# Lecture 21 Neural Network I

## ECE 625: Data Analysis and Knowledge Discovery

Di Niu

Department of Electrical and Computer Engineering
University of Alberta

March 30, 2021

# Outline

# Projection Pursuit

- Suppose that the vector $\mathbf{x}$ of independent variables is (possibly) of high dimension $p$.
- Are there interesting linear combinations $\alpha^T\mathbf{x}$ and possibly nonlinear transformations $f(\cdot)$ such that we might profitably model the data as

$$y = \sum_{m=1}^{M} f_m\left(\underbrace{\alpha_m^T\mathbf{x}}_{Zm}\right) + \varepsilon$$

Neural Network:
f_m= f (dont' need to learn)

for some small value of M?

- We assume that all $\|\alpha\| = 1$ so that the terms are possibly of comparable scales.
- Even then, there is a problem if the $x$'s are not measured in the same units.
- We typically scale the $x_j$ so that at least their magnitudes are comparable.

# Projection pursuit

- We call $\alpha^T \mathbf{x}$ the projection in the direction $\alpha$; hence the name *projection pursuit regression (PPR)*.
- For $M = 1$, the model is known as single index model in economics.
- The model is very general; as well as picking out individual $x$'s (e.g. $\alpha = (1, 0, \cdots, 0)^T$) we can model interactions and many other forms of terms.
- For instance

$$
\begin{aligned}
x_1 x_2 &= \frac{1}{2}\left(\frac{x_1 + x_2}{\sqrt{2}}\right)^2 - \frac{1}{2}\left(\frac{x_1 - x_2}{\sqrt{2}}\right)^2 \\
&= f_1\left(\alpha_1^T \mathbf{x}\right) + f_2\left(\alpha_2^T \mathbf{x}\right) \text{ for} \\
\alpha_1^T &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \alpha_2^T = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right), \\
f_1(t) &= \frac{t^2}{2}, f_2(t) = -\frac{t^2}{2}.
\end{aligned}
$$

# Algorithm

- A forward stage-wise strategy is used to minimize

$$\sum_{i=1}^{n}\left(y_i - \sum_{m=1}^{M} f_m\left(\alpha_m^T \mathbf{x}_i\right)\right)^2 .$$

- First suppose $M = 1$, so that $\sum_{i=1}^{n}\left(y_i - f_1\left(\alpha_1^T \mathbf{x}_i\right)\right)^2$ is to be minimized.
- If $\alpha_1^T$ is given, then $f_1(\cdot)$ can be obtained by nonparametric techniques, like spline smoothing or kernel smoothing.
- On the other hand if $f_1$ is given, we can update $\alpha_1$ using gradient descent.

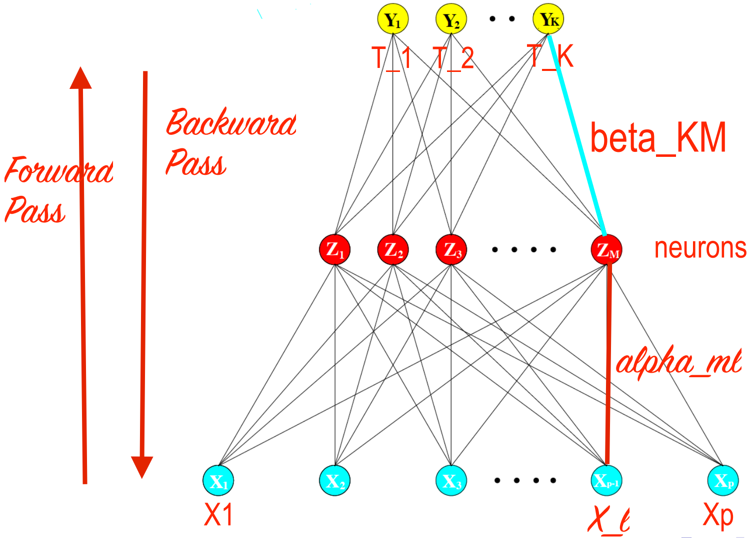
# Algorithm

- For $M > 1$, the problem can be solved in a stage-wise manner.
- Fit each $f_m$ to the residual of $f_1, \ldots, f_{m-1}$.
- At each step, the $f_m$ from previous steps can be readjusted using the backfitting procedure.
- The value $M$ can be chosen by stopping when the addition of another term does not improve the fit appreciably.
- The number of terms $M$ is usually estimated as part of the forward stage-wise strategy, or by cross validation.

