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 13, 2016

Contents

Preface

This work is expected to be used as a supplementary material for Weatherwax and Epstein's solution manual [?], 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 [?].

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 [?].

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

Ex. 12.1

Firstly, we prove that for (12.8), the optimal solution must satisfy $\hat{\xi}_i = [1 - y_i(x_i^T\hat{\beta} + \hat{\beta}_0)]_+$. To see this, from the constraints in (12.8), we have $\hat{\xi}_i \geq [1 - y_i(x_i^T\hat{\beta} + \hat{\beta}_0)]_+$. Assume for contradiction that $\exists i$ such that $\hat{\xi} > [1-y_i(x_i^T\hat{\beta}+\hat{\beta}_0)]_+$, then setting $\hat{\xi}_i \leftarrow [1-y_i(x_i^T\hat{\beta}+\hat{\beta}_0)]_+$ results in smaller objective in (12.8), which is in contradiction to the fact that $\hat{\xi}_i$ is from an optimal solution.

On the other hand, $\xi_i = [1 - y_i(x_i^T\beta + \beta_0)]_+ \Rightarrow \xi_i \geq 0$, $y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i$. Therefore, the solution to (12.8) is the same as

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \tag{12.1}$$

$$\text{s.t. } \xi_i = [1 - y_i(x_i^T\beta + \beta_0)]_+, \ \forall i \tag{12.2}$$

which is exactly the same as (12.25).

Ex. 12.2

Define kernel $K(a,b) = \sum_{j=1}^{p} a_j b_j$, i.e. $\psi_j(x) = x_j, \gamma_j = 1$ for $j = 1,\ldots,p$. Consequently, $g(x) = \sum_{j=1}^{p}\beta_j x_j \Leftrightarrow g(x) \in \mathcal{H}_K$. Consequently,

$$(12.25) \Leftrightarrow \min_{g,\beta_0}\sum_{i=1}^{N}[1 - y_i(g(x_i) + \beta_0)]_+ + \frac{\lambda}{2}\|g\|^2_{\mathcal{H}_K} \tag{12.3}$$

Denote $L(y_i, g(x_i); \beta_0) = [1 - y_i(g(x_i) + \beta_0)]_+ = L_i(\beta_0)$, then

$$(12.25) \Leftrightarrow \min_{\beta_0}\left\{\min_{g}\sum_{i=1}^{N} L_i(\beta_0) + \frac{\lambda}{2}\|g\|^2_{\mathcal{H}_K}\right\}. \tag{12.4}$$

where the inner min must have a solution in the form of $g(x) = \sum_{i=1}^{N}\alpha_i K(x, x_i)$ as per

(5.50)(5.51), and we have $\|g\|^2_{\mathcal{H}_K} = \alpha^T K\alpha$. Therefore

$$(12.25) \Leftrightarrow \min_{\beta_0} \left\{ \min_{\alpha} \sum_{i=1}^{N} [1 - y_i(\sum_{j=1}^{N} \alpha_j K(x_j, x_i) + \beta_0)]_+ + \frac{\lambda}{2}\alpha^T K\alpha \right\} \tag{12.5}$$

$$\Leftrightarrow \min_{\beta_0,\alpha} \sum_{i=1}^{N} [1 - y_i(\sum_{j=1}^{N} \alpha_j K(x_j, x_i) + \beta_0)]_+ + \frac{\lambda}{2}\alpha^T K\alpha \tag{12.6}$$

Ex. 12.3

Similar to Ex. (12.2). Denote $g(x) = \sum_{m=1}^{M} \beta_m h_m(x)$. Without penalting the constant term, we have

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - \beta_0 - g(x_i)) + \frac{\lambda}{2}\sum_{m=1}^{M} \beta_m \tag{12.7}$$

Again we break the minimization problem into 2 steps:

$$\min_{\beta_0,\beta} H(\beta, \beta_0) = \min_{\beta_0} \left\{ \min_{\beta|\beta_0} H(\beta, \beta_0) \right\} \tag{12.8}$$

