A Partial Solution Manual for: *The Elements of Statistical Learning* by Jerome Friedman, Trevor Hastie, and Robert Tibshirani

Wenhao Wu

wnhwu@ucdavis.edu

Dept. ECE, UC Davis

February 5, 2016

Contents

Preface

This work is expected to be used as a supplementary material for Weatherwax and Epstein's solution manual [1], which I found to be very helpful when self-studying this popular textbook. The numbering of chapters and problems are based on the 2nd edition (10th printing with corrections, Jan 2013) available online [2].

The author was not able to solve all the excercises. Even for the solutions included we expect many mistakes and shortcomings. It would be of great help if people could suggest possible solutions or help us find and correct the errors so this solution manual can be continuously improved to benefit more interested readers. We are also open to all comments and criticisms. Our contact information can be found at the website holding this draft [3].

Acknowledgment

Chapter 2

Overview of Supervised Learning

Chapter 3

Linear Methods for Regression

Chapter 4

Linear Methods for Classification

Chapter 5

Basis Expansions and Regularization

Chapter 6

Kernel Smoothing Methods

Chapter 7

Model Assessment and Selection

Chapter 8

Model Inference and Averaging

Chapter 9

Additive Models, Trees, and Related Methods

Chapter 10

Boosting and Additive Trees

Chapter 11

Neural Networks

Ex. 11.1

In (11.5), set $K = 1$, $g_1(T) = T$, we have

$$f_1(X) = \beta_{01} + \beta_1^T Z = \beta_{01} + \sum_{m=1}^{M} \beta_{m1}\sigma(\alpha_{0m} + \alpha_m^T X) \tag{11.1}$$

The correspondence between (11.1) and (11.5) becomes clearer, as enumerated in Table 11.1

Table 11.1: Correspondonce between the project pursuit regression and the neural network

(11.1)	(11.5)
ω_m	α_m
$g_m(\cdot)$	β_{01}, $\beta_{m1}\sigma(\alpha_{0m} + \alpha_m^T X)$

Ex. 11.2

$$\frac{\partial f}{\partial X} = \sum_{m=1}^{M} \beta_m[\sigma(\cdot)(\sigma(\cdot) - 1)]\alpha_m \tag{11.2}$$

$$\frac{\partial^2 f}{\partial X \partial X^T} = \sum_{m=1}^{M} \beta_m[(2\sigma(\cdot) - 1)(\sigma(\cdot) - 1)\sigma(\cdot)]\alpha_m\alpha_m^T \tag{11.3}$$

Since $\sigma(\alpha_{0m} + \alpha_m^T X) \approx 1/2$ when $\alpha_{0m} \approx 0$ and $\alpha_m \approx 0$, therefore $\frac{\partial^2 f}{\partial X \partial X^T} \approx 0$, i.e. the resulting model is nearly linear.

Ex. 11.3

$$R(\theta) = -\sum_{i=1}^{N} R_i(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij} \log g_j(T) \tag{11.4}$$

Note that different from regression, each softmax function $g_j(T)$, $j = 1, \ldots, K$ is a function of all $T_1, \ldots, T_K$.

$$\frac{\partial R_i}{\partial \beta_{km}} = -\sum_{j=1}^{K} \frac{y_{ij}}{g_j} \frac{\partial g_j}{\partial T_k} z_{mi} = \delta_{ki} z_{mi} \tag{11.5a}$$

$$\begin{aligned} \frac{\partial R_i}{\partial \alpha_{ml}} &= -\sum_{j=1}^{K} \frac{y_{ij}}{g_j} \sum_{k=1}^{K} \frac{\partial g_j}{\partial T_k} \beta_{km} \sigma'(\alpha_m^T x_i) x_{il} \\ &= \left[\sigma'(\alpha_m^T x_i) \sum_{k=1}^{K} \beta_{km} \delta_{ki} \right] x_{il} = s_{mi} x_{il} \end{aligned} \tag{11.5b}$$

It is noted that

$$\frac{1}{g_j} \frac{\partial g_j}{\partial T_k} = \begin{cases} 1 - g_j & j = k \\ -g_k / \exp(T_j) & j \neq k \end{cases} \tag{11.6}$$

As a result, although $g_j(T)$ depends on all $T_1, \ldots, T_K$, $(\partial g_j / \partial T_k)/g_j$ can still be locally evaluated and propagated downward over the link (T_k, g_j). Consequently, the forward and backward propagation equations are pretty much the same as those for the square error loss function. In the forward pass for record x_i, $i = 1, \ldots, N$, the weights β_{km} and α_{ml} are fixed and the predicted $\hat{g}_j(T_i)$ are evaluated. In the backward pass, $(y_{ij}/g_j)(\partial g_j / \partial T_k)$ are evaluated and propagated to T_k, where δ_{ki} is computed, and then back-propagated to give s_{mi} at Z_m. Then the gradients are evaluated as in Eq. (11.5). The gradient descent update is exactly the same as (11.13).

Ex. 11.4

If the network has no hidden layer, we have

$$g_j(x) = \frac{\exp(T_j)}{\sum_{k=1}^{K} \exp(T_k)} = \frac{\exp\left(\beta_j^T x\right)}{\sum_{k=1}^{K} \exp(\beta_k^T x)}, \tag{11.7}$$

exactly the same as the multinomial logistic model.

Ex. 11.5 (Program)

Ex. 11.6 (Program)

Ex. 11.7 (Program)

Chapter 12

Support Vector Machines and Flexible Discriminants

Chapter 13

Prototype Methods and Nearest-Neighbors

Chapter 14

Unsupervised Learning

Chapter 15

Random Forests

Chapter 16

Ensemble Learning

Chapter 17

Undirected Graphical Models

Chapter 18

High-Dimensional Problems

References

[1] J. L. Weatherwax and D. Epstein, "A solution manual and notes for: The elements of statistical learning by jerome friedman, trevor hastie, and robert tibshirani," June 2013. [Online]. Available: http://waxworksmath.com/Authors/G_M/Hastie/hastie.html

[2] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer series in statistics Springer, Berlin, 2009.

[3] W. Wu, "A partial solution manual for: The elements of statistical learning by jerome friedman, trevor hastie, and robert tibshirani," 2016. [Online]. Available: https://github.com/huragok/IDA