Introduction to Machine Learning
Homework 5

Student Name:Lu Hongyu   Student ID: 2023202269

June 4, 2025

Question 1

Answer

Using least squares to fit one-hot label vectors amounts to solving

$$\min_W \| Y - XW \|_F^2, \quad W_{\text{LS}} = (X^\top X)^{-1} X^\top Y.$$

For a new feature vector x, one computes $\hat{y} = x^\top W_{\text{LS}}$ and predicts the class $\arg\max_j \hat{y}_j$. This *least-squares classification* generally only performs well under the following conditions:

- Each class-conditional density is a Gaussian with the *same* covariance matrix:

$$p(x \mid y = k) = \mathcal{N}(x; \mu_k, \Sigma), \quad \forall k, \text{ and all share } \Sigma.$$

- The class priors are equal (or approximately equal) so that each class has roughly the same number of training examples.

Under these assumptions, Linear Discriminant Analysis (LDA) produces linear decision boundaries that coincide with those from least-squares fitting of one-hot labels. Concretely, when $p(x \mid y = k) = \mathcal{N}(x; \mu_k, \Sigma)$ with Σ common and $\Pr(y = k) = 1/C$, the Bayes optimal decision rule reduces to

$$\text{decide } i \text{ if } x^\top \Sigma^{-1}(\mu_i - \mu_j) \ > \ \tfrac{1}{2}\left(\mu_i^\top \Sigma^{-1} \mu_i - \mu_j^\top \Sigma^{-1} \mu_j\right),$$

which is an affine hyperplane. In this scenario, least squares on one-hot labels recovers the same boundary as LDA. Therefore, *when each class is generated by a homoscedastic Gaussian and priors are equal, regression on one-hot vectors yields the correct classifier.* Otherwise, if class covariances differ or priors are highly imbalanced, least-squares classification may give suboptimal decision boundaries.

Question 2

Answer

(1) Modified dual formulation (Equation (15)):

$$\begin{aligned} L(w, b, \xi, c, \beta) &= \frac{1}{2}\|w\|_2^2 \ + \ \frac{1}{2N\lambda}\sum_{n=1}^{N} \xi_n \ + \ \sum_{n=1}^{N} c_n \left[1 - \xi_n - y_n(w^\top x_n - b) \right] \\ & c_n \geq 0, \ \ n = 1, \ldots, N. \end{aligned} \tag{15}$$

(2) Stationarity conditions for the inner minimization over w, b, ξ:

- w**:**

$$\frac{\partial L}{\partial w} = w \; - \; \sum_{n=1}^{N} c_n \, y_n \, x_n = 0 \;\; \Longrightarrow \;\; w = \sum_{n=1}^{N} c_n \, y_n \, x_n. \tag{16a}$$

- b**:**

$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} c_n \, y_n = 0 \;\; \Longrightarrow \;\; \sum_{n=1}^{N} c_n \, y_n = 0. \tag{16b}$$

- ξ_n**:**

$$\frac{\partial L}{\partial \xi_n} = \frac{1}{2N\lambda} \; - \; c_n \geq 0$$

(3) Substituting w and b back into the Lagrangian and dropping ξ_n-terms: From (16a) and (16b) we know

$$w = \sum_{n=1}^{N} c_n \, y_n \, x_n, \quad \sum_{n=1}^{N} c_n \, y_n = 0, \quad 0 \leq c_n \leq \frac{1}{2N\lambda}.$$

In $L(w, b, \xi, c, \beta)$, split into two parts:

$$L = \underbrace{\frac{1}{2}\|w\|_2^2 \; + \; \sum_{n=1}^{N} c_n \big[\, 1 - y_n(w^\mathsf{T} x_n - b) \big]}_{(A)} \; + \; \underbrace{\sum_{n=1}^{N} \big(\tfrac{1}{2N\lambda} - c_n \big) \, \xi_n}_{(B)} \, .$$

Since part (B) could be minimized to 0, it can be ignored. Thus only part (A) contributes:

$$\begin{aligned}
(A) &= \frac{1}{2} \Big\| \sum_{n=1}^{N} c_n \, y_n \, x_n \Big\|_2^2 + \sum_{n=1}^{N} c_n \; - \; \sum_{n=1}^{N} c_n \, y_n \Big(\sum_{m=1}^{N} c_m \, y_m \, x_m \Big)^\mathsf{T} x_n + b \sum_{n=1}^{N} c_n \, y_n \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} c_n \, c_m \, y_n \, y_m \, x_n^\mathsf{T} x_m + \sum_{n=1}^{N} c_n - \sum_{n=1}^{N} \sum_{m=1}^{N} c_n \, c_m \, y_n \, y_m \, x_n^\mathsf{T} x_m + b \cdot 0 \\
&= \sum_{n=1}^{N} c_n \; - \; \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} c_n \, c_m \, y_n \, y_m \, x_n^\mathsf{T} x_m,
\end{aligned}$$

Therefore, the dual becomes

$$\max_{\{c_n\}} \; \sum_{n=1}^{N} c_n \; - \; \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n \, c_n \, (x_n^\mathsf{T} x_m) \, y_m \, c_m,$$

$$\text{s.t.} \quad \sum_{n=1}^{N} c_n \, y_n = 0, \quad 0 \leq c_n \leq \frac{1}{2N\lambda}, \;\; n = 1, \ldots, N.$$

This matches Equation (17) in the slides. In particular, all ξ_n-related terms vanish because $\xi_n^\star = 0$, showing that they are indeed ignorable in the final dual formulation.