Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Fall 2024

Writing Assignment 2

Issued: Friday 18th October, 2024 **Due:** Friday 1st November, 2024

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. **Naive Bayes Parameter Learning** (3 points)

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:

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

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.

2.3. **Comparison of Generative and Discriminative Models**

In a binary classification problem, we can use a generative model such as Gaussian Discriminant Analysis (GDA) and a discriminative model such as Logistic Regression for classification. Assume we have a binary classification problem with samples $\{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})\}$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ is a d-dimensional feature vector, and $y^{(i)} \in \{0, 1\}$ is the class label.

(a) (2 points) Parameter Estimation. Derive the parameter estimation for GDA with a shared covariance matrix, also known as LDA. Assume that the feature $\boldsymbol{x}$ given the class y follows a Gaussian distribution:

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ \boldsymbol{x} \mid y = 0 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \\ \boldsymbol{x} \mid y = 1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \end{aligned}$$

(b) (1 point) Given the following dataset, compute the parameters for a given dataset.

$$\boldsymbol{x}^{(1)} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad y^{(1)} = 0$$

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

$$\boldsymbol{x}^{(2)} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \quad y^{(2)} = 0$$

$$\boldsymbol{x}^{(3)} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad y^{(3)} = 1$$

$$\boldsymbol{x}^{(4)} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \quad y^{(4)} = 1$$

Calculate the values of $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}$, and ϕ.

(c) (1 point) Decision Boundary for LDA. Derive the equation for the decision boundary for LDA. Assume the decision rule is:

$$\hat{y} = \arg\max_y P(y \mid \boldsymbol{x})$$

Show how the decision boundary can be derived from the class-conditional Gaussian distributions and the prior probabilities.

(d) (1 point) Compare the decision boundaries of LDA and logistic regression. Assuming that the covariance matrix $\boldsymbol{\Sigma}$ is the identity matrix in LDA, discuss whether the decision boundary of logistic regression is linear or nonlinear and explain why.

(e) (1 point) Suppose you have a very small dataset. Discuss how LDA and logistic regression might perform differently in this case. Given the properties of generative and discriminative models, explain why a generative model might perform better on small datasets.