Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2019

Homework 1

TIAN Chenyu — September 30, 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 template by myself.

1.1. 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) 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) Based on (1), verify that

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

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

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

where $C > 0$ is a fixed parameter.

(a) 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) 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 $\theta \in [0,1]$.

1.3. You may find `https://en.wikibooks.org/wiki/LaTeX` useful.

(a) Writing LaTeX online may be easier for beginners:

i. ShareLaTeX: `https://www.sharelatex.com/`.

ii. Overleaf: `https://www.overleaf.com/`.

1.4. You may need aligned equations for your homework, here are several examples:

Total propability rule:

$$\begin{aligned} \mathbb{P}(\mathsf{x} = x) &= \sum_{y\in\mathcal{Y}} \mathbb{P}(\mathsf{x} = x, \mathsf{y} = y) \\ &= \sum_{y\in\mathcal{Y}} \mathbb{P}(\mathsf{x} = x|\mathsf{y} = y)\,\mathbb{P}(\mathsf{y} = y), \end{aligned}$$

or

$$\begin{aligned} & P_{\mathsf{x}}(x) \\ &= \sum_{y\in\mathcal{Y}} P_{\mathsf{xy}}(x,y) \\ &= \sum_{y\in\mathcal{Y}} P_{\mathsf{x}|\mathsf{y}}(x|y)P_{\mathsf{y}}(y). \end{aligned}$$

Indicator function:

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

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

1.5. You may need to add figure and source codes in your homework. Figure 1 is an example that compares the empirical distribution (histogram) and probability density function of the Gaussian random variable.

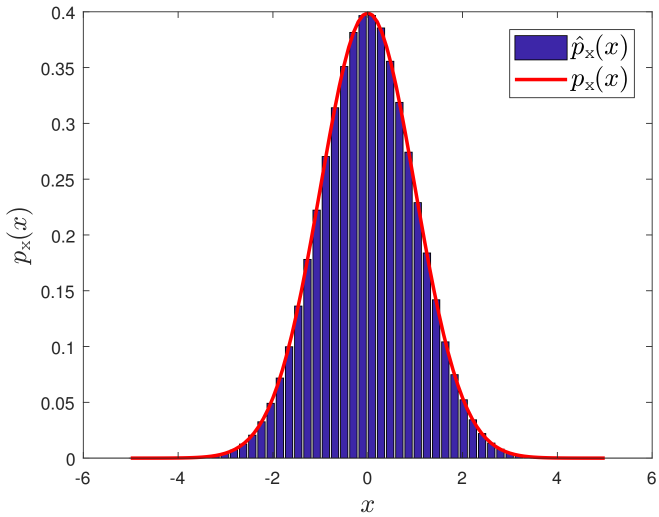

Figure 1: Gaussian PDF and histogram of samples

The source code to plot Figure 1 could be found in Appendix **??**. Here are the core codes:

```
4 [cnt, x_hist] = hist(data, nbins); % not to plot, only to get
      emperical distribution.
```

```
6 cnt = cnt / n / (x_hist(2) - x_hist(1)); % normalization, be
      careful :)
7 bar(x_hist, cnt); % plot the hist using bar()
```

To understand line 6, note that if we have n samples of x denoted by $x^{(i)}, i = 1, 2, \cdots, n$, then the probability density function p_{x} could be estimated as

$$
\begin{aligned}
p_{\mathsf{x}}(x_0) &= \frac{\mathrm{d}}{\mathrm{d}x} \mathbb{P}(\mathsf{x} \le x) \Big|_{x=x_0} \\
&\approx \frac{\mathbb{P}(x_0 - \Delta x < \mathsf{x} \le x_0)}{\Delta x} \\
&\approx \frac{1}{n\Delta x} \sum_{i=1}^{n} \mathbb{1}_{x^{(i)} \in (x_0 - \Delta x, x_0]} \,.
\end{aligned}
$$