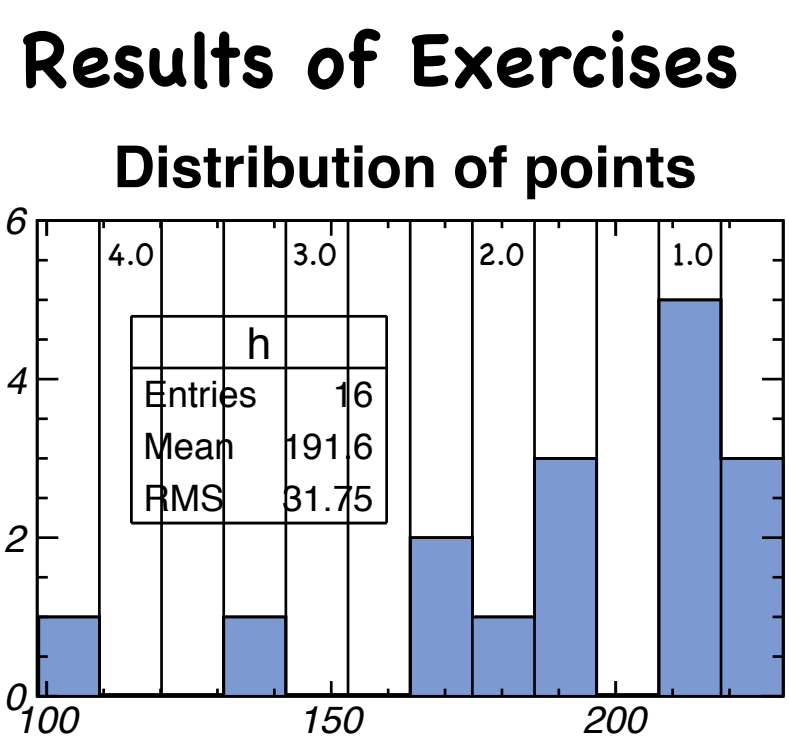
Statistical Methods of Data Analysis

Ulrich Husemann

Humboldt-Universität zu Berlin
Winter Semester 2010/2011

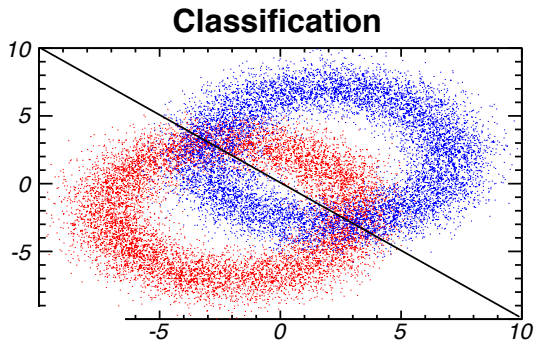
Results of Exercises

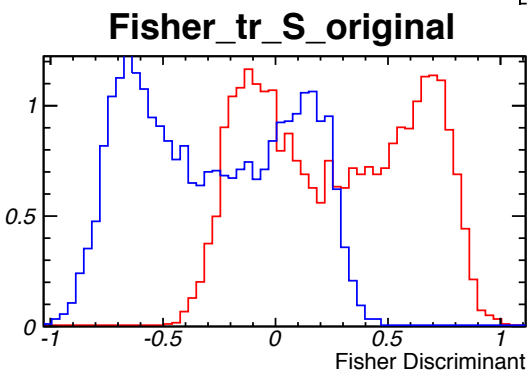


Grade scale renormalized to 95% of full number of points = 218.5

Fisher Discriminant

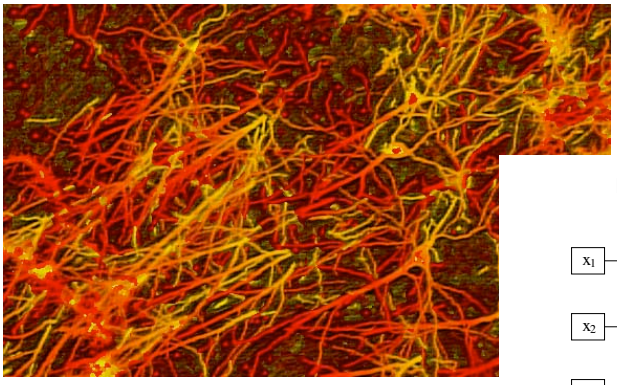
- TMVA Fisher discriminant:
  - Fisher x: -0.067
  - Fisher y: -0.067
  - Fisher offset: +0.005



