Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2020

Writing Assignment 2

Issued: Saturday 17th October, 2020 **Due:** Friday 30th October, 2020

POLICIES

- **Acknowledgments:** We expect you to make an honest effort to solve the problems individually. As we sometimes reuse problem set questions from previous years, covered by papers and web pages, we expect the students **NOT** to copy, refer to, or look at the solutions in preparing their answers (relating to an unauthorized material is considered a violation of the honor principle). Similarly, we expect to not to google directly for answers (though you are free to google for knowledge about the topic). If you do happen to use other material, it must be acknowledged here, with a citation on the submitted solution.
- **Required homework submission format:** You can submit homework either as one single PDF document or as handwritten papers. Written homework needs to be provided during the class in the due date, and PDF document needs to be submitted through Tsinghua's Web Learning (`http://learn.tsinghua.edu.cn/`) before the end of due date.

  It is encouraged you LaTeX all your work, and we would provide a LaTeX template for your homework.
- **Collaborators:** In a separate section (before your answers), list the names of all people you collaborated with and for which question(s). If you did the HW entirely on your own, **PLEASE STATE THIS**. Each student must understand, write, and hand in answers of their own.

2.1. (2.5 points)(Naive Bayes Parameter Learning) Suppose we are given 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 with entry $\boldsymbol{x}_j \in \{0, 1\}$, and $y^{(i)} \in \{0, 1\}$. We will model the joint distribution of $(\boldsymbol{x}, y)$ according to:

$$\begin{aligned} y^{(i)} &\sim \mathrm{Bernoulli}(\phi_\mathrm{y}) \\ \boldsymbol{x}_j^{(i)} | y^{(i)} = b &\sim \mathrm{Bernoulli}(\phi_{\mathrm{j|y=b}}), \mathrm{b} = 0, 1 \end{aligned}$$

where the parameters $\phi_y \overset{\text{def}}{=} p(y = 1)$ and $\phi_{j|y=b} \overset{\text{def}}{=} p(\boldsymbol{x}_j = 1 | y^{(i)} = b)$. Under Naive Bayes (NB) assumption, the probability of observing $\boldsymbol{x}_j | y = b, j = 1, \ldots, n$ are independent which means $p(x_1, \cdots, x_n | y) = \Pi_{j=1}^n p(x_j | y)$. Calculate the maximum likelihood estimation of those parameters.

2.2. (2.5 points)(Quadratic Discriminant Analysis) Suppose we are given a dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)}) \colon i = 1, 2, \dots, m\}$ consisting of m independent examples, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^n$ are n-dimension vector, and $y^{(i)} \in \{1, 2, \dots, k\}$. We will model the joint distribution of $(\boldsymbol{x}, y)$ according to:

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

where the parameter ϕ_j gives $p(y^{(i)} = j)$ for each $j \in \{1, 2, \dots, k\}$.

In Gaussian Discriminant Analysis (GDA), Linear Discriminant Analysis (LDA) just assume that the classes have a common covariance matrix $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}, \forall j$. If the $\boldsymbol{\Sigma}_j$ 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 $\boldsymbol{\Sigma}_j$ in the case that $k = 2$.

2.3. (Soft-SVM) When the data are not linearly separable, consider the soft-margin SVM given by

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

where $C > 0$ is a fixed parameter.

(a) (1 point) Show that (3) is equivalent[1] to

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

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

(b) (Bonus question, 2 points) When we do optimization problem, the first thing to consider is whether the optimal point exist and unique. Generally speaking, convex optimization has been well studied and possess good properties. Show that the objective function of (2), 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 $\theta \in [0, 1]$.

2.4. (Kernel-SVM) When the data are not linearly separable, consider the Kernel-SVM given by

$$\begin{aligned} \underset{\boldsymbol{w}, b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b) \geq 1, \quad i = 1, \dots, l, \end{aligned} \tag{3}$$

where $\phi(\boldsymbol{x})$ is a mapping function $\phi(\boldsymbol{x}) : (x_1, x_2) \mapsto \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$.

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

(a) (1 point) Prove that $\Phi(\boldsymbol{x}, \boldsymbol{y}) \overset{\text{def}}{=} \phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{y})$ is positive definite symmetric.

(b) (2 points) Given data set $\big\{((1, \sqrt{2})^{\mathrm{T}}, 1), ((\sqrt{2}, 1)^{\mathrm{T}}, 1), ((2, \sqrt{2})^{\mathrm{T}}, -1)\big\}$, derive the optimal value of $\boldsymbol{w}^*$ and b^* in (3).

(c) (1 point) In (b), for new sample $((4\sqrt{2}, 1)^{\mathrm{T}}, 1)$, make your decision of classification.