# Neural Network

- The term neural network has evolved to encompass a large class of models and learning methods.
- Here we describe the most widely used Vanilla neural net, sometimes called the single hidden layer back-propagation network, or single layer perceptron. multiple layer perception (MLP)
- A neural network is a two-stage regression or classification model, typically represented by a network diagram.
- For regression, typically $K = 1$ and there is only one output unit $Y_1$ at the top.
- For $K$-class classification, there are $K$ units at the top, with the $k$th unit modeling the probability of class $k$. There are K target measurements $Y_k, k = 1, \cdots, K$ each being coded as a $0 - 1$ variable for the $k$th class.

# Neural Network



# Neural Network

alpha_m = (alpha_1m, ..., beta_pm)

beta_k = (beta_1k, ..., beta_Mk)

- Derived features $Z_m$ are created from linear combinations of the inputs, and then the target $Y_k$ is modeled as a function of linear combinations of the $Z_m$.
- That is

For classification g_k(T) is a softmax

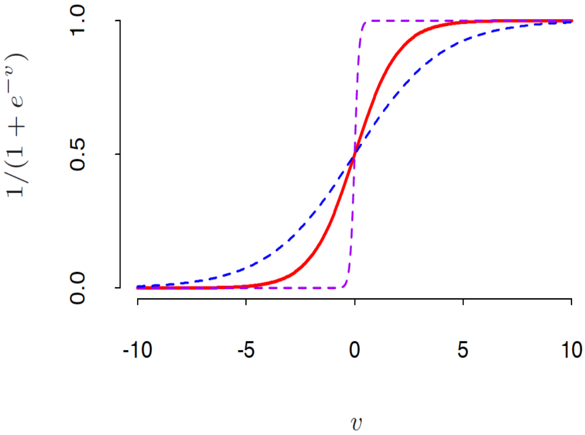
X: the input vector

$$
\begin{aligned}
Z_m &= \sigma(\alpha_{0m} + \alpha_m^T X), \ m = 1, \cdots, M, \\
T_k &= \beta_{0k} + \beta_k^T Z, \ k = 1, \cdots, K, \\
f_k(X) &= g_k(T), \ k = 1, \cdots, K,
\end{aligned}
$$

where $Z = (Z_1, \cdots, Z_M)$, and $T = (T_1, \cdots, T_K)$.

- The activation function $\sigma(v)$ is usually chosen to be the sigmoid $\sigma(v) = 1/(1 + e^{-v})$.
- Sometimes, Gaussian basis function can be used, producing what is known as a radial basis function network.

# Neural Network




Plot of $\sigma(sv)$ for $s = 1$ (red), $s = 1/2$ (blue) and $s = 10$ (purple), where $s$ controls activation rate.

# Neural Network

- The output function $g_k(T)$ allows a final transformation of the vector of outputs $T$.
- For regression we typically choose the identity function $g_k(T) = T_k$. k: kth target (response)
- For $K$-class classification, we choose the softmax function

Convert scores into a probability vector

$$g_K(T) = e^{T_k} / \sum_{k=1}^{K} e^{T_k}.$$

- The units in the middle of the network, computing the derived features $Z_m$, are called hidden units because the values $Z_M$ are not directly observed.
- The neural network model with one hidden layer has exactly the same form as the projection pursuit model with different link functions.
- The name neural networks derives from the fact that they were first developed as models for the human brain.

# Summary and Remark

- Projection pursuit
- Neural network
- Read textbook Chapter 11 and R code
- Do R lab