Tsinghua-Berkeley Shenzhen Institute
Learning from Data
Fall 2019

Writing Homework 2

TIAN Chenyu October 23, 2019

- **Acknowledgments:** This template takes some materials from course CSE 547/Stat 548 of Washington University: `https://courses.cs.washington.edu/courses/cse547/17sp/index.html`.

- **Collaborators:** I finish this homework by myself.

2.1. Define $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$, for a given vector $\boldsymbol{v}$

$$\boldsymbol{v} = \boldsymbol{P}\boldsymbol{v} + (\boldsymbol{v} - \boldsymbol{P}\boldsymbol{v})$$

If we can prove that $\boldsymbol{v} - \boldsymbol{P}\boldsymbol{v}$ is orthogonal to both $\boldsymbol{P}\boldsymbol{v}$ and the column space of $\boldsymbol{X}$, we can prove that matrix $\boldsymbol{P}$ project $\boldsymbol{v}$ onto column space of $\boldsymbol{X}$.
So this problem is equvalent to prove:

$$\begin{aligned}(\boldsymbol{P}\boldsymbol{v})^T(\boldsymbol{v} - \boldsymbol{P}\boldsymbol{v}) &= 0\\ \boldsymbol{X}^T(\boldsymbol{v} - \boldsymbol{P}\boldsymbol{v}) &= 0\end{aligned}$$

Proof:

$$\begin{aligned}(\boldsymbol{P}\boldsymbol{v})^T(\boldsymbol{v} - \boldsymbol{P}\boldsymbol{v}) &= (\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v})^T\boldsymbol{v}\\ &\quad - (\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v})^T(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v})\\ &= \boldsymbol{v}^T\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X})^{-1})^T\boldsymbol{X}^T\boldsymbol{v}\\ &\quad - \boldsymbol{v}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{v}\\ &= \boldsymbol{v}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}\\ &\quad - \boldsymbol{v}^T\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}\\ &= 0\end{aligned}$$

$$\begin{aligned}\boldsymbol{X}^T(\boldsymbol{v} - \boldsymbol{P}\boldsymbol{v}) &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{v} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{v}\\ &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{v} \ - \ \boldsymbol{X}^T\ \boldsymbol{X}\ (\boldsymbol{X}^T\ \boldsymbol{X})^{-1}\ \boldsymbol{X}^T\ \boldsymbol{v}\\ &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{v} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}\\ &= 0\end{aligned}$$

Thus, $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ project $\boldsymbol{v}$ onto column space of $\boldsymbol{X}$.

I think that $\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$ correspond to an orthogonal projection of the vector $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

2.2.

$$p(\boldsymbol{x}|y=0) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma_0}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_0})^T \boldsymbol{\Sigma_0}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_0})\right)$$
$$p(\boldsymbol{x}|y=1) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma_1}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})^T \boldsymbol{\Sigma_1}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})\right)$$

The log likelihood function of QDA is

$$\begin{aligned} l(\phi, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}) &= \log \prod_{i=1}^{m} p\left(\boldsymbol{x}^{(i)}, y^{(i)}; \phi, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}\right) \\ &= \log \prod_{i=1}^{m} p\left(\boldsymbol{x}^{(i)}|y^{(i)}; \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}\right) \phi_{y^{(i)}} \end{aligned}$$

For $\boldsymbol{\Sigma_0}$, we have

$$\begin{aligned} \frac{\partial l(\phi, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1})}{\partial \boldsymbol{\Sigma_0}} &= -\frac{\sum_{i=1}^{m} \mathbb{1}(y^i=0)}{2} \frac{\partial}{\partial \boldsymbol{\Sigma_0}} \log|\boldsymbol{\Sigma_0}| \\ &\quad - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma_0}} \sum_{i=1}^{m} \mathbb{1}(y^i=0) \left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_0}\right)^T \boldsymbol{\Sigma_0}^{-1} \left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_0}\right) \\ &= -\frac{\sum_{i=1}^{m} \mathbb{1}(y^i=0)}{2} \boldsymbol{\Sigma_0}^{-1} \\ &\quad + \frac{1}{2}\boldsymbol{\Sigma_0}^{-1}\left[\sum_{i=1}^{m} \mathbb{1}(y^i=0)\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_0}\right)\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_0}\right)^{\mathrm{T}}\right]\boldsymbol{\Sigma_0}^{-1} \\ &= O \end{aligned}$$