Consider saquare error loss $V(r) = r^2$, the inner min problem is in the form of

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - H\beta)^2 + \frac{\lambda}{2}\beta^T\beta \tag{12.9}$$

$$\Leftrightarrow \min_{\alpha} \|\mathbf{y} - \beta_0\mathbf{1} - \mathbf{K}\boldsymbol{\alpha}\|_F^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \tag{12.10}$$

whose solution is $\hat{\alpha} = (\mathbf{K} + \lambda\mathbf{I}/2)^{-1}\mathbf{y}_{\beta_0}$, $\mathbf{y}_{\beta_0} = \mathbf{y} - \beta_0\mathbf{1}$. Consequently, the outer min problem w.r.t β_0 is in the form of

$$\min_{\beta_0} \mathbf{y}_{\beta_0}^T \left[\mathbf{I} - (\mathbf{K} + \lambda\mathbf{I}/2)^{-1}\mathbf{K}\right] \mathbf{y}_{\beta_0} \tag{12.11}$$

which is a quadratic problem.

Ex. 12.4

(a)

$$\text{Left} = (x - \bar{x}_k)^T UU^T (x - \bar{x}_k) - (x - \bar{x}_{k'})^T UU^T (x - \bar{x}_{k'}) \tag{12.12}$$

where $U = W^{-1/2}V^*$, the L columns of V^* are the eigen vectors of $B^* = (W^{-1/2})^T BW^{-1/2}$, where B is the between-class covariance.

$$\text{Right} = (x - \bar{x}_k)^T W^{-1}(x - \bar{x}_k) - (x - \bar{x}_{k'})^T W^{-1}(x - \bar{x}_{k'}) \tag{12.13}$$

Consequently,

$$\begin{aligned}
&\text{Left} - \text{Right} \\
=&2(\bar{x}_k - \bar{x}_{k'})^T (W^{-1} - UU^T)x + (\bar{x}_k - \bar{x}_{k'})^T [W^{-1} - UU^T](\bar{x}_k + \bar{x}_{k'}) && (12.14)\\
=&2(\bar{x}_k - \bar{x}_{k'})^T W^{-1/2}(I - V^*(V^*)^T)(W^{-1/2})^T x \\
&+(\bar{x}_k - \bar{x}_{k'})^T W^{-1/2}(I - V^*(V^*)^T)(W^{-1/2})^T(\bar{x}_k + \bar{x}_{k'}) && (12.15)
\end{aligned}$$

Since $(\bar{x}_k - \bar{x}_{k'})^T \in R(M)$ (row space), $(\bar{x}_k - \bar{x}_{k'})^T W^{-1/2} \in R(M^*)$, therefore $(W^{-1/2})^T(\bar{x}_k - \bar{x}_{k'}) \in C(V^*)$ (column space). Therefore

$$(\bar{x}_k - \bar{x}_{k'})^T W^{-1/2}(I - V^*(V^*)^T) = 0 \tag{12.16}$$

thus Left = Right.

(b) ???

Ex. 12.5 (Program)

Ex. 12.6

(a) The i-th row of $\mathbf{Y}\theta$ is

$$(\mathbf{Y}\theta)_i = \sum_{j=1}^{K} 1(Y_{ij} = 1)\theta_j = \theta(g_i) \tag{12.17}$$

(since there are exactly one j where $_{ij} = 1$ for each i). The i-th row of $\mathbf{H}\beta$ is

$$(\mathbf{H}\beta)_i = \sum_{j=1}^{K} \beta_j h_j(x_i) \tag{12.18}$$

therefore

$$\sum_{i=1}^{N} (\theta(g_i) - \beta^T h(x_i))^2 = \|\mathbf{Y}\theta - \mathbf{H}\beta\|^2 \tag{12.19}$$

(b) According to the definition, $(\mathbf{D}_\pi)_{kk}$ is the empirical frequency of class k, and θ_k is the score for class k. $\theta^T \mathbf{D}_\pi 1 = 0$ implies that the average score over the N records is 0; $\theta^T \mathbf{D}_\pi \theta = 1$ means the variance of the over the N records is 1.

