EE 559 Homework 6

Posted: March 28, 2023,
Due: March 31, 2023, 11:59 PM PDT
Keith M. Chugg, B. Keith Jenkins
version 4

1 SVM using `sklearn`

In this problem, you will use the `sklearn` implementation of SVM solver, which is based on LIBSVM. You will try different hyperparameters of SVM models on the training and testing datasets 1 (linearly separable) and 3 (not linearly separable) from Homework 1 . You are provided with a routine called `plotSVMBoundaries()` that will allow you to visualize the decision boundaries and regions learned by the SVM.

In all cases below:

- Use class `sklearn.svm.SVC`
- Always report train and test accuracies.
- For linear kernels, also give weight vector $\underline{w}$ including offset $\underline{w_0}$.
- For all plots, show the data (training or testing data), and decision regions using `plotSVMBoundaries()` (separate plots for training data and testing data).
- For some plots, you will be asked to show which training vectors are support vectors, (on the training-data plots only).
- Tip: A definition and description of support vectors is summarized at the end of this problem.

In problems 1(a) - 1(b), you will be using Homework 1 dataset 1.

(a) Use the `Linear Kernel` and try different values of slack variable parameter C. What is the meaning of parameter C and how will it impact your classification? Set $C = 0.01$ and $C = 1$. Report the above items and also provide the support vectors in the plots for each value of C. Discuss your results and explain the performance and the differences for the different values of C.

You will provide 4 *plots in total;* 2 *(for train and test sets) using* $C = 0.01$ *and* 2 *(for train and test sets) for* $C = 1$.

(b) Use a `Gaussian (RBF) Kernel` with C parameter set to $C = 0.01$. Set $\gamma = 1, 3, 10, 50$. Report the above items and also show the support vectors in the training-data plots for each value of γ. Explain the linearity or nonlinearity of the decision boundary, and explain the difference in decision regions for the various values of γ. State where (if anywhere) you observe underfitting or overfitting.

You will provide 8 *plots in total.*

In problems 1(c) - 1(e), you will be using Homework 1 dataset 3.

(c) Use the `Linear Kernel` and try different values of slack variable parameter C. Set $C = 1$ and $C = 100$. Report the above items for each value of C. Discuss your results. *You will provide* 4 *plots in total.*

(d) Use a `Gaussian (RBF) Kernel` with C parameter set to $C = 1$. Set $\gamma = 0.1, 10, 200$. Report the above items and also show the support vectors in the training-data plots for each value of γ. Explain the difference in decision regions for the different values of γ. Tip: you might want to try plots at other values of γ to help you understand its effects (no need to include these extra plots in your solution). Do you observe any overfitting or underfitting for any of the given values of γ? *You will provide* 6 *plots in total.*

(e) Use a `Gaussian (RBF) Kernel` and pick the γ parameter from part (d) (from the 3 given values) that results in the minimum test error. Set $C = 0.01, 1, 100$. Report the above items and also provide the support vectors in the plots for each value of C. Explain your observations in the different decision boundaries and the support vectors for the different values of C. *You will provide* 6 *plots in total.*

Support vectors are the data points (from the training set) that determine the position and orientation of the max-margin decision-boundary hyperplane. The support vectors are mathematically defined as the training data points that have Lagrange multiplier value (at the solution) of $\lambda_i \neq 0$. For linearly separable data, the support vectors are the vectors that lie on the margin boundary (for nonlinear kernels, this applies in the expanded dimensional feature space). For datasets that are not linearly separable in the expanded dimensional feature space, the support vectors are the vectors that lie on the margin boundary or on the wrong side of the margin boundary. (You can infer this behavior from the KKT conditions in the SVM solution (c.f. lecture notes).) The `sklearn SVC` model outputs the support vectors with the attribute `support_vectors_`.

2 Backprop Initialization for Multi-class Classification

Let $\underline{a} = \underline{a}^{(L)}$ denote the pre-activation at the final layer of an ANN and let $\underline{\hat{p}} = \underline{v}^{(L)} = \underline{\text{SoftMax}}(\underline{a})$ be the final output when the output layer is a SoftMax layer. Recall that multi-class cross-entropy cost is given as

$$J = -\sum_{i=1}^{C} p_i \ln \hat{p}_i \tag{1}$$

where $\underline{p}$ is the label vector. Finally, recall that delta vector recursion is initialized via

$$\underline{\delta}^{(L)} = \frac{\partial \underline{\hat{p}}}{\partial \underline{a}} \nabla_{\underline{\hat{p}}} J \tag{2}$$

where the denominator convention and right-to-left chain rule for vector differentiation has been used.

