Machine and Statistical Learning

library(neuralnet) from: Günther, Frauke, and Stefan Fritsch. "Neuralnet: Training of Neural Networks." *The R Journal* 2, no. 1 (2010): 30-38.

- Use of french and english words in the abstracts could actually be helpful.
- Multi-meaning words in english may be subdivided in french

Project 3

- Presentations on the last day of class
- Groups
- presentations are 8 minutes + 2 for questions
- 2 flavours: Beer and Maple Syrup!

Beer flavour!

- Expand on the beer project. This time your task is to visit all of the breweries in the lower mainland. On this tour, patrons will start at brewery 1 (your choice) and then hop aboard a bus to visit all of the lower mainland breweries (everything from Chilliwack west). Patrons may not survive a trip to all breweries so breweries should be location/quality prioritized. The route may not be the shortest but it should be the best. You will need to provide a guide with analytics so that patrons will know what they should try at each stop. Patrons should also be able to decide when they want to abandon the tour by looking at distance and quality measures. Emphasis is on the analytical value you add to the brewery tour company because that is what sells tickets.

~~Maple Syrup~~ SSC Abstracts

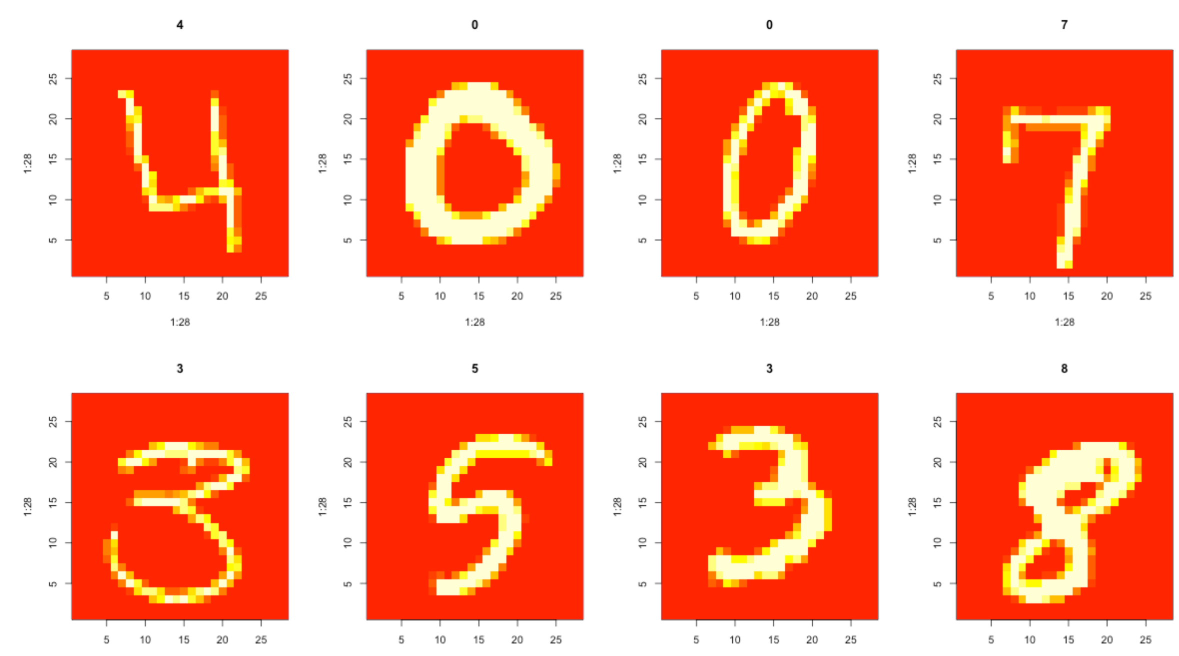
- Consider all of the abstracts from 2011-2015 and model the topics that are contained therein. Are there any signs of changing interests at the SSC? Could consider spatial models of topics. Must compare the topics studied from at least the 10 most dominant presenter groups (also include SFU if we are not one of them). Produce a grad school recommendation engine. Emphasis is on the value extracted from the topics and universities; topics by university, universities within topics… Comparisons should be easy and insightful

- Best submission here will be used to approach the SSC publication ‘Liaison’.