(c) Fixing θ the optimal β is

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Y}\theta \tag{12.20}$$

therefore (12.65) can be rewritten as

$$\min_\theta \|(\mathbf{I}-\mathbf{S})\mathbf{Y}\theta\|^2 \Leftrightarrow \min_\theta \theta^T\mathbf{Y}^T\mathbf{Y}\theta - \theta^T\mathbf{Y}^T\mathbf{S}\mathbf{Y}\theta \tag{12.21}$$

where $\mathbf{S} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$. Since $\theta^T\mathbf{Y}^T\mathbf{Y}\theta = N$, this minimization is equivalent to

$$\max_\theta \theta^T\mathbf{Y}^T\mathbf{S}\mathbf{Y}\theta \tag{12.22}$$

(d) Suppose that the SVD of $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, then $\mathbf{S} = \mathbf{U}\mathbf{U}^T$ where $\mathbf{U}$ is a N-by-l orthonormal matrix. Therefore $\mathbf{S}$ has L eigenvalues of 1 and $N-L$ eigenvalues of 0. Since constant function is included in h_j, $\mathbf{H} \neq 0$, therefore $L > 0$, so the largest eigenvalue is 1.

(e) (12.53) can be rewritten as

$$ASR = \frac{1}{N}\|\mathbf{Y}\mathbf{\Theta} - \mathbf{H}\mathbf{B}\|_F^2 \tag{12.23}$$

Similar to (c) the solution is the same as

$$\max_\Theta \operatorname{tr}\left\{\mathbf{\Theta}^T\mathbf{Y}^T\mathbf{S}\mathbf{Y}\mathbf{\Theta}\right\} \tag{12.24}$$

$$\text{s.t. } \mathbf{\Theta}^T\mathbf{Y}^T\mathbf{Y}\mathbf{\Theta} = \mathbf{I} \tag{12.25}$$

Therefore $\mathbf{Y}\Theta$ are the K largest eigenvectors of $\mathbf{S}$.

Ex. 12.7

The penalized optimal scoring problem is in the form of

$$\min_{\Theta,\mathbf{B}} \|\mathbf{Y}\mathbf{\Theta} - \mathbf{H}\mathbf{B}\|_F^2 + \lambda \operatorname{tr}\left(\mathbf{B}^T\mathbf{\Omega}\mathbf{B}\right) \tag{12.26}$$

Given $\mathbf{\Theta}$, the optimal $\mathbf{B}$ is

$$\hat{\mathbf{B}} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{\Omega})^{-1}\mathbf{H}^T\mathbf{Y}\mathbf{\Theta} \tag{12.27}$$

Substitute into Eq. (12.26), we have

$$\min_{\mathbf{\Theta}} \operatorname{tr}\left(\mathbf{\Theta}^T\mathbf{Y}^T\mathbf{Y}\mathbf{\Theta} - \mathbf{\Theta}^T\mathbf{Y}^T\mathbf{S}\mathbf{Y}\mathbf{\Theta}\right) \tag{12.28}$$

$$\text{s.t. } \mathbf{\Theta}^T\mathbf{D}_\pi\mathbf{\Theta} = \mathbf{I} \tag{12.29}$$

where $\mathbf{S} = \mathbf{H}(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{\Omega})^{-1}\mathbf{H}^T$. Therefore $\mathbf{Y\Theta}$ are still the eigenvectors of $\mathbf{S}$.