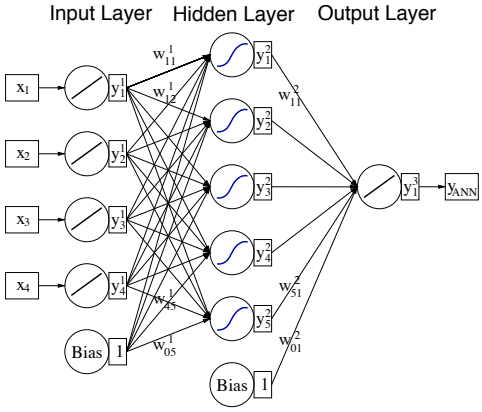


Neural Networks

[http://www.willamette.edu/~gorr/classes/cs449/brain.html]



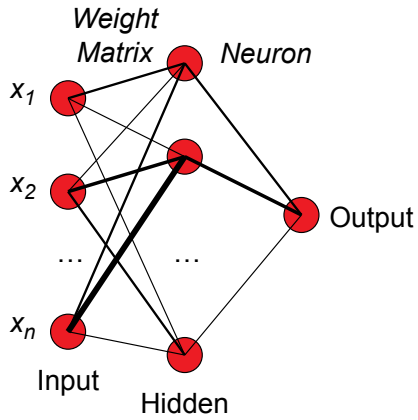
Human brain



[TMVA Users Guide]

Artificial Neural Networks

- Artificial neural network (ANN):
  - Used in many branches of data analysis (imaging, finance, ...)
  - Particle physics: classification of events as signal/background
- Typical ANN in particle physics: multilayer perceptron
  - Input layer: most sensitive kinematic variables (momenta, ...)
  - Hidden layer(s): linear combination of outputs of previous layer
  - Output: typically single discriminating variable, e.g. $\in[0,1]$



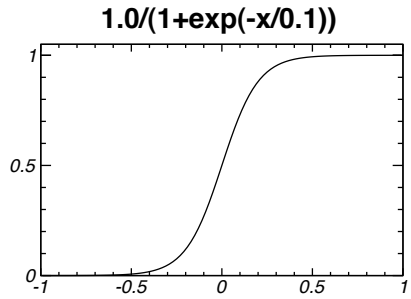
Output of each neuron j:

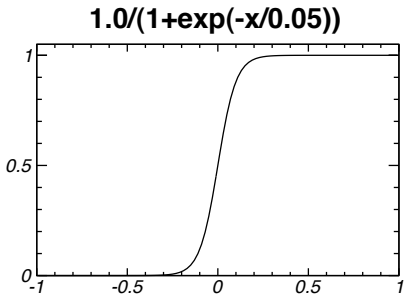
$$y_j = f\left(\sum_i w_{ij} x_i - T_j\right)$$

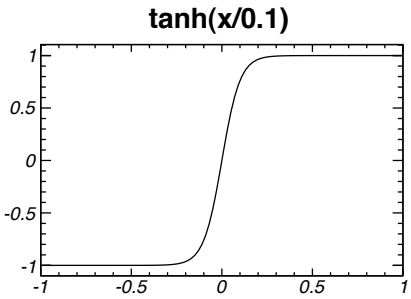
with w_{ij} weight, T_j threshold, sigmoid function for turn-on:

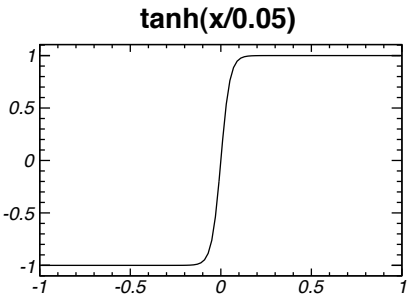
$$f(x) = \frac{1}{1+e^{-x}}$$

Transfer Functions



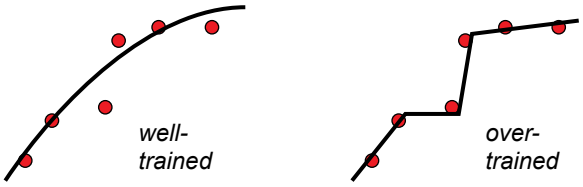






Artificial Neural Networks

- ANN training: determine weights w_{ij} and threshold T_j
  - Feed ANN with simulated data (signal & background)
  - Iterative optimization of weights & thresholds by "error back-propagation" → minimize differences between actual and desired output (1 for signal, 0 for background)
- ANNs are very powerful but "black box"
  - Input variables (& correlations) must be well-understood
  - Danger of over-training → ANN "snaps on" particular features of training sample



There's More...

- Recent comprehensive book on statistical learning:
  - T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning, Springer (2009)
  - http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- Extensions of ROOT
  - TMVA: multivariate analysis classes
  - RooFit: classes for PDFs and fitting
  - RooStats: classes for statistical data analysis
- Standard programming language for statisticians: R
  - http://www.r-project.org/
  - Many external packages available