which yields that

$$\boldsymbol{\Sigma_0} = \frac{1}{\sum_{i=1}^{m} \mathbb{1}(y^i=0)} \sum_{i=1}^{m} \mathbb{1}(y^i=0)\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_0}\right)\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_0}\right)^{\mathrm{T}}$$

With same derivation

$$\boldsymbol{\Sigma_1} = \frac{1}{\sum_{i=1}^{m} \mathbb{1}(y^i=1)} \sum_{i=1}^{m} \mathbb{1}(y^i=1)\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_1}\right)\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu_1}\right)^{\mathrm{T}}$$

2.3. (a) Since the data is separable, there exist support vectors which $y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) = 1$.
When $y_i = 1$, $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \geq 1$, and $\min\limits_{i:y_i=1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + b^\star = 1$;
When $y_i = -1$, $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \leq -1$, and $\max\limits_{i:y_i=-1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + b^\star = -1$;
Therefore,

$$\begin{aligned} &\max_{i:y_i=-1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + b^\star + \min_{i:y_i=1} \boldsymbol{w}^{\star\mathrm{T}}\boldsymbol{x}_i + b^\star = 0 \\ \Rightarrow \quad & b^\star = -\frac{1}{2}\left(\max_{i:y_i=-1} \boldsymbol{w}^{\star\mathbf{T}}\boldsymbol{x}_i + \min_{i:y_i=1} \boldsymbol{w}^{\star\mathbf{T}}\boldsymbol{x}_i\right) \end{aligned}$$

(b) From the KKT condition, here exists:

$$
\begin{aligned}
& \sum_{i=1}^{l} \alpha_i^\star \left[y_i \left(\boldsymbol{w}^{\star\mathrm{T}} \boldsymbol{x}_i + b^\star \right) - 1 \right] = 0 \\
\Rightarrow \quad & \sum_{i=1}^{l} \alpha_i^\star y_i \boldsymbol{w}^{\star\mathrm{T}} \boldsymbol{x}_i + \sum_{i=1}^{l} \alpha_i^\star y_i b^\star = \sum_{i=1}^{l} \alpha_i^\star \\
\Rightarrow \quad & \sum_{i=1}^{l} \alpha_i^\star y_i \boldsymbol{w}^{\star\mathrm{T}} \boldsymbol{x}_i + b^\star \sum_{i=1}^{l} \alpha_i^\star y_i = \sum_{i=1}^{l} \alpha_i^\star \\
\Rightarrow \quad & \sum_{i=1}^{l} \alpha_i^\star y_i \boldsymbol{w}^{\star\mathrm{T}} \boldsymbol{x}_i = \sum_{i=1}^{l} \alpha_i^\star \\
\Rightarrow \quad & \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i^\star \alpha_j^\star y_i y_j \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle = \sum_{i=1}^{l} \alpha_i^\star
\end{aligned}
$$

Then, using the equation above, it has

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

2.4. (a) The original problem is

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

Then generate its Lagrange function

$$
L(\boldsymbol{w}, b, \lambda_i, \mu_i) = \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \lambda_i \xi_i - \sum_{i=1}^{l} \mu_i (y_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b \right) - 1 - \xi_i)
$$

The optimal solution satisfies KKT condition, with

$$
\begin{aligned}
& C - \lambda_i - \mu_i = 0 \\
& \xi_i \geq 0 \\
& \lambda_i \xi_i = 0 \\
& y_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b \right) \geq 1 - \xi_i \\
& \mu_i (y_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b \right) - 1 - \xi_i) = 0
\end{aligned}
$$

when find the optimal solution, here exist

$$
\min L(\boldsymbol{w}^*, b^*, \lambda_i^*, \mu_i^*) = \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^{l} \xi_i
$$