Ex. 12.8

I found the proof to this problem on [?]. I am trying to follow it the best I can and here is my interpretation. Assuming that $\bar{x} = 0$. We first perform the generalized SVD:

$$(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{UDV}^T \tag{12.30}$$

$$\text{s.t. } \mathbf{U}^T\mathbf{Y}^T\mathbf{YU} = \mathbf{I},\ \mathbf{V}^T\mathbf{X}^T\mathbf{XV} = \mathbf{I} \tag{12.31}$$

Later we will show that both β_l and v_l are proportional to the colunns of $\mathbf{V}$. From the GSVD, $\mathbf{U}$ and $\mathbf{V}$ satisfy the following 2 equations:

$$\mathbf{U}^T\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{YU} = \mathbf{D}^2 \tag{12.32a}$$

$$\mathbf{V}^T\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{XV} = \mathbf{D}^2 \tag{12.32b}$$

of which the proof is trivial. First we show that the LDA's discriminant directions v_l are parallel to the columns of $\mathbf{V}$:

Proposition 12.1. *For the LDA problem (Fisher)*

$$\max_{\mathbf{A}} \operatorname{tr}\big(\mathbf{A}^T\mathbf{BA}\big),\ \textit{s.t. } \mathbf{A}^T\mathbf{WA} = \mathbf{I} \tag{12.33}$$

where

$$\mathbf{B} = \mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X} \tag{12.34a}$$

$$\mathbf{W} = \mathbf{T} - \mathbf{B} \tag{12.34b}$$

$$\mathbf{T} = \mathbf{X}^T\mathbf{X} \tag{12.34c}$$

are the between-class, within-class and total variance (up to normalization), the solution is

$$\hat{\mathbf{A}} = \mathbf{V}(\mathbf{I} - \mathbf{D}^2)^{-1/2} \tag{12.35}$$

Proof. From Eq. (12.32b) and the second constraint of the GSVD, it is easy to see $\hat{\mathbf{A}}^T\mathbf{W}\hat{\mathbf{A}} = \mathbf{I}$. On the other hand, $\hat{\mathbf{A}}$ diagonalizes $\mathbf{B}$ by $\hat{\mathbf{A}}^T\mathbf{B}\hat{\mathbf{A}} = (\mathbf{I} - \mathbf{D}^2)^{-1}\mathbf{D}^2$. □

Next we show that the β_l from optimal scoring are also parallel to the columns of $\mathbf{V}$

Proposition 12.2. *The optimal scoring problem as in Eq.* (12.24) *has solution* $\hat{\mathbf{\Theta}} = \mathbf{U}$.

Proof. From the first constraint of GSVD, obviously $\mathbf{U}^T\mathbf{Y}^T\mathbf{YU} = \mathbf{I}$. from Eq. (12.32a), $\mathbf{U}$ diagonalizes $\mathbf{Y}^T\mathbf{SY}$ by $\mathbf{U}^T\mathbf{Y}^T\mathbf{SYU} = \mathbf{D}^2$. □

Consequently, we have $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{U} = \mathbf{V}\mathbf{D}$. We can see that v_l (columns of $\hat{\mathbf{A}}$) and β_l (columns of $\hat{\mathbf{B}}$) differ by only a diagonal matrix $(\mathbf{I} - \mathbf{D}^2)^{-1}\mathbf{D}$.

Ex. 12.9

The reduced features are simply

$$\mathbf{X}^* = \mathbf{X}\hat{\mathbf{B}} = \mathbf{S}\mathbf{Y} \tag{12.36}$$

therefore the optimal scoring can be computed by

$$\max_{\boldsymbol{\Theta}} \boldsymbol{\Theta}^T\mathbf{Y}^T[\mathbf{S}\mathbf{Y}(\mathbf{Y}^T\mathbf{S}\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{S}^T]\mathbf{Y}\boldsymbol{\Theta} \tag{12.37}$$

$$\text{s.t. } \boldsymbol{\Theta}^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\Theta} = N\mathbf{I} \tag{12.38}$$

with trivial manipulations one can see that the objective function is exactly the same as optimal scoring on original features.

