Tsinghua-Berkeley Shenzhen Institute
Learning from Data
Fall 2019

Written Assignment 2

Issued: Wednesday 16th October, 2019 **Due:** Wednesday 30th 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 $\boldsymbol{X}_i$.

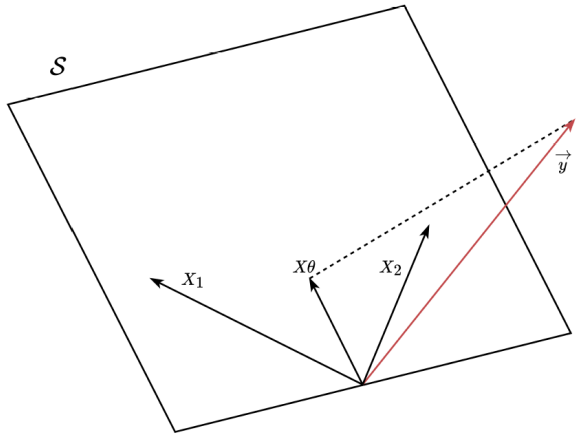

Figure 1: Projection of $\boldsymbol{y}$ on column space of $\boldsymbol{X}$

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 $\boldsymbol{X}$.

Solution: $\boldsymbol{X}$ can be written as $[\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n]$.
Therefore,

$$\boldsymbol{X\theta} = \sum_{i=1}^{n} \theta_i \boldsymbol{X}_i$$

It is on the column space of $\boldsymbol{X}$
Note that

$$\begin{aligned}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\theta} &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\end{aligned}$$

which means

$$\langle \boldsymbol{X}_i, \boldsymbol{X}\boldsymbol{\theta} \rangle = \langle \boldsymbol{X}_i, \boldsymbol{y} \rangle$$

Therefore,

$$\boldsymbol{X}\boldsymbol{\theta} = \sum_{i=1}^{n} \langle \boldsymbol{X}_i, \boldsymbol{X}\boldsymbol{\theta} \rangle \boldsymbol{X}_i = \sum_{i=1}^{n} \langle \boldsymbol{X}_i, \boldsymbol{y} \rangle \boldsymbol{X}_i$$

$\boldsymbol{X}\boldsymbol{\theta}$ is the orthogonal projection of the vector $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

2.2. (2 points) Suppose we are given a dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)}) : i = 1, 2, \ldots, m\}$ 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:

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

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 $k = 2$.

Solution:

$$\begin{aligned}
\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi) &= \log \prod_{i=1}^{m} P(\boldsymbol{x}_i | y_i; \boldsymbol{\mu}_{y_i}, \boldsymbol{\Sigma}_{y_i}) P(y_i; \phi_i) \\
&= \log \textstyle\prod_{i=1}^{m} \prod_{j=1}^{k} \mathbf{1}\{y_i = j\} \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)\Big) P(y_i = k; \phi_k) \\
&= \textstyle\sum_{i=1}^{m} \sum_{j=1}^{k} \mathbf{1}\{y_i = j\} \Big(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}_j| + \log P(y_i = j; \phi_k)\Big)
\end{aligned}$$

Note that

$$\frac{\partial \log |\boldsymbol{\Sigma}|}{\partial \boldsymbol{\Sigma}} = \frac{1}{|\boldsymbol{\Sigma}|} \frac{\partial |\boldsymbol{\Sigma}|}{\partial \boldsymbol{\Sigma}} = \frac{|\boldsymbol{\Sigma}| \boldsymbol{\Sigma}^{-1}}{|\boldsymbol{\Sigma}|} = \boldsymbol{\Sigma}^{-1}$$
$$\frac{\partial \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{v}}{\partial \boldsymbol{\Sigma}} = -\boldsymbol{\Sigma}^{-1} \boldsymbol{v} \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1^{\mathrm{T}}}$$

The equation above may need some procedures to be proved.

$$\begin{aligned}
\left(\frac{\partial \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{v}}{\partial \boldsymbol{\Sigma}}\right)_{ij} &= \frac{\partial \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{v}}{\partial \boldsymbol{\Sigma}_{ij}} = \boldsymbol{v}^{\mathrm{T}} \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \boldsymbol{\Sigma}_{ij}} \boldsymbol{v} \\
&= -\boldsymbol{v}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\Sigma}_{ij}} \boldsymbol{\Sigma}^{-1} \boldsymbol{v} \\
&= -\boldsymbol{v}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1^{\mathrm{T}}} \frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\Sigma}_{ij}} \boldsymbol{\Sigma}^{-1} \boldsymbol{v} \\
&= -\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{v} \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1^{\mathrm{T}}}\right)_{ij}
\end{aligned}$$

Therefore, the derivative on likelyhood function is

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\Sigma}_j} &= \frac{1}{2}\sum_{i=1}^{m} \mathbf{1}\left\{y_i = j\right\}\left(-\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1^{\mathrm{T}}}\right) \\ &= \mathbf{0}\end{aligned}$$

The result is

$$\boldsymbol{\Sigma}_j = \frac{\sum_{j=1}^{m} \mathbf{1}\left\{y_i = j\right\}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}}{\sum_{j=1}^{m} \mathbf{1}\left\{y_i = j\right\}}$$

where $j = 1, 2$

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

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

Solution:

(a) *The key point is to show that $\gamma_1 = \gamma_2 = 1$ (γ is defined below). Geometrically, it means that we can always find samples lying on the two boundaries $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = \pm 1$ when the parameters are optimal.*

