial a_k^L} \cdot \frac{\partial a_k^L}{\partial z_j^L} \\
&= \frac{\partial \mathcal{L}}{\partial a_j^L} \cdot \frac{\partial a_j^L}{\partial z_j^L} \\
&= \frac{\partial \mathcal{L}}{\partial a_j^L} \cdot \sigma^{'}(z_j^L).
\end{aligned}
$$

An equation for the error in the output layer!

Backward propagation of errors

- The derivatives wrt z^l can be written as a function of derivatives wrt z^{l+1}.

- Using the chain rule

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial z_j^l} &= \sum_k \frac{\partial \mathcal{L}}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial z_j^l} \\
&= \sum_k \frac{\partial \mathcal{L}}{\partial z_k^{l+1}} \cdot w_{kj}^{l+1} \sigma^{'}(z_j^l),
\end{aligned}
$$

where we use

$$
z_k^{l+1} = \sum_j w_{kj}^{l+1} a_j^l + b_k^{l+1} = \sum_j w_{kj}^{l+1} \sigma(z_j^l) + b_k^{l+1}.
$$

Backward propagation of errors

- The vanilla implementation of the backpropagation alogrithm:

  1. Input x, set a^0 for the input layer.
  2. Feedforward: for each $l = 2, 3, \ldots, L$ compute
  $$z^l \quad = \quad < w^l, a^{l-1} > + b^l \text{ and } a^l = \sigma(z^l).$$
  3. Output error: compute $\frac{\partial \mathcal{L}}{\partial z_j^L} = \frac{\partial \mathcal{L}}{\partial a_j^L} \cdot \sigma^{'}(z_j^L)$.
  4. Backpropagate the error: for each $l = L-1, \ldots, 2$ compute $\frac{\partial \mathcal{L}}{\partial z_j^l} = \sum_k \frac{\partial \mathcal{L}}{\partial z_k^{l+1}} \cdot w_{kj}^{l+1} \sigma^{'}(z_j^l)$.
  5. Output: the gradient of the loss function $\frac{\partial \mathcal{L}}{\partial w_{jk}^l} = \frac{\partial \mathcal{L}}{\partial z_j^l} \cdot a_k^{l-1}$ and $\frac{\partial \mathcal{L}}{\partial b_j^l} = \frac{\partial \mathcal{L}}{\partial z_j^l}$.

Improving SGD

- Several improvements covered in the literature
  - to reduce overfitting
  - to reduce getting stuck in local saddle points
  - to converge in a smaller number of epochs.

Improving SGD

Reduce overfitting

- Methods to reduce overfitting:
  - different weight initialization
  - early stopping
  - regularization
  - dropout.

Improving SGD

Reduce overfitting: weight initialization, early stopping

- Update the weight initialization $\rightsquigarrow$ might drastically improve performance of a model with many layers.
- Early stopping to prevent overfitting:
  - calculate validation performance after each epoch
  - stop when this no longer improves
  - NNs are motivated by an optimization problem, but do not attempt to solve the optimization task.

Improving SGD

Reduce overfitting: regularization

- Prevent overfitting via regularization (cfr. lecture on Lasso and friends)
  - add (e.g.) ℓ_2-norm

    $$f_\lambda(w,b) \quad = \quad f(w,b) + \frac{\lambda}{2} \cdot \|w\|_2^2,$$

    with weights w and bias terms b, where only weights are penalized
  - with gradient

    $$\nabla_w f_\lambda \quad = \quad \nabla_w f(w,b) + \lambda \cdot w$$

    such that

    $$\begin{aligned} w_{\text{new}} \quad &\leftarrow \quad w_{\text{old}} - \eta \cdot \nabla_w f_\lambda(w_{\text{old}}) \\ &\leftarrow \quad w_{\text{old}} - \eta \cdot [\nabla_w f(w_{\text{old}}) + \lambda \cdot w] \\ &\leftarrow \quad [1 - \eta \cdot \lambda] \cdot w_{\text{old}} - \eta \cdot f(w_{\text{old}}). \end{aligned}$$

Improving SGD

Reduce overfitting: dropout

- Dropout:
  - randomly set activations to zero, with fixed probability p
  - both in forward propagation as well as backpropagation
  - only in training, all nodes turned on during prediction.