Ex. 12.10

The derivation for the general K-class GDA can be found in [?]. The kernel LDA (Fisher) is in the form of

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} \tag{12.39}$$

where

$$\mathbf{B} = (\bar{\mathbf{h}}_1 - \bar{\mathbf{h}}_2)(\bar{\mathbf{h}}_1 - \bar{\mathbf{h}}_2)^T \tag{12.40}$$

$$\bar{\mathbf{h}}_1 = \frac{1}{N_1}\sum_{i\in\mathcal{C}_1}\mathbf{h}(x_i) \tag{12.41}$$

$$\bar{\mathbf{h}}_2 = \frac{1}{N_2}\sum_{i\in\mathcal{C}_2}\mathbf{h}(x_i) \tag{12.42}$$

$$\tag{12.43}$$

(up to constant) and $N_1 = |\mathcal{C}_1|$, $N_2 = |\mathcal{C}_2|$ are the numbers of data poitns in class 1, 2, respectively. The discriminant vector $\mathbf{a}$ is a linear combination

$$\mathbf{a} = \sum_{i=1}^{N}\alpha_i\mathbf{h}(x_i) \tag{12.44}$$

therefore

$$\mathbf{a}^T\mathbf{B}\mathbf{a} = \boldsymbol{\alpha}^T(\mathbf{k}_1 - \mathbf{k}_2)(\mathbf{k}_1 - \mathbf{k}_2)^T\boldsymbol{\alpha} \tag{12.45}$$

where

$$\{\mathbf{k}_1\}_i = \frac{1}{N_1}\sum_{j\in\mathcal{C}_1} K_{ij},\ \{\mathbf{k}_2\}_i = \frac{1}{N_2}\sum_{j\in\mathcal{C}_2} K_{ij},\ i = 1,\ldots,N \tag{12.46}$$

On the other hand, $\mathbf{W} = \mathbf{W}_h + \gamma\mathbf{I}$, where

$$\mathbf{W}_h = \sum_{i\in\mathcal{C}_1}\left(\mathbf{h}(x_i)\mathbf{h}(x_i)^T - \bar{\mathbf{h}}_1\bar{\mathbf{h}}_1^T\right) + \sum_{i\in\mathcal{C}_2}\left(\mathbf{h}(x_i)\mathbf{h}(x_i)^T - \bar{\mathbf{h}}_2\bar{\mathbf{h}}_2^T\right) \tag{12.47}$$

(up to constant) therefore

$$\mathbf{a}^T\mathbf{W}_h\mathbf{a} = \boldsymbol{\alpha}^T\mathbf{K}^2\boldsymbol{\alpha} - N_1\boldsymbol{\alpha}^T\mathbf{k}_1\mathbf{k}_1^T\boldsymbol{\alpha} - N_2\boldsymbol{\alpha}^T\mathbf{k}_2\mathbf{k}_2^T\boldsymbol{\alpha} \tag{12.48}$$

and $\mathbf{a}^T\mathbf{a} = \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$. Consequently, the model depend on $h(\cdot)$ only via the N-by-N matrix $\mathbf{K}$.

Ex. 12.11

(a)

$$\begin{aligned}
P(X = x|G = k) &= \frac{P(X = x, G = k)}{\int_x P(X = x, G = k)dx} \\
&= \frac{\sum_{r=1}^R \pi_r P_r(G = k)\phi(x;\mu_r,\Sigma)}{\sum_{r=1}^R \pi_r P_r(G = k)\int_x \phi(x;\mu_r,\Sigma)dx} \\
&= \frac{\sum_{r=1}^R \pi_r P_r(G = k)\phi(x;\mu_r,\Sigma)}{\sum_{r=1}^R \pi_r P_r(G = k)}
\end{aligned} \tag{12.49}$$

Compare with (12.59), we can see that, by setting

$$\pi_{kr} = \frac{\sum_{r=1}^R \pi_r P_r(G = k)}{\sum_{r=1}^R \pi_r P_r(G = k)},\ R_k = R,\ \mu_{kr} = \mu_r \tag{12.50}$$

MDA2 is a generalization of MDA.

