Pattern Recognition and Machine Learning: Homework 3

Qingru Hu 2020012996

March 13, 2023

Problem 1

The loss function for LDA (using Fisher's condition) is:

$$l_{\text{LDA}} = \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i)^2 \tag{1}$$

The loss function for logistic regression is:

$$l_{\text{LG}}(x) = \begin{cases} -\log(\theta(x)) & y_i = 1 \\ -\log(1-\theta(x)) & y_i = 0 \end{cases}$$

where $\theta(x) = \frac{e^{\boldsymbol{w}^\top \boldsymbol{x}_i + b}}{1+e^{\boldsymbol{w}^\top \boldsymbol{x}_i + b}}$. The loss function of soft SVM is:

$$l_{\text{hinge}}(x) = \max(1 - y_i \cdot (\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0) \tag{2}$$

For a concrete class $y = 1$, plot curves of the three loss functions in Fig.1.

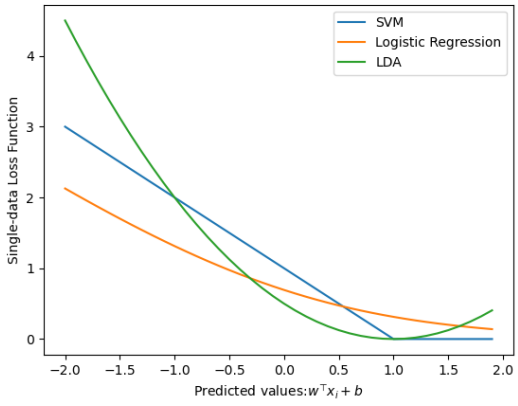

Figure 1: The single-data loss function for the three methods

How does that affect these methods' behavior and make them different?

Problem 2

The **hard-margin** problem is:

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t.} \quad y_i \cdot (\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ 0 \leq i \leq n. \end{aligned}$$

The Lagrangian function is:

$$\begin{aligned} L(\boldsymbol{w}, b, \alpha, \xi, \mu) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i [1 - y_i(\boldsymbol{w}^\top x_i + b)] \\ \alpha_i \geq 0, i = 1, \ldots, n \end{aligned}$$

Take the partial derivatives of Lagrangian w.r.t $\boldsymbol{w}$,b and set to zero:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = 0 \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

Pluge the above relations into the original Lagrangian function and we can get:

$$\begin{aligned} L(\boldsymbol{w}, b, \alpha, \xi, \mu) &= \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i (\boldsymbol{w}^\top x_i + b) \\ &= \frac{1}{2}(\sum_{i=1}^{n} \alpha_i y_i x_i)^\top (\sum_{i=1}^{n} \alpha_i y_i x_i) + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} (\alpha_i y_i (\sum_{j=1}^{n} \alpha_j y_j x_j)^\top x_i) \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \end{aligned}$$

Therefore, the original optimal problem is equivalent to the dual problem:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \\ &\text{s.t.} \sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, \ldots, n \end{aligned}$$

Denote the solution in the first problem as $\boldsymbol{w}^*$ and solution in the second problem as $\boldsymbol{\alpha}^*$, and the solutions satisfy:

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

Problem 2 continued on next page…

From the complementary slackness of KKT conditions, we have:

$$\alpha_i[1 - y_i(\boldsymbol{w}^\top x_i + b)] = 0$$

If $\alpha_i > 0$, then $y_i(\boldsymbol{w}^\top x_i + b) = 1$. Notice that $y_i = \pm 1$, so $\boldsymbol{w}^\top x_i + b = y_i$.

Problem 3

Search the paper and conduct a literature review for how to add different regularizations to SVM, and more sophisticated forms of kernel function.

Problem 4

(a) Use SVM to classify the train set and predict on the test set

Use Linear SVM to classify the samples on the training set, and the accuracy on the test set is 97%.

(b) Different kernel functions and penalty strength C

Kernel	Penalty Strength	Accuracy(%)
Linear	0.1	96
Poly	0.1	79
RBF	0.1	84
Sigmoid	0.1	86
Linear	1	96
Poly	1	83
RBF	1	97
Sigmoid	1	81
Linear	10	96
Poly	10	90
RBF	10	97
Sigmoid	10	76

Table 1: Different kernel functions and penalty strength for SVM

The best composition is the RBF kernel with a penalty strength of 1 or 10, which both have an accuracy of 97%.

(c) The supporting vectors