- Neural Network Models
- Feature Extraction
- Convolution Networks.
- auto encoders
- Backpropagation

Mixed National Institute of Standards and Technology database (MNIST) database

- 28 x 28 pixel images = 784 variables
- 10 different digits to be classified
- A **mix** of American Census Bureau employees and American high school students
- Goal is to classify these digits to human accuracy but faster and cheaper than humans can do it.



- Identify the wine producer from different wine traits.
- 12 wine variables:
- 1) Alcohol 2) Malic acid 3) Ash
  4) Alcalinity of ash 5) Magnesium
  6) Total phenols 7) Flavanoids
  8) Nonflavanoid phenols 9) Proanthocyanins
  10)Color intensity 11)Hue 12)OD280/OD315 of diluted wines 13)Proline

Species identification

- Iris data set is the kindergarten of classification systems
- 3 species of Iris flowers from the Gaspé penninsula

- library(neuralnet)

- data —> many parallel neurons —> predictions

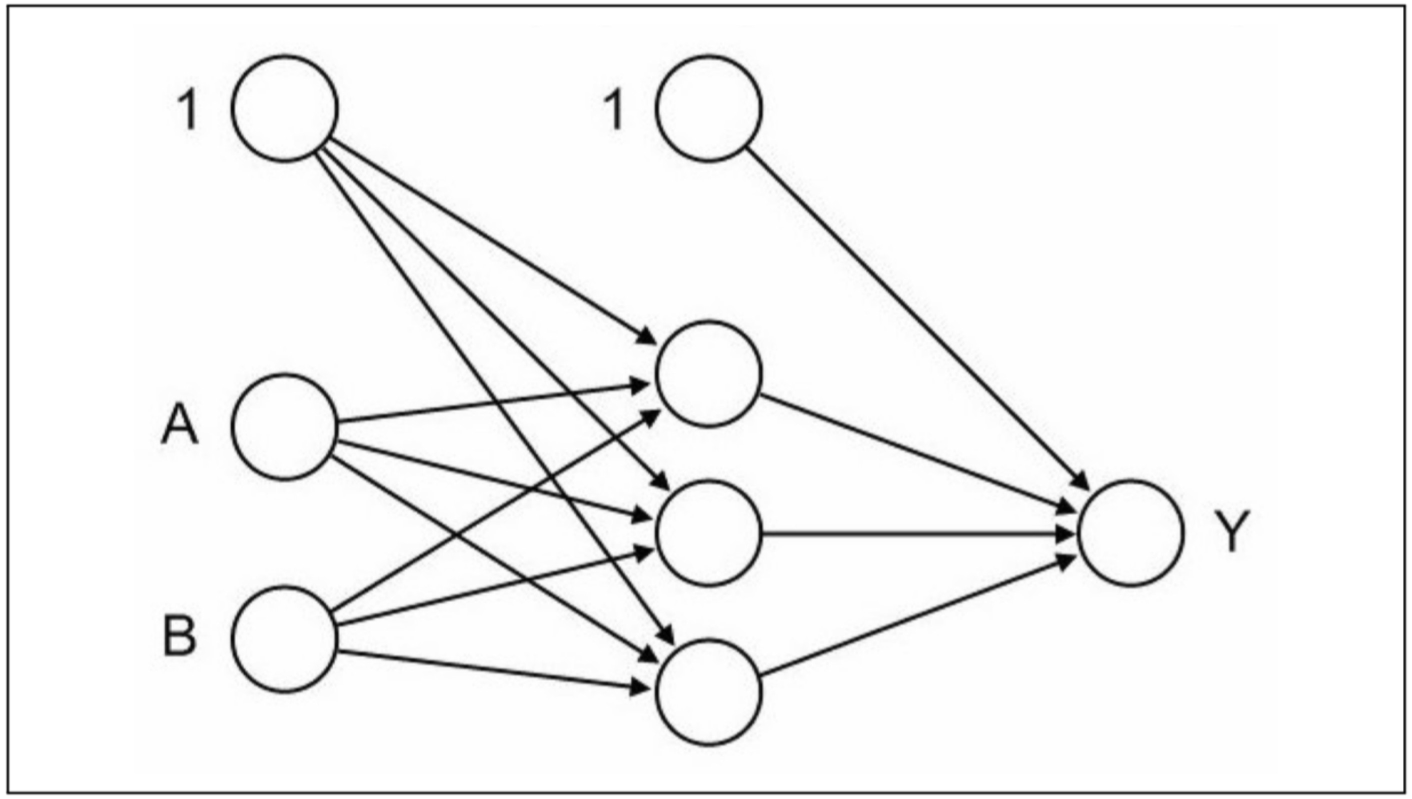


Figure 1: Example of a neural network with two input neurons (A and B), one output neuron (Y) and one hidden layer consisting of three hidden neurons.