Improving SGD and regularization
Dropout

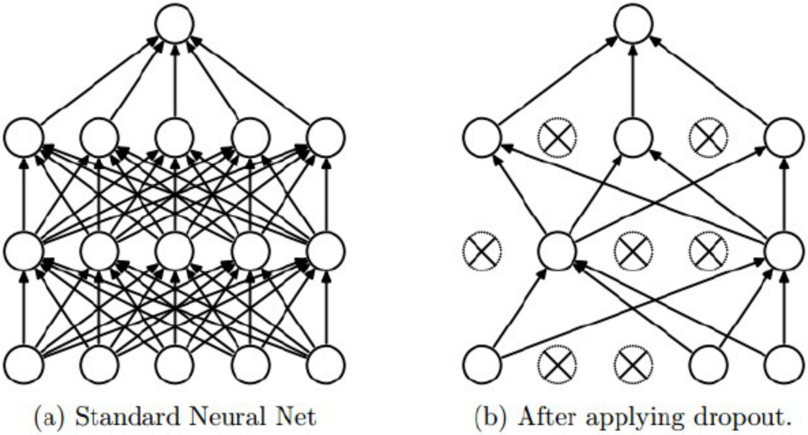


Figure 5: Dropout - source: http://blog.christianperone.com/

Improving SGD

Saddle points & faster convergence

- There is a fast proliferation of improved optimization algorithms.
- Examples:
  - momentum
  - RMSprop
  - adagrad, adamdelta
  - adam, adamax
  - natural gradient descent

Where to finetune your ANN?

- A list of tuning parameters/architecture choices:
  - learning rate η
  - regularization parameter λ
  - number of epochs
  - mini-batch size
  - number of layers
  - number of hidden neurons per layer
  - activation functions, choice of optimization strategy.

Classification with Artificial neural networks

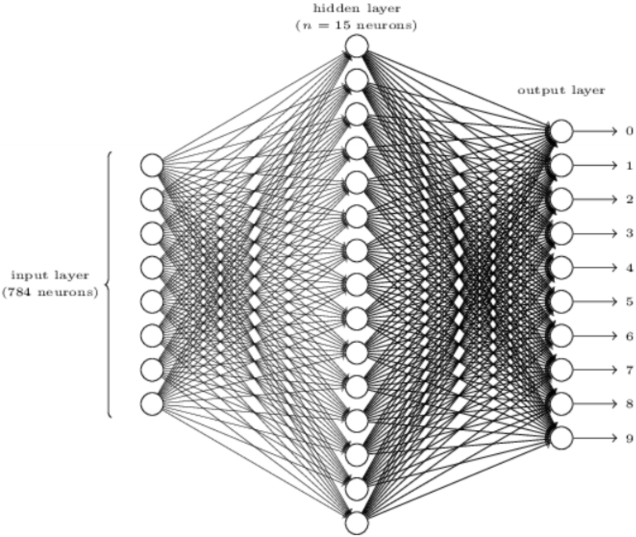
Back to the famous challenge

- **The goal**: try to recognize individual hand-written digits
  - take e.g. 28 by 28 greyscale image
  - $784 = 28 \times 28$ input neurons, with intensities between 0 (white) and 1 (black)
  - output layer has 10 layers

    neuron with highest activation value fires.

Classification with artificial neural networks

Back to the famous challenge



Classification with artificial neural networks

- Use a multi–valued output layer for classification tasks.
- One-hot encoding applied to output vector y

$$\begin{pmatrix} 2 \\ 4 \\ \vdots \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

- Use softmax function as activation in final layer

$$\begin{aligned} a_j^L &= \text{softmax}(z_j^L) \\ &= \frac{e^{z_j}}{\sum_k e^{z_k}}, \end{aligned}$$

for the jth output neuron. The last layer is then easily interpreted as sequence of probabilities.

Classification with artificial neural networks

- Instead of using squared error loss, use categorical cross-entropy

$$\mathcal{L}(a^L, y) \quad = \quad -\frac{1}{n}\sum_i \sum_k y_{ik} \cdot \log\left(a_k^L\right),$$

  as the loss function, where i runs over training data and k runs over the class labels.

- Think: softmax output layer and negative log-likelihood cost function.

- Work out the derivatives to set-up the backpropagation with this loss function.

Convolutional neural networks

The motivation

- Prediction tasks with images as inputs are a popular application of NNs.
- With a limited number of input pixels, put a weight on each pixel and use ANN.
- With large images use convolutional layers.

Convolutional neural networks

A convolution?

