6.867: Homework 2

1. Logistic regression

In this section, we explore logistic regression with L1 and L2 regularization. We use gradient descent to compare the resulting weight vectors under different regularizers and regularization parameters, and we evaluate the effect of these choices in the context of multiple data sets.

1.1. L2 regularization

We first consider L2 regularization, in which the objective function to minimize is

$$E_{LR}(w, w_0) = \text{NLL}(w, w_0) + \lambda |w|^2$$

where

$$\text{NLL}(w, w_0) = \sum_i \log(1 + \exp(-y^{(i)}(wx^{(i)} + w_0)))$$

and in the case of L2 regularization,

$$|w| = |w|_2 = \sqrt{w_1^2 + ... + w_n^2}$$

Gradient descent was run with this objective function on the training dataset `data1_train.csv` with $\lambda = 0$. Interestingly, as the number of gradient descent iterations, controlled by the step size and the convergence criterion, increased, the weight vector ------

1.2. L1 regularization

In the case of L1 regularization, we get that in the above equation,

$$|w| = |w|_1 = \sum_{i=1}^{n} |w_i|$$

We can evaluate the different regularization techniques under different values of λ in the context of the weight vectors, the decision boundary, and the classification error rate in each of the training data sets.

Dataset	Best regularizer	Best λ	Test performance
1	5	6	1
2	5	6	1
3	5	6	1
4	5	6	1

Table 1. Optimal regularizer and λ for datasets

1.2.1. Weight vector

1.2.2. Decision boundary

1.2.3. Classification error rate

1.3. Optimization

By using the training and validation data sets, we can identify the best regularizer and value for λ for each of the four data sets. These results are presented in Table 1.

2. Support Vector Machine

In this section, we explore various versions of the dual form of support vector machines, first with slack variables and then with generalized kernel functions.

2.1. Dual form with slack variables

We here implement a dual form of linear SVMs with slack variables. More specifically, we solve the following optimization problem with respect to α:

$$\max_{\alpha} -\frac{1}{2}|\sum_i \alpha_i y^{(i)} x^{(i)}|^2 + \sum_i \alpha_i$$

$$\text{s.t.} \sum_i \alpha_i y^{(i)} = 0$$

$$0 \leq \alpha_i \leq C, 1 \leq i \leq n$$

Written another way, this maximization problem can be framed as a minimiation problem:

$$\min_{\alpha} \frac{1}{2} x^T P x + q^T x$$

$$\text{s.t.} Gx \leq h$$

$$Ax = b$$

where $b = 0$ and

$$x = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}, q = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}, A^T = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}, G = \begin{bmatrix} I \\ -I \end{bmatrix},$$

where I and $-I$ are the identity and negative identity matrix respectively. Furthermore,

$$P = \begin{bmatrix} x_0^2 y_0^2 & x_0 y_0 x_1 y_1 & \dots & x_0 y_0 x_{n-1} y_{n-1} \\ x_1 y_1 x_0 y_0 & x_1^2 y_1^2 & \dots & x_1 y_1 x_{n-1} y_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-1} y_{n-1} x_0 y_0 & x_{n-1} y_{n-1} x_1 y_1 & \dots & x_{n-1}^2 y_{n-1}^2 \end{bmatrix}.$$

$$h^T = \begin{bmatrix} C & \dots & C & 0 & \dots & 0 \end{bmatrix}$$

In the context of the four-point 2D problem, we seek to solve the following optimization:

$$\min_{\alpha} \frac{1}{2} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}^T \begin{bmatrix} 16 & 24 & 0 & 24 \\ 24 & 36 & 0 & 36 \\ 0 & 0 & 0 & 0 \\ 24 & 36 & 0 & 36 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

$$\text{s.t.} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \leq \begin{bmatrix} C \\ C \\ C \\ C \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ -1 \\ -2 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = 0$$

Solving for x when $C = 1$, we get that $\begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}^T = \begin{bmatrix} 0.1875 & 0 & 0.3750 & 0 \end{bmatrix}^T$. This indicates that the first and third samples are support vectors because $0 \leq \alpha_i \leq C$.

Our implementation of the dual form of SVMs with slack variables was run on each of the four 2D datasets provided. With $C = 1$, the decision boundaries and classification error rates were determined. This information is illustrated and summarized in Figure 1 and Table 2, respectively. As can be seen from the classification error rates in training and validation, depending on the values of the trained parameters and the spread of the datasets, the classification error could be greater for the training data (as in dataset 2), for the validation data (as in dataset 4), or equal (as in dataset 1).

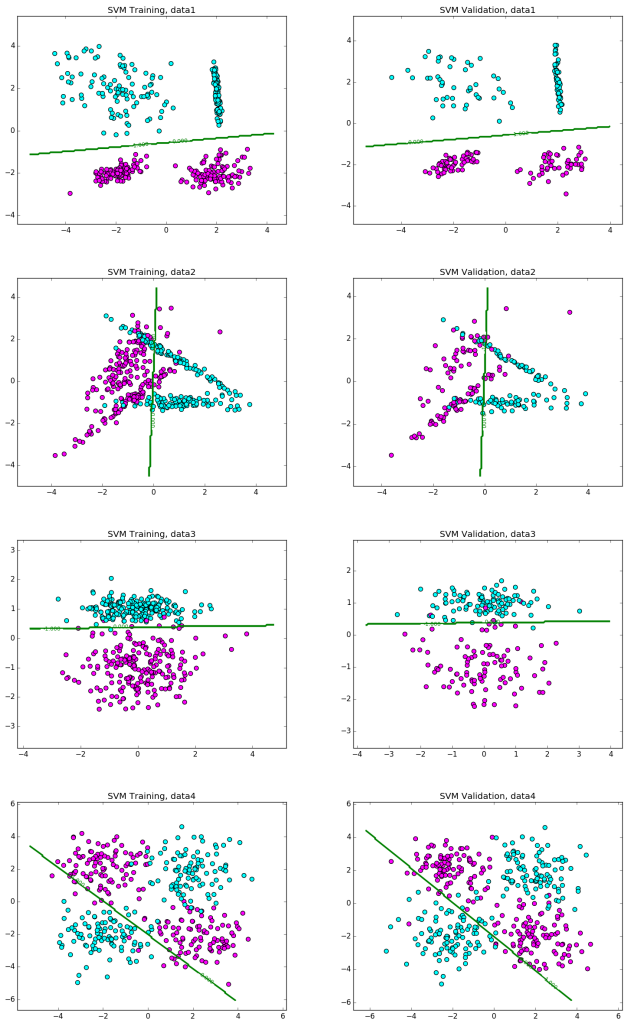

Figure 1. Training and validation decision boundary

Dataset	Training error rate	Validation error rate
1	0.0	0.0
2	0.1775	0.09
3	0.02	0.015
4	0.3	0.305

Table 2. Training and validation error rates

2.2. Dual form with kernels

3. Pegasos

4. Handwritten digit recognition