Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Spring 2024

Writing Assignment 2

Issued: Monday 1st April, 2024 **Due:** Monday 15th April, 2024

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 you to not 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 on the due date, and a PDF document needs to be submitted through Tsinghua's Web Learning (`http://learn.tsinghua.edu.cn/`) before the end of the due date.
- **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. **SVM and logistic regression** (4 points)

Support Vector Machine (SVM) is a powerful and effective supervised machine learning algorithm. Given m samples $\left(\boldsymbol{x}^{(1)}, y^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(m)}, y^{(m)}\right), \boldsymbol{x}^{(i)} \in \mathbb{R}^n, y^{(i)} \in \mathbb{R}, i = 1, \ldots, m$, we have learnt that the optimal parameters ($\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$) can be derived by solving the optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \\ &\text{s.t. } \; y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned} \tag{1}$$

The constraints in (1) indicates a hard punishment of incorrect classification. As an alternative form, the optimization problem above can be re-written into the minimization of the following function

$$\sum_{i=1}^{m} E_\infty(y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b)) + \lambda||\boldsymbol{w}||^2.$$

(a) (0.5 + 0.5 points) Give the definition of function $E_\infty(\cdot)$ and the constraint for the regularization parameter λ.

(b) (1 point) Consider the logistic regression model with a target variable $y \in \{-1, 1\}$, and we have $p(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b)$, where $\sigma(\cdot)$ is the Sigmoid function. Show that the negative log-likelihood, with the addition of a quadratic regularizer, take the form

$$\sum_{i=1}^{m} E_{LR}(y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b)) + \lambda||\boldsymbol{w}||^2,$$

and give the definition of function $E_{LR}(\cdot)$.

(c) (Bonus 1 points) In real-world applications, there might exist overlap between the class-conditional distributions, making an exact separation of training data unfeasible and inadequate. To avoid such limitation, SVM is modified to allow for some training points to be misclassified. Specifically, we introduce slack variables $\xi^{(i)} \geq 0$, such that the constraints in (1) are replaced with

$$y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) \geq 1 - \xi^{(i)}, \quad i = 1, \ldots, m,$$

and we therefore minimize

$$C \sum_{i=1}^{m} \xi^{(i)} + \frac{1}{2}||\boldsymbol{w}||^2, \tag{2}$$

where the parameter $C > 0$. Show that (2) can also be written in the form

$$\sum_{i=1}^{m} E_{SV}(y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b)) + \lambda||\boldsymbol{w}||^2,$$

and give the definition of function $E_{SV}(\cdot)$ and regularization parameter λ.
Hint: you may need to discuss the relationship of $y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b)$ and $\xi^{(i)}$. A possible way is to write down the Lagrangian for soft SVM and use its KKT conditions.

(d) (Bonus 1 points) Plot the error functions $E_\infty(\cdot)$, $E_{LR}(\cdot)$ and $E_{SV}(\cdot)$ in one graph [1]. Conclude your findings and discuss what may happen if we replace the error function with other functions.

2.2. **Poisson regression and the exponential family** (3 points)

(a) (1 point) Consider the Poisson distribution parameterized by λ:

$$p(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

Show that the Poisson distribution is in the exponential family, and clearly state what are $b(y), \eta, T(y)$, and $a(\eta)$.

[1] Function input as x-axis and output as y-axis. You may use different colors or line styles to represent different functions.

(b) (1 point) Consider performing regression using a GLM model with a Poisson response variable. What is the canonical response function for the family? (You may use the fact that a Poisson random variable with a parameter λ has mean λ.)

(c) (1 point) For a training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, let the log-likelihood of an example $\log p(y^{(i)}|x^{(i)}; \theta)$. By taking the derivative of the log-likelihood with respect to θ_j, derive the stochastic gradient ascent rule for learning using a GLM model with Poisson responses y and the canonical response function.

2.3. **Gaussian discriminant analysis** (4 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:

$$\begin{aligned} y^{(i)} &\sim \text{Multinomial}(\phi_1, \ldots, \phi_k) \\ \boldsymbol{x}^{(i)}|y^{(i)} = j \quad &\sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \end{aligned}$$

where the parameter ϕ_j gives $p(y^{(i)} = j)$ for each $j \in \{1, 2, \ldots, k\}$.
In Gaussian Discriminant Analysis (GDA), Linear Discriminant Analysis (LDA) assumes 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 when $k = 2$.