(b) E-step: compute the responsibility of subclass c_{kr} within class k for each class-k observation ($g_i = k$):

$$W(c_{kr}|x_i, g_i) = \frac{\pi_{kr}\phi(x_i;\mu_r,\Sigma)}{\sum_{r=1}^R \pi_{kr}\phi(x_i;\mu_r,\Sigma)} \tag{12.51}$$

M-step: MLE on μ_r and Σ

$$\mu_r = \frac{\sum_{k=1}^K \sum_{i:g_i=k} W(c_{kr}|x_i, g_i)x_i}{\sum_{k=1}^K \sum_{i:g_i=k} W(c_{kr}|x_i, g_i)} \tag{12.52}$$

$$\Sigma = \frac{\sum_{k=1}^K \sum_{i:g_i=k} \sum_{r=1}^R W(c_{kr}|x_i, g_i)(x_i - \mu_r)(x_i - \mu_r)^T}{\sum_{k=1}^K \sum_{i:g_i=k} \sum_{r=1}^R W(c_{kr}|x_i, g_i)} \tag{12.53}$$

(c) ???

Chapter 13

Prototype Methods and Nearest-Neighbors

Ex. 13.1

Again $k = 1, \dots, K$ are the indices of clusters/classes, $r = 1, \dots, R$ are the indices of cluster centers/Gaussian components.

For each precictor labeled to class k, namely $\{x_i | g_i = k\}$, in terms of the E-step,

- For EM algorithm, its responsibility to component r is evaluated as

$$\gamma_{kr}(x_i) = \frac{\pi_{kr}\phi(x_i; \mu_{kr}, \Sigma)}{\sum_{s=1}^{R} \pi_{kr}\phi(x_i; \mu_{ks}, \Sigma)} \tag{13.1}$$

- For k-means algorithm, each predictor belongs to exactly 1 cluster center, therefore its counterpart of responsibility is binary-valued:

$$\gamma_{kr}(x_i) = \begin{cases} 1 & \text{if } \|x_i - \mu_{kr}\| \leq \|x_i - \mu_{ks}\|, s = 1, \dots, R \\ 0 & \text{otherwise} \end{cases} \tag{13.2}$$

In terms of the M-step,

- For EM algorithm, the component means are updated as a weighted average

$$\mu_{kr} = \frac{\sum_{i:g_i=k} \gamma_{kr}(x_i) x_i}{\sum_{i:g_i=k} \gamma_{kr}} \tag{13.3}$$

  and the mixing probability is updated as

$$\pi_{kr} = \sum_{i:g_i=k} \frac{\gamma_{kr}(x_i)}{N_k} \tag{13.4}$$

  where N_k is the number of records labeld to class k.

- For k-means, the cluster center is taken as an unweighted average over all x_i closest to it. Using the binary-valued responsibility definition, μ_{kr} is exactly the same as Eq. (13.3).

To draw a connection between EM and k-means, as $\sigma \to 0$, $\phi(x_i; \mu_{kr}, \Sigma) \gg \phi(x_i; \mu_{ks}, \Sigma)$ if $\|x_i - \mu_{kr}\| < \|x_i - \mu_{ks}\|$, therefore the responsibility for EM approaches that of k-means. And π_{kr} becomes the proportion of points in $\{x_i | g_i = k\}$ that are closer to r than any other components centers.

Ex. 13.2

This problem is similar as Ex 2.3. Denote $P_0(r, N)$ as the probability of the following event: "Among the N i.i.d uniformly distributed points, there is none within the ball centered at 0 with radius of r." thus $P_0(r, N) = P_0(r)^N$, where $P_0(r) = P_0(r, 1)$. Since the points are uniformly distributed within the p-dim cube of edge length 1, $P_0(r, 1)$ simply equals to the ratio between the volume of spaces out of the r-ball but within the cube, and the volume of the cube. Consequently,

