Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2020

Writing Assignment 2

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

2.1. (2.5 points)(Naive Bayes Parameter Learning) Suppose we are given dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)}), i = 1, 2, \dots, m\}$ 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:

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

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 $\boldsymbol{x}_j | y = b, j = 1, \dots, n$ 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.

Solution: Based on the Naive Bayes assumption, the log-likelihood function

$$\begin{aligned}
\ell\left(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}\right) &= \sum_{i=1}^m \log(p(\boldsymbol{x}^{(i)}, y^{(i)})) \\
&= \sum_{i=1}^m \log(p(\boldsymbol{x}^{(i)} | y^{(i)})) + \log(p(y^{(i)})) \\
&= \sum_{i=1}^m \sum_{j=1}^n \log(p(\boldsymbol{x}_j^{(i)} | y^{(i)})) + \sum_{i=1}^m \log(p(y^{(i)})) \\
&= \sum_{i=1}^m \sum_{j=1}^n (\mathbf{1}\{y^{(i)} = 0\} + \mathbf{1}\{y^{(i)} = 1\}) \log(p(\boldsymbol{x}_j^{(i)} | y^{(i)})) \\
&\quad + \sum_{i=1}^m y^{(i)} \log(\phi_y) + (1 - y^{(i)}) \log(1 - \phi_y) \\
&= \sum_{i=1}^m \sum_{j=1}^n \sum_{b=0}^1 \mathbf{1}\{y^{(i)} = b\}(\boldsymbol{x}_j^{(i)} \log(\phi_{j|y=b}) + (1 - \boldsymbol{x}_j^{(i)}) \log(1 - \phi_{j|y=b})) \\
&\quad + \sum_{i=1}^m y^{(i)} \log(\phi_y) + (1 - y^{(i)}) \log(1 - \phi_y).
\end{aligned}$$

Taking derivatives, we have

$$\frac{\partial \ell}{\partial \phi_y} = \frac{\sum\limits_{i=1}^m y^{(i)}}{\phi_y} - \frac{\sum\limits_{i=1}^m (1 - y^{(i)})}{1 - \phi_y} = 0$$

$$\frac{\partial \ell}{\partial \phi_{j|y=b}} = \frac{\sum\limits_{i=1}^m \mathbf{1}\{x_j^{(i)} = 1, y^i = b\}}{\phi_{j|y=b}} - \frac{\sum\limits_{i=1}^m \mathbf{1}\{y^{(i)} = b\} - \mathbf{1}\{x_j^{(i)} = 1, y^i = b\}}{1 - \phi_{j|y=b}} = 0.$$

Thus,

$$\phi_y^* = \frac{\sum_{i=1}^m \mathbf{1}\{y^i = 1\}}{m},$$
$$\phi_{j|y=b}^* = \frac{\sum_{i=1}^m \mathbf{1}\{x_j^{(i)} = 1, y^i = b\}}{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = b\}}, b = 0, 1.$$

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:

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

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

Solution:

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

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.

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

Therefore, the derivative on likelyhood function is

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

The result is

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

where $j = 1, 2$

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

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

where $C > 0$ is a fixed parameter.

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

Solution:

(a) Consider the optimal value of $\boldsymbol{\xi}$ for given $(\boldsymbol{w}, b)$ in (1). The constraints are equivalent to

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

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

Hence, we have

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

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 (1), 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 (1) becomes the optimization problem (2).

(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.*

Since (2) is a convex optimization problem without constraints, it can be solved efficiently.

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

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

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

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

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*)

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

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

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

(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 $\left\{((1, \sqrt{2})^{\mathrm{T}}, 1), ((\sqrt{2}, 1)^{\mathrm{T}}, 1), ((2, \sqrt{2})^{\mathrm{T}}, -1)\right\}$, derive the optimal value of $\boldsymbol{w}^*$ and b^* in (3).

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

Solution:

(a) Actually, a kernel function is said to be positive definite symmetric if for any $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$, the kernel matrix $\boldsymbol{K} = [\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{ij} \in \mathbb{R}^n$ is symmetric positive semidefinite.

For any vector $\boldsymbol{v} \in \mathbb{R}^l$, we have

$$\begin{aligned}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{v} &= \boldsymbol{v}^{\mathrm{T}}(\phi(\boldsymbol{x}^{(1)}), \cdots, \phi(\boldsymbol{x}^{(l)}))(\phi(\boldsymbol{x}^{(1)}), \cdots, \phi(\boldsymbol{x}^{(l)}))^{\mathrm{T}}\boldsymbol{v} \\ &= ((\phi(\boldsymbol{x}^{(1)}), \cdots, \phi(\boldsymbol{x}^{(l)}))^{\mathrm{T}}\boldsymbol{v})^{\mathrm{T}}(\phi(\boldsymbol{x}^{(1)}), \cdots, \phi(\boldsymbol{x}^{(l)}))^{\mathrm{T}}\boldsymbol{v} \\ &\geq 0\end{aligned}$$

Thus it is a symmetric definite kernel function.

(b) We form the Lagrangian function as

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||_2^2 - \sum_{i=1}^{l}\alpha_i[y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b) - 1]$$

Letting its partial derivatives with respect to $\boldsymbol{w}$ and b be zero, we have

$$\begin{aligned}\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} &= 0 \longrightarrow \boldsymbol{w} = \sum_{i=1}^{l}\alpha_i y_i \phi(x_i) \\ \frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial b} &= 0 \longrightarrow \sum_{i=1}^{l}\alpha_i y_i = 0\end{aligned}$$

Eliminating the primal decision variables $\boldsymbol{w}$ and b, we have the objective of the lagrange dual problem as

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j \alpha_i \alpha_j [\phi(\boldsymbol{x}_i)]^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$

The whole dual problem can then be written as

$$\begin{aligned}&\max_{\boldsymbol{\alpha}} \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j \alpha_i \alpha_j [\phi(\boldsymbol{x}_i)]^{\mathrm{T}}\phi(\boldsymbol{x}_j) \\ &\text{subject to} \quad \alpha_i \geq 0, \quad i = 1, ...l \\ &\qquad\qquad \sum_{i=1}^{l}\alpha_i y_i = 0\end{aligned}$$

Implement the data within it, we have

$$\begin{aligned}&\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) \overset{\text{def}}{=} 2\alpha_1 + 2\alpha_2 - \frac{13}{2}\alpha_1^2 - 10\alpha_1\alpha_2 - \frac{9}{2}\alpha_2^2 \\ &\text{subject to} \quad \alpha_i \geq 0, \quad i = 1, 2\end{aligned}$$

The optimal $\alpha_1^* = 0, \alpha_2^* = \frac{2}{9}, \alpha_3^* = \frac{2}{9}$. Thus the optimal

$$\begin{aligned}\boldsymbol{w}^* &= \sum_{i=1}^{3} \alpha_i y_i \phi(\boldsymbol{x}^i) = (-\frac{4}{9}, -\frac{4}{9}, -\frac{2}{9})^{\mathrm{T}} \\ b^* &= \frac{1}{|S|} \sum_{i \in S} [y_i - (\boldsymbol{w}^*)^{\mathrm{T}} \phi(\boldsymbol{x}_i)] = 3,\end{aligned}$$

in which S denotes the set of indices of all support vectors and $|S|$ is the cardinality of S.

(c) $sign[(\boldsymbol{w}^*)^{\mathrm{T}} \phi(\boldsymbol{z}) + b^*] = sign(-15) = -1$