1. What is the dimension of each of the following quantities: $\frac{\partial \underline{\hat{p}}}{\partial \underline{a}}$, $\nabla_{\underline{\hat{p}}} J$, and $\underline{\delta}^{(L)}$?

2. Find $\frac{\partial \underline{\hat{p}}}{\partial \underline{a}}$.

3. Find $\nabla_{\underline{\hat{p}}} J$.

4. Show that $\underline{\delta}^{(L)} = \underline{\hat{p}} - \underline{p}$

This problem starts from

$$\boldsymbol{\delta} = \nabla_{\mathbf{s}} C = \dot{\mathbf{A}} \dot{C}(\mathbf{a}) \tag{3}$$

with

$$\dot{\mathbf{A}} = \frac{d\mathbf{h}(\mathbf{s})}{d\mathbf{s}} = \begin{bmatrix} \frac{\partial a_0}{\partial s_0} & \cdots & \frac{\partial a_{M-1}}{\partial s_0} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_0}{\partial s_{M-1}} & \cdots & \frac{\partial a_{m-1}}{\partial s_{M-1}} \end{bmatrix} \tag{4}$$

where the simplification that $\nabla_{\mathbf{a}} C = \dot{C}(\mathbf{a})$ for this MCE loss.

Let's first find the $\dot{\mathbf{A}}$ matrix. The softmax can be written as

$$a_j = \frac{e^{s_j}}{D(\mathbf{s})} \tag{5}$$

where

$$D(\mathbf{s}) = \sum_{m=0}^{M-1} e^{s_m} \tag{6}$$

Therefore, we have

$$\frac{\partial a_j}{\partial s_i} = \frac{\frac{\partial s_j}{\partial s_i}}{D(\mathbf{s})} - \frac{e^{s_j}}{[D(\mathbf{s})]^2} \frac{\partial D(\mathbf{s})}{\partial s_i} \tag{7}$$

First, let's find the partial of $D(\mathbf{s})$ with respect to s_i

$$\frac{\partial D(\mathbf{s})}{\partial s_i} = e^{s_i} \tag{8}$$

and the other term is

$$\frac{\partial s_j}{\partial s_i} = e^{s_i} \delta_{i,j} \tag{9}$$

where $\delta_{i,j} = 0$ if $i \neq j$ and 1 if $i = j$. Substituting these back into (7)

$$\frac{\partial a_j}{\partial s_i} = \frac{e^{s_j}}{D(\mathbf{s})} \delta_{i,j} - \frac{e^{s_j}}{[D(\mathbf{s})]^2} e^{s_i} \tag{10}$$

$$= a_j \delta_{i,j} - a_i a_j \tag{11}$$

$$= \begin{cases} a_i(1 - a_i) & i = j \\ -a_i a_j & i \neq j \end{cases} \tag{12}$$

It follows that

$$\left[\dot{\mathbf{A}}\right]_{i,j} = \begin{cases} a_i(1 - a_i) & i = j \\ -a_i a_j & i \neq j \end{cases} \tag{13}$$

or

$$\dot{\mathbf{A}} = \begin{bmatrix} a_0(1-a_0) & -a_0a_1 & \cdots & -a_0a_{M-1} \\ -a_0a_1 & a_1(1-a_1) & \cdots & -a_1a_{M-1} \\ \vdots & & \ddots & \vdots \\ -a_0a_{M-1} & \cdots & -a_{M-1}a_{M-2} & a_{M-1}(1-a_{M-1}) \end{bmatrix} \tag{14}$$

The j^{th} component of $\dot{C}(\mathbf{a})$ is given by

$$\frac{\partial C}{\partial a_j} = \frac{\partial}{\partial a_j}\left(-\sum_{i=1}^{n} y_i \ln a_i\right) = \frac{-y_j}{a_j} \tag{15}$$

or, in vector form

$$\dot{C}(\mathbf{a}) = -\begin{bmatrix} y_0/a_0 \\ y_1/a_1 \\ \vdots \\ y_{M-1}/a_{M-1} \end{bmatrix} \tag{16}$$