$$P_0(r, N) = (1 - v_p r^p)^N \tag{13.5}$$

The median R_{med} satisfy $P_0(R_{med}, N) = 1/2$, therefore

$$R_{med} = v_p^{-1/p}(1 - 2^{-1/N})^{1/p} \tag{13.6}$$

Ex. 13.3

Since $\sum_{k \neq k^*} p_k(x) = 1 - p_{k^*}(x)$

$$\begin{aligned}
\sum_{k=1}^{K} p_k(x)(1 - p_k(x)) &= p_{k^*}(x)(1 - p_{k^*}(x)) + \sum_{k \neq k^*} p_k(x)(1 - p_k(x)) \\
&= p_{k^*}(x)(1 - p_{k^*}(x)) + (1 - p_{k^*}(x)) - \sum_{k \neq k^*} p_k(x)^2 \\
&\leq p_{k^*}(x)(1 - p_{k^*}(x)) + (1 - p_{k^*}(x)) - \frac{(\sum_{k \neq k^*} p_k(x))^2}{K - 1} \\
& (1 - p_{k^*}(x)) - \frac{K}{K-1}(1 - p_{k^*}(x))^2
\end{aligned} \tag{13.7}$$

where we made use of Cauchy-Schwarz inequality.

Ex. 13.4

(a)

$$\begin{aligned}
\mathbf{A} &= \mathbf{QR} \\
&= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix}
\end{aligned} \tag{13.8}$$

where θ represents the angle of rotation, a, b represent the scale in x and y direction and λ represent the shear.

(b) Denote $\mathbf{J}$ as the Jacobian (1-by-2) of F at $c + x_0 + \mathbf{A}(\mathbf{x} - \mathbf{x}_0)$, we have

$$\frac{\partial F}{\partial \theta} = \mathbf{J} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \mathbf{A}(\mathbf{x} - \mathbf{x}_0) \tag{13.9a}$$

$$\frac{\partial F}{\partial a} = \mathbf{J} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \begin{bmatrix} 1 & \lambda \end{bmatrix} (\mathbf{x} - \mathbf{x}_0) \tag{13.9b}$$

$$\frac{\partial F}{\partial b} = \mathbf{J} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} (\mathbf{x} - \mathbf{x}_0) \tag{13.9c}$$

$$\frac{\partial F}{\partial \lambda} = \mathbf{J} \begin{bmatrix} a\cos\theta \\ a\sin\theta \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} (\mathbf{x} - \mathbf{x}_0) \tag{13.9d}$$

(c) It seems that the key to this procedure is to evaluate $\mathbf{J}$ given a coordinate $\mathbf{x}$. Denote the 2-D kernel smoother as $K(\mathbf{u}, \mathbf{v})$, then we solve the locally weighted regresion at $\mathbf{x}$:

$$\min_{\alpha(\mathbf{x}), \boldsymbol{\beta}(\mathbf{x})} \sum_{i=1}^{256} K(\mathbf{x}, \mathbf{x}_i)[F(\mathbf{x}_i) - \alpha(\mathbf{x}) - \boldsymbol{\beta}(\mathbf{x})^T \mathbf{x}_i] \tag{13.10}$$

Then we can use use $\boldsymbol{\beta}(\mathbf{x})^T$ as $\mathbf{J}(\mathbf{x})$ to compute the tangent space.

Ex. 13.5
Since

$$N \operatorname{tr}\left(\bar{\mathbf{B}}\bar{\mathbf{B}}^T\right) = \sum_{i=1}^{N} \operatorname{tr}\left(\mathbf{B}_i \bar{\mathbf{B}}^T\right) = \sum_{i=1}^{N} \operatorname{tr}\left(\bar{\mathbf{B}}\mathbf{B}_i^T\right) \tag{13.11a}$$

$$N \operatorname{tr}\left(\bar{\mathbf{B}}\mathbf{M}^T\right) = N \operatorname{tr}\left(\mathbf{M}\bar{\mathbf{B}}^T\right) = \sum_{i=1}^{N} \operatorname{tr}\left(\mathbf{B}_i \mathbf{M}^T\right) = \sum_{i=1}^{N} \operatorname{tr}\left(\mathbf{M}\mathbf{B}_i^T\right), \tag{13.11b}$$

it is easy to show that

$$\sum_{i=1}^{N} \operatorname{tr}\left[(\mathbf{B}_i - \mathbf{M})^2\right] = \sum_{i=1}^{N} \operatorname{tr}\left[(\mathbf{B}_i - \bar{\mathbf{B}})^2\right] + N \operatorname{tr}\left[(\mathbf{M} - \bar{\mathbf{B}})^2\right] \tag{13.12}$$