Günther, Frauke, and Stefan Fritsch. "Neuralnet: Training of Neural Networks." *The R Journal* 2, no. 1 (2010): 30-38.

- with 2 input covariates (**x = {A,B}**) and n=3 neurons the output function o(**x**) is:

$$o(\mathbf{x}) = f\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$$

Multiple layers

- Often we build a hierarchical model:

$$o(\mathbf{x}) = f\left(w_0 + \sum_{j=1}^{J} w_j * f\left(w_{0,j} + \sum_{i=1}^{n} w_{i,j} x_{i,j}\right)\right)$$

- data —> many parallel neurons —> many parallel neurons —> …—>predictions

- In the simplest case the inner **f** is the single internal layer and the outer **f** is the output node

- library(neuralnet) uses the same **f** everywhere

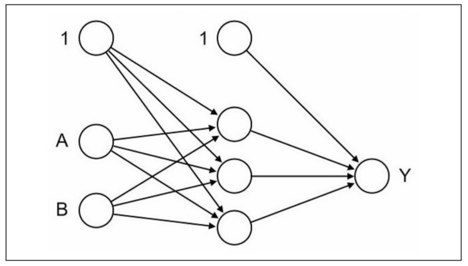
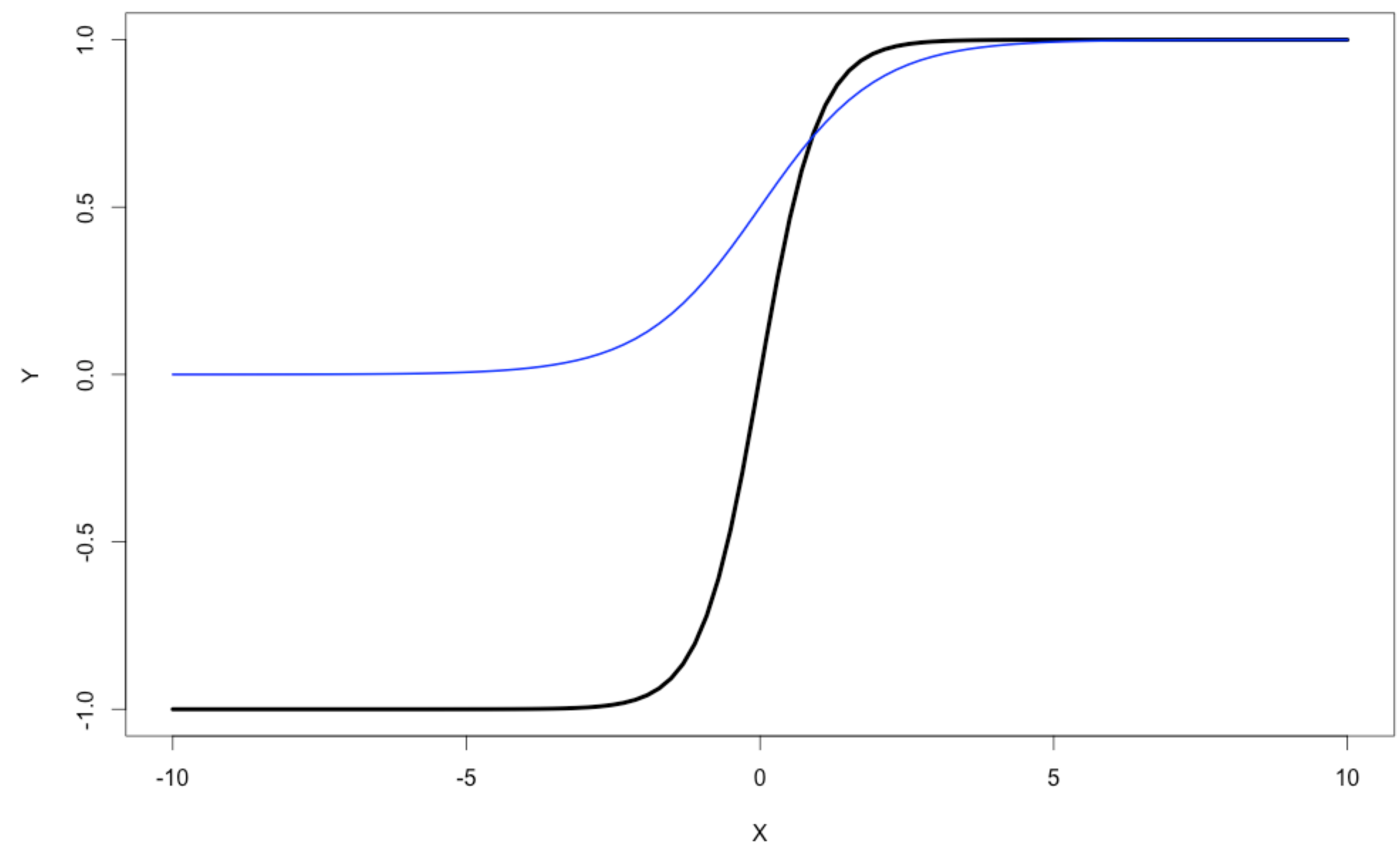


