Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2019

Written Assignment 2

Issued: Wednesday 16$^{\text{th}}$ October, 2019 **Due:** Wednesday 30$^{\text{th}}$ October, 2019

2.1. (2 points) Define the **design matrix** $\boldsymbol{X}$ to be the m-by-n matrix that the training examples input values in its rows. Geometrically the solution of least-squares could be interpreted of as a vector in an M-dimensional space whose coordinates are $\boldsymbol{y} = \left[y^{(1)}, y^{(2)}, ..., y^{(m)}\right]^{\mathrm{T}}$. The least-squares regression function is obtained by finding the orthogonal projection of the target vector $\boldsymbol{y}$ onto the subspace spanned by the column vectors of $\boldsymbol{X}$, in which the i-th column vector is denoted as $\boldsymbol{X}_i$.

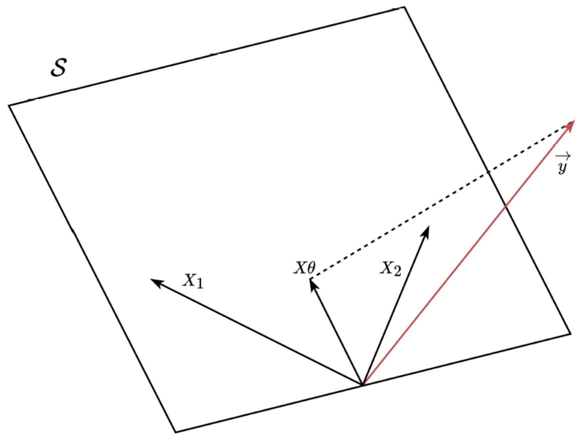

Figure 1: Projection of $\boldsymbol{y}$ on column space of $\boldsymbol{X}$

As shown in Figure 1, please show that the matrix

$$\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$$

takes any vector $\boldsymbol{v}$ and projects it onto the space spanned by the columns of $\boldsymbol{X}$. Use this result to show that the least-squares solution $\boldsymbol{\theta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ correspond to an orthogonal projection of the vector $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

2.2. (2 points) Suppose we are given a dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)})\colon i = 1, 2, \ldots, m\}$ consisting of m independent examples, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^n$ are n-dimension vector, and $y^{(i)} \in \{1, 2, \ldots, k\}$.We will model the joint distribution of $(\boldsymbol{x}, y)$ according to:

$$\begin{aligned} y^{(i)} &\sim \text{Multinomial}(\phi) \\ \boldsymbol{x}^{(i)}|y^{(i)} = j &\sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \end{aligned}$$

where the parameter ϕ_j gives $p(y^{(i)} = j)$ for each $j \in \{1, 2, \ldots, k\}$.
In Gaussian Discriminant Analysis (GDA), Linear Discriminant Analysis (LDA) just assume that the classes have a common covariance matrix $\Sigma_j = \Sigma, \forall j$. If the Σ_j are not assumed to be equal,we get Quadratic Discriminant Analysis (QDA). The estimates for QDA are similar to those for LDA, except that separate covariance matrices must be estimated for each class. Give the maximum likelihood estimate of Σ_j in the case that $k = 2$.

2.3. Suppose the data are linearly separable. The optimization problem of SVM is

$$\begin{aligned} \underset{\boldsymbol{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1, \quad i = 1, \ldots, l, \end{aligned} \tag{P}$$

and let $(\boldsymbol{w}^\star, b^\star)$ denote its optimal solution.

(a) (2 points) Show that

$$b^\star = -\frac{1}{2}\left(\max_{i:\, y_i=-1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + \min_{i:\, y_i=1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i\right).$$

The corresponding Lagrange dual problem is given by

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \\ \text{subject to} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, l, \\ & \sum_{i=1}^{l}\alpha_i y_i = 0. \end{aligned} \tag{D}$$

Suppose the optimal solution of (D) is $\boldsymbol{\alpha}^\star = (\alpha_1^\star, \cdots, \alpha_l^\star)^{\mathrm{T}}$, from the KKT conditions we know that

$$\begin{aligned} \boldsymbol{w}^\star &= \sum_{i=1}^{l}\alpha_i^\star y_i \boldsymbol{x}_i, \\ \sum_{i=1}^{l}\alpha_i^\star \left[y_i(\boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + b^\star) - 1\right] &= 0. \end{aligned} \tag{1}$$

(b) (1 point) Based on (1), verify that

$$\frac{1}{2}\|\boldsymbol{w}^\star\|_2^2 = \sum_{i=1}^{l}\alpha_i^\star - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i^\star\alpha_j^\star y_i y_j\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle = \frac{1}{2}\sum_{i=1}^{l}\alpha_i^\star.$$

2.4. When the data are not linearly separable, consider the soft-margin SVM given by

$$\begin{aligned} \underset{\boldsymbol{w},b,\boldsymbol{\xi}}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \quad i = 1, \ldots, l, \\ & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l, \end{aligned} \tag{2}$$

where $C > 0$ is a fixed parameter.

(a) (1 point) Show that (2) is equivalent[1] to

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{l}\ell(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b), \tag{3}$$

where $\ell(\cdot,\cdot)$ is the hinge loss defined by $\ell(y, z) \triangleq \max\{1 - yz, 0\}$.

[1] Two optimization problems are called equivalent if from a solution of one, a solution of the other is readily found, and vice versa.

(b) (2 points) Show that the objective function of (3), denoted by $f(\boldsymbol{w}, b)$, is convex, i.e.,

$$f(\theta \boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2, \theta b_1 + (1-\theta) b_2) \leq \theta f(\boldsymbol{w}_1, b_1) + (1-\theta) f(\boldsymbol{w}_2, b_2)$$

for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^n, b_1, b_2 \in \mathbb{R}$, and $\theta \in [0,1]$.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.