When $\mu_i = 0$, here exists $\lambda = C$ and $\xi_i = 0$, and
$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \leq 0$,
so $\ell\left(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$ =max{0,$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$}=0;
When $\mu_i \neq 0$, $\xi_i = 1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$, and
$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) = \xi_i \geq 0$,
so $\ell\left(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$ =max{0,$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$}=$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$.
Thus, the problem is equivalent to

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

(b) To prove a convex function

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

proof:

$$\begin{aligned}
&\|\theta\boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2\|_2^2 - \theta\|\boldsymbol{w}_1\|_2^2 - (1-\theta)\|\boldsymbol{w}_2\|_2^2 \\
=&\theta^2\|\boldsymbol{w}_1\|_2^2 + 2\theta(1-\theta)\boldsymbol{w}_1^T\boldsymbol{w}_2 + (1-\theta)^2\|\boldsymbol{w}_2\|_2^2 - \theta\|\boldsymbol{w}_1\|_2^2 - (1-\theta)\|\boldsymbol{w}_2\|_2^2 \\
=&2\theta(1-\theta)\boldsymbol{w}_1^T\boldsymbol{w}_2 - \theta(1-\theta)\|\boldsymbol{w}_1\|_2^2 - \theta(1-\theta)\|\boldsymbol{w}_2\|_2^2 \\
\leq&2\theta(1-\theta)\boldsymbol{w}_1^T\boldsymbol{w}_2 - \theta\|\boldsymbol{w}_1\|_2^2 - \theta\|\boldsymbol{w}_2\|_2^2 \\
\leq& - \theta\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 \\
\leq&0 \\
\Rightarrow \quad &\|\theta\boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2\|_2^2 \leq \theta\|\boldsymbol{w}_1\|_2^2 + (1-\theta)\|\boldsymbol{w}_2\|_2^2
\end{aligned}$$

So $\|\boldsymbol{w}\|_2^2$ is a convex function.

$$\begin{aligned}
&\ell\left(y_i, (\theta\boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2)^T\boldsymbol{x_i} + \theta b_1 + (1-\theta)b_2\right) \\
=&\max\{1 - y_i\left((\theta\boldsymbol{w}_1 + (1-\theta)\boldsymbol{w}_2)^T\boldsymbol{x_i} + \theta b_1 + (1-\theta)b_2\right), 0\} \\
\leq&\max\{\theta - y_i(\theta\boldsymbol{w}_1^T\boldsymbol{x_i} + \theta b_1) + (1-\theta) - y_i((1-\theta)\boldsymbol{w}_2^T\boldsymbol{x_i} + (1-\theta)b_2), 0\} \\
\leq&\max\{\theta - y_i(\theta\boldsymbol{w}_1^T\boldsymbol{x_i} + \theta b_1)), 0\} + \max\{(1-\theta) - y_i((1-\theta)\boldsymbol{w}_2^T\boldsymbol{x_i} + (1-\theta)b_2), 0\} \\
\leq&\theta\max\{1 - y_i(\boldsymbol{w}_1^T\boldsymbol{x_i} + b_1)), 0\} + (1-\theta)\max\{1 - y_i\boldsymbol{w}_2^T\boldsymbol{x_i} + b_2), 0\} \\
\leq&\theta\ell\left(y_i, \boldsymbol{w}_1^T\boldsymbol{x}_i + b_1\right) + (1-\theta)\ell\left(y_i, \boldsymbol{w}_2^T\boldsymbol{x}_i + b_2\right)
\end{aligned}$$

So $\ell(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)$ is a convex function.
The non-negative weighted sum of convex functions is still a convex function. And $C \geq 0$.
Thus the objective function
$f(\boldsymbol{\omega}, b) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{l}\ell\left(y_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$ is convex.