Figure 1: Example of a neural network with two input neurons (A and B), one output neuron (Y) and one hidden layer consisting of three hidden neurons.

Günther, Frauke, and Stefan Fritsch. "Neuralnet: Training of Neural Networks." *The R Journal* 2, no. 1 (2010): 30-38.

$$o(\mathbf{x}) = f\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$$

- f is bounded, monotone, and differentiable.
- Often logistic: $$f(u) = \frac{1}{1 + e^{-u}}$$
- or Hyperbolic Tangent: $$tanh(u) = \frac{e^{2u} - 1}{e^{2u} + 1}$$



Models

$$E = \frac{1}{2}\sum_{l=1}^{L}\sum_{h=1}^{H}(o_{l,h} - y_{l,h})^2$$

$$o(\mathbf{x}) = f\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$$

$$f(u) = \frac{1}{1 + e^{-u}}$$

- How does this relate to GLMs?

Model Training

- Fitting criteria, is often Squared Error Loss (i.e. Gaussian likelihood)

$$E = \frac{1}{2} \sum_{l=1}^{L} \sum_{h=1}^{H} (o_{l,h} - y_{l,h})^2$$

- Or Cross Entropy (i.e. log Binomial likelihood):

$$E = -\sum_{l=1}^{L} \sum_{h=1}^{H} [y_{l,h} log(o_{l,h}) + (1 - y_{l,h}) log(1 - o_{l,h})]$$

- for observation *l* at output node *h*

Model Training

- Optimization is usually gradient based.

$$\frac{\partial E}{\partial w} = 0|_{w=\hat{w}}$$

- Both **E** and **f** are differentiable, so gradients are analytic and often auto-differentiated (call this use of the chain rule *back propagation*)

$$\frac{\partial E}{\partial w} = \frac{dE}{do}\frac{do}{df}\frac{df}{dw}$$

- Often use CG variant or (random) subsets of dimensions to optimize at a time

Model Training

- Usually BFGS, CG, … define a step based on curvature.

- Back propagation uses learning rate **η**

$$w_k^{(t+1)} = w_k^{(t)} - \eta_k^{(t)} \left(\frac{\partial E^{(t)}}{\partial w_k^{(t)}} \right)$$

- iteration **t** and weight **k**

- library(neuralnet) uses this instead:

$$w_k^{(t+1)} = w_k^{(t)} - \eta_k^{(t)} * sign \left(\frac{\partial E^{(t)}}{\partial w_k^{(t)}} \right)$$

RStudio

- Iris
- Wine
- MNIST

Improvements

- Shutting off some weights via thresholding: sparse auto-encoders
- Imposing known structure: Convolution Neural Nets
- (much) better software

Real ML software

- Tensorflow, Theano, H2O,Caffee…, etc subdivide the neural net pieces to different GPU cores.
- Model Evaluation and optimization occur very quickly using parallel disjoint model segments and gradients.
- Several orders of magnitude speedup over R is typical