Therefore the rank-L approximation of $\mathbf{B}_i$ is equivalent to the rank-L approximation of $\bar{\mathbf{B}}$, namely $\bar{\mathbf{B}}_{[L]}$.

Ex. 13.6
($\mathbf{L}_j, j = 1, \ldots, M$ are the coordinates of the "black" parts of the cursive letter.) As the

optimal $\mathbf{A}_j = \mathbf{V}^T\mathbf{L}_j$, we have

$$\begin{aligned}
\sum_{j=1}^{M} \min_{\mathbf{A}_j} \|\mathbf{L}_j - \mathbf{V}\mathbf{A}_j\|^2 &= \sum_{j=1}^{M} \|(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{L}_j\|^2 \\
&= \sum_{j=1}^{M} \operatorname{tr}\left[(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{L}_j\mathbf{L}_j^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\right] \\
&= \sum_{j=1}^{M} \operatorname{tr}\left[(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{L}_j\mathbf{L}_j^T\right] \\
&= \sum_{j=1}^{M} \operatorname{tr}\left[\mathbf{U}_j\mathbf{\Sigma}_j^2\mathbf{U}_j^T\right] - \operatorname{tr}\left[\mathbf{V}^T\left(\sum_{j=1}^{M}\mathbf{U}_j\mathbf{\Sigma}_j^2\mathbf{U}_j^T\right)\mathbf{V}\right] \qquad (13.13)
\end{aligned}$$

where $\mathbf{L}_j = \mathbf{U}_j\mathbf{\Sigma}_j\mathbf{V}_j^T$ is the SVD of $\mathbf{L}_j$. Therefore $\mathbf{V}$ corresponds to the 2 largest eigenvectors of $\sum_{j=1}^{M}\mathbf{U}_j\mathbf{\Sigma}_j^2\mathbf{U}_j^T$.

For the alternative approach,

$$\begin{aligned}
\sum_{j=1}^{M} \min_{\mathbf{A}_j} \|\mathbf{L}_j\mathbf{A}_j^T - \mathbf{V}\|^2 &= \sum_{j=1}^{M} \|(\mathbf{I} - \mathbf{S}_j)\mathbf{V}\|^2 \\
&= \sum_{j=1}^{M} \operatorname{tr}\left[\mathbf{V}^T(\mathbf{I} - \mathbf{S}_j)(\mathbf{I} - \mathbf{S}_j)^T\mathbf{V}\right] \\
&= \sum_{j=1}^{M} \operatorname{tr}\left[\mathbf{V}^T(\mathbf{I} - \mathbf{S}_j)\mathbf{V}\right] \\
&= 2M - \operatorname{tr}\left[\mathbf{V}^T\left(\sum_{j=1}^{M}\mathbf{S}_j\right)\mathbf{V}\right] \qquad (13.14)
\end{aligned}$$

where $\mathbf{S}_j = \mathbf{L}_j(\mathbf{L}_j^T\mathbf{L}_j)^{-1}\mathbf{L}_j^T = \mathbf{U}_j\mathbf{U}_j^T$. Therefore $\mathbf{V}$ corresponds to the 2 largest eigenvectors of $\sum_{j=1}^{M}\mathbf{U}_j\mathbf{U}_j^T$.

Chapter 14

Unsupervised Learning

Chapter 15

Random Forests

Chapter 16

Ensemble Learning

Chapter 17

Undirected Graphical Models

Chapter 18

High-Dimensional Problems