- The discrete convolution between f and g is

$$(f * g)(x) \;=\; \sum_t f(t) \cdot g(x+t).$$

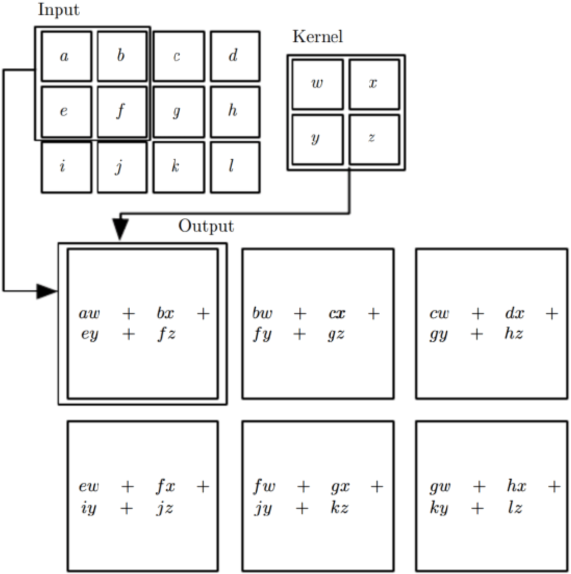
- With 2-dimensional signals (e.g. images), consider 2D-convolutions

$$(K * I)(i,j) \;=\; \sum_{m,n} K(m,n) \cdot I(i+m, j+n),$$

with K a convolutional kernel applied to a $2D$ signal (or image) I.

Convolutional neural networks

A convolution pictured



Convolutional neural networks

- With large images use convolutional layers:
  - apply small set of weights to subsections of the image
  - same weights across the image
  - use a kernel matrix, e.g. for a black and white image

$$K = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

  Take pixel value and subtract from this the pixel value to its immediate lower right.

Convolutional neural networks

- In a CNN:
  - apply several such kernels in a layer
  - with weights learned during training of the algorithm
  - different convolutions pick up different features (e.g. edges, texture, basic object types).
- Pooling layer: simplify information in output from convolutional layer.
- With input image in color: kernel matrix K is 3-dimensional array with weights applied to each color channel.

Recurrent neural networks

- To infer sequential data such as text or time series.
- Mathematically,

$$\begin{aligned} h_t &= Wx_t + b + Uh_{t-1} \\ &= Wx_t + b + UWx_{t-1} + Ub + U^2 h_{t-2} \\ &= \dots \end{aligned}$$

Recurrent neural networks
Unrolled

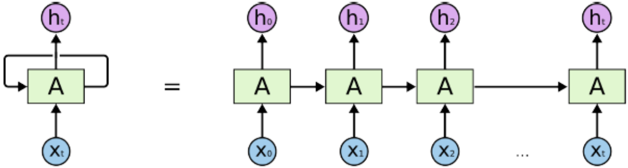


Figure 19: Unrolled representation of a RNN. Source : Understanding LSTM Networks by Christopher Olah - http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Yes, we CANN!

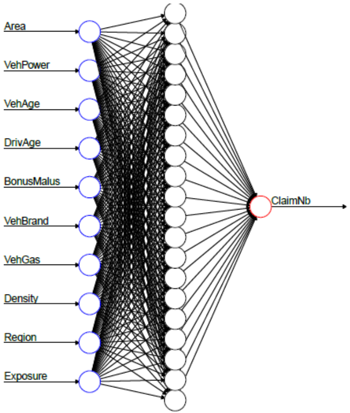
- Starting from the Poisson regression model

$$N_i \sim \text{POI}(\lambda(\boldsymbol{x}_i, d_i)),$$

with d_i exposure and $\boldsymbol{x}_i$ the features for insured i.

- Build e.g.
  - a fully-connected single hidden layer feed-forward neural network with 20 hidden neurons for modelling the regression function
  - a more general architecture with K hidden layers.

Yes, we CANN!



A shallow (left) and a deep (right) neural network.

Yes, we CANN!

- Some comments:
  - output layer has a single neuron, and exponential activation function
  - some preprocessing steps:

    ordered categorical features as continuous

    dummy coding or one-hot encoding to construct binary representation of nominal features

    put continuous covariates on the same scale
  - use Poisson loss function in `keras`.

Yes, we CANN!

Listing 3: R script for fitting networks in Keras

```
library(keras)

model <- keras_model_sequential()
model %>%
  layer_dense(units = q1, activation = 'tanh', input_shape = c(ncol(Xlearn))) %>%
  layer_dense(units = 1, activation = k_exp)

summary(model)

model %>% compile(
  loss = 'poisson',
  optimizer ='sgd'
)

fit <- model %>% fit(Xlearn, learn$ClaimNb, epochs=100, batch_size=10000)
```

`keras` implementation of a neural network for frequencies.

Yes, we CANN!

- Explore the following links
  - Case study: French MTPL claims
  - Insights from inside neural networks
  - Data analytics for non-life insurance pricing
  - Editorial: Yes we CANN!