Finally, multiplying $\dot{C}(\mathbf{a})$ by $\dot{\mathbf{A}}$ yields

$$\boldsymbol{\delta} = \dot{\mathbf{A}}\dot{C}(\mathbf{a}) = \begin{bmatrix} -(1-a_0)y_0 + a_0y_1 + a_0y_2 + \cdots a_0y_{M-1} \\ a_1y_0 - (1-a_1)y_1 + a_1y_2 + \cdots a_1y_{M-1} \\ \vdots \\ a_{M-1}y_0 + a_{M-1}y_1 + a_{M-1}y_2 + \cdots - (1-a_{M-1})y_{M-1} \end{bmatrix} \tag{17}$$

Consider the first element in this vector

$$\delta_0 = -(1-a_0)y_0 + a_0y_1 + a_0y_2 + \cdots a_0y_{M-1} \tag{18}$$

$$= y_0 + a_0\sum_{m=0}^{M-1} y_m \tag{19}$$

$$= a_0 - y_0 \tag{20}$$

where the fact that $\{y_m\}$ is a probability mass function was used. This applies to each row of this vector, hence

$$\boldsymbol{\delta} = \mathbf{a} - \mathbf{y} \tag{21}$$

Note that we never made the assumption of one-hot or hard-labels – *i.e.,* the result is valid for soft-labels too!

3 Backprop by Hand

An MLP has three input nodes, two hidden layers, and three outputs. The activation for the hidden layers is ReLu. The output layer is softmax. The weights and biases for this MLP are:

$$\underline{\underline{W}}^{(1)} = \begin{bmatrix} 1 & -2 & 1 \\ 3 & 4 & -2 \end{bmatrix}, \quad \underline{b}^{(1)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$\underline{\underline{W}}^{(2)} = \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix}, \quad \underline{b}^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\underline{\underline{W}}^{(3)} = \begin{bmatrix} 2 & 2 \\ 3 & -3 \\ 2 & 1 \end{bmatrix}, \quad \underline{b}^{(3)} = \begin{bmatrix} 0 \\ -4 \\ -2 \end{bmatrix}$$

1. **Feedforward Computation:** In this part you will perform the feedforward computation for the input vector $\underline{x} = [\ +1\ \ -1\ \ +1\ ]^T$. The standard notation used in the slides and handouts is used. Specifically, $\underline{a}^{(l)}$ is the linear activation – *i.e.,* $\underline{v}^{(l)} = \mathbf{h}(\underline{a}^{(l)})$ – and $\dot{\underline{v}}^{(l)} = \dot{\mathbf{h}}(\underline{a}^{(l)})$. Find $\underline{a}^{(l)}$ and $\underline{v}^{(l)}$ for $l = 1, 2, 3$ and $\dot{\underline{v}}^{(l)}$ for $l = 1, 2$. In the process, if you need the value of the derivitive of ReLU(a) at $a = 0$, use $1/2$.

2. **Back-propagation Computation:** Run the standard SGD back-propagation for this input assuming a multi-category cross-entropy loss function and the one-hot labeled target: $\underline{p} = [\ 0\ \ 0\ \ 1\ ]^T$. Again, our standard notation is used so that $\underline{\delta}^{(l)} = \nabla_{\underline{a}^{(l)}} J$. Determine $\underline{\delta}^{(l)}$ for $l = 3, 2, 1$ and provide the updated weights and biases assuming a learning rate of $\eta = 0.5$ – *i.e.,* find $\underline{\underline{W}}^{(l)}(i+1)$ and $\underline{b}^{(l)}(i+1)$ for $l = 1, 2, 3$ assuming that weights and biases provided are the values at iteration i.

3. For this fixed input $\underline{x}$, show that the mapping from $\underline{x}$ to to $\underline{a}^{(3)}$ is is of the form$\underline{a}^{(3)} = \underline{\underline{W}}_{\text{eff}}\, \underline{x} + \underline{b}_{\text{eff}}$, where $\underline{\underline{W}}_{\text{eff}}$ and bias $\underline{b}_{\text{eff}}$ are the "effective" weight matrix and bias vector. Are these effective weights and biases a function of $\underline{x}$? Explain how this allows one to interpret an ANN with ReLu activations.