Lecture 21 Neural Network I

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

Outline

Projection Pursuit

- Suppose that the vector **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
$$y = \sum_{m=1}^{M} f_m \left(\alpha_m^T \mathbf{x}\right) + \varepsilon$$
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 xs are not measured in the same units.
- We typically scale the x_j so that at least their magnitudes are comparable.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Projection pursuit

- We call $\alpha^T\mathbf{x}$ the projection in the direction α; 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 xs (e.g. $\alpha = (1, 0, \cdots, 0)^T$) we can model interactions and many other terms.
- For instance

$$
\begin{aligned}
x_1x_2 &= \frac{1}{2}\left(\frac{x_1+x_2}{\sqrt{2}}\right)^2 - \frac{1}{2}\left(\frac{x_1-x_2}{\sqrt{2}}\right)^2 \\
&= f_1\left(\alpha_1^T\mathbf{x}\right) + f_2\left(\alpha_2^T\mathbf{x}\right) \text{ for} \\
\alpha_1^T &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \alpha_2^T = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right), \\
f_1(t) &= \frac{t^2}{2}, f_2(t) = -\frac{t^2}{2}.
\end{aligned}
$$

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 α_1^T is given, then $f_1(\cdot)$ can be gotten by the nonparametric techniques, like basis expansion or kernel smoothing.
- On the other hand if f_1 is given, and we have a trial value $\alpha_{(0)}$ of α, then it can be updated through a weighted least square (WLS).

$$\alpha_{(1)} = \alpha_{(0)} + \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z},$$

with

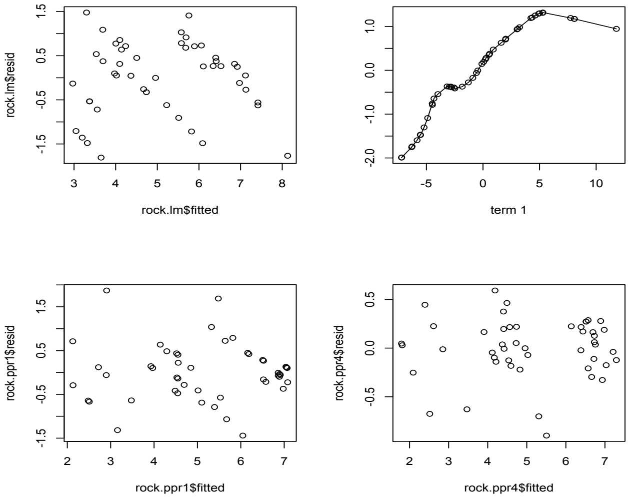
$$z_i = \frac{y_i - f_1\left(\alpha_{(0)}^T \mathbf{x}_i\right)}{f_1'\left(\alpha_{(0)}^T \mathbf{x}_i\right)}, \text{ and } w_i = \left[f_1'\left(\alpha_{(0)}^T \mathbf{x}_i\right)\right]^2.$$

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Algorithm

- For $M > 1$ this is applied by nonparametric techniques, using the residuals from all $M - 1$ other fits, at each stage.
- The value M can be chosen by stopping when the addition of another term does not improve the fit appreciably.
- At each step, the f_m from previous steps can be readjusted using the backfitting procedure.
- The number of terms M is usually estimated as part of the forward stage-wise strategy, or by cross validation.

Rock data

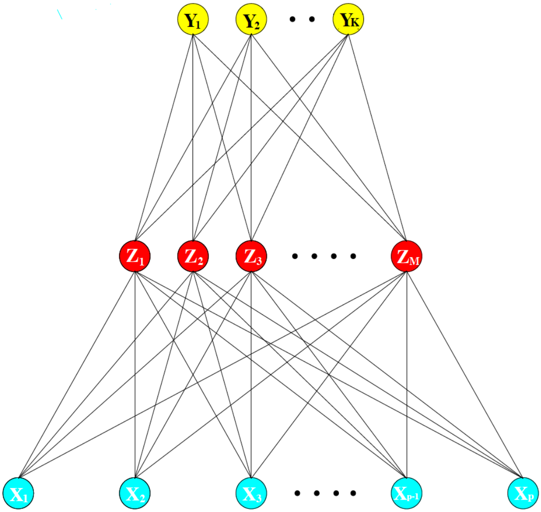

rock.lm$resid
rock.lm$fitted
term 1
rock.ppr1$resid
rock.ppr1$fitted
rock.ppr4$resid
rock.ppr4$fitted

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

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.
- 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 kth 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 kth class.

Neural Network



UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

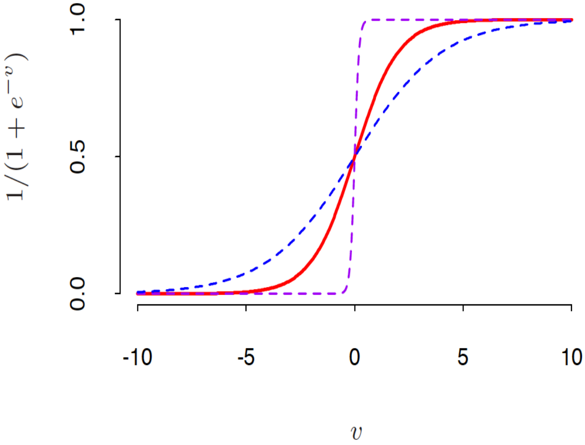
Neural Network

- 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

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

  where $Z = (Z_1, \cdots, Z_M)^T$, and $T = (T_1, \cdots, T_K)^T$.
- 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 $g_k(T) = T_k$.
- For K-class classification, we choose the softmax function

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

- 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.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Summary and Remark

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