Since the data are linearly separable, we have $\|\boldsymbol{w}^\star\| < \infty$. Define γ_1, γ_2 as

$$\gamma_1 \triangleq \min_{i:\, y_i=1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + b^\star,$$
$$\gamma_2 \triangleq -\left(\max_{i:\, y_i=-1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + b^\star\right),$$

then it suffices to prove $\gamma_1 = \gamma_2 = 1$. Note that the constraints of (P) imply $\gamma_1 \geq 1, \gamma_2 \geq 1$. If $\gamma_1 + \gamma_2 > 2$, we can define $\boldsymbol{w}^\dagger, b^\dagger$ as

$$\hat{\boldsymbol{w}}^\dagger \triangleq \frac{2}{\gamma_1+\gamma_2}\boldsymbol{w}^\star, \quad b^\dagger \triangleq \frac{2}{\gamma_1+\gamma_2}\left(b^\star - \frac{\gamma_1-\gamma_2}{2}\right),$$

such that $(\boldsymbol{w}^\dagger, b^\dagger)$ is a feasible solution of (P) and $\|\boldsymbol{w}^\dagger\|_2^2 \leq \|\boldsymbol{w}^\star\|_2^2$, which contradicts to the optimality of $(\boldsymbol{w}^\star, b^\star)$.

(b) *For convex optimization problems, the KKT conditions are sufficient conditions for strong duality.*

From the first equality of (1), we have

$$\boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{w}^\star = \left(\sum_{i=1}^{l}\alpha_i^\star y_i \boldsymbol{x}_i\right)^{\mathrm{T}}\left(\sum_{j=1}^{l}\alpha_j^\star y_j \boldsymbol{x}_j\right) = \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i^\star\alpha_j^\star y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle$$

Moreover,

$$\boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{w}^\star = \boldsymbol{w}^{\star\mathrm{T}}\left(\sum_{i=1}^{l}\alpha_i^\star y_i \boldsymbol{x}_i\right) = \sum_{i=1}^{l}\alpha_i^\star y_i \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i = \sum_{i=1}^{l}\alpha_i^\star,$$

where the last equality has used the second equation of (1). As a result,

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

which immediately yields the result.

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 $\ell(y,z) \triangleq \max\{1-yz, 0\}$.

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

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

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

Solution:

(a) Consider the optimal value of $\boldsymbol{\xi}$ for given $(\boldsymbol{w}, b)$ in (2). The constraints are equivalent to

$$\xi_i \ge \max\{0, 1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)\} = \ell(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b), \quad \forall i = 1, \ldots, l.$$

Hence, we have

$$\frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{l}\xi_i \ge \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{l}\ell(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b),$$

where the equality holds if and only if $\xi_i = \ell(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)$ for all $i = 1, \ldots, l$. As a consequence, to minimize the objective function in (2), the optimal ξ_i shall be chosen as $\xi_i = \ell(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)$, then the optimization problem (2) becomes the optimization problem (3).

(b) *It can be shown that both $\|\boldsymbol{w}\|_2^2$ and $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b$ are convex functions of $(\boldsymbol{w}, b)$. Then the conclusion is obvious by noting that the nonnegative weighted sum and pointwise maximum are operations that preserve convexity. See Chapter 3.2 of [1] for more details.*

Since (3) is a convex optimization problem without constraints, it can be solved efficiently.

The function $\|\boldsymbol{w}\|_2^2$ is convex since

$$\theta\|\boldsymbol{w}_1\|_2^2 + (1-\theta)\|\boldsymbol{w}_2\|_2^2 - \|\theta\boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2\|_2^2 = \theta(1-\theta)\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 \ge 0.$$

Let $t^+ \triangleq \max\{0, t\}$, then the hinge loss $\ell(y, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)$ as a function of $(\boldsymbol{w}, b)$ is convex, since

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

$$
\begin{aligned}
&\ell(y, (\theta \boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2)^{\mathrm{T}}\boldsymbol{x} + \theta b_1 + (1-\theta)b_2) \\
&= \max\{0, 1 - y\left[(\theta \boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2)^{\mathrm{T}}\boldsymbol{x} + \theta b_1 + (1-\theta)b_2\right]\} \\
&= \max\{0, \theta[1 - y(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} + b_1)] + (1-\theta)[1 - y(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} + b_2)]\} \\
&\leq \max\left\{0, \left(\theta[1 - y(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} + b_1)]\right)^+ + \left((1-\theta)[1 - y(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} + b_2)]\right)^+\right\} \\
&= \left(\theta[1 - y(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} + b_1)]\right)^+ + \left((1-\theta)[1 - y(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} + b_2)]\right)^+ \\
&= \theta\left(1 - y(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} + b_1)\right)^+ + (1-\theta)\left(1 - y(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} + b_2)\right)^+ \\
&= \theta \cdot \ell(y, \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} + b_1) + (1-\theta) \cdot \ell(y, \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} + b_2),
\end{aligned}
$$

where the inequality follows from $t \leq t^+$, and the penultimate equality follows from the fact that $(\theta t)^+ = \theta \cdot t^+$ for $\theta \geq 0$.

Finally, the convexity of the objective function $f(\boldsymbol{w}, b)$ follows from the fact that $f(\boldsymbol{w}, b)$ is a nonnegative weighted sum of convex functions $\|\boldsymbol{w}\|_2^2$ and $\ell(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b), i = 1, \ldots, l$. (*Easily verified by definition*)

References

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