CS 189 Introduction to Machine Learning
Fall 2018 DIS9

1 Generalization of SVM: piecewise linear boundaries with ReLUs

In this problem we explore a generalization of the support vector machine for binary classification. In particular, we will see how a neural net with one hidden layer and ReLU activations can express a piecewise linear decision boundary, which can help find a good decision boundary for data sets that are not linearly separable. To be more precise, in this problem we consider two dimensional data points $\mathbf{x} \in \mathbb{R}^2$ which have labels $y \in \{-1, 1\}$. Given a training set of such points we are interested to find a model, of the form

$$\hat{y}_{\mathbf{w},b,\alpha}(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j \max\{\langle \mathbf{w}_j, \mathbf{x} \rangle + b_j, 0\}, \tag{1}$$

which fits the data, where the coefficients α_j and b_j are real valued scalars, the coefficients $\mathbf{w}_j$ are two dimensional, and k denotes the number of hidden units of the neural network. The coefficient λ quantifies the amount of regularization on $\mathbf{w}_j$ and α_j. The subscript of $\hat{y}_{\mathbf{w},b,\alpha}$ is a reminder that the prediction depends on the coefficients of the neural network.

From now on we use the shorthand notation $(\gamma)_+ = \max\{\gamma, 0\}$, for any real value γ. During training we are interested in minimizing the following hinge loss objective

$$F(\mathbf{w}_1, \ldots, \mathbf{w}_k, b_1, \ldots, b_k, \alpha_1, \ldots \alpha_k) = \frac{1}{n} \sum_{i=1}^{n} \left[(1 - y_i \cdot \hat{y}_{w,b,\alpha}(\mathbf{x}_i))_+ + \frac{\lambda}{2} \sum_{j=1}^{k} (\|\mathbf{w}_j\|_2^2 + \alpha_j^2) \right]. \tag{2}$$

We denote

$$f_i(\mathbf{w}, b, \alpha) = (1 - y_i \cdot \hat{y}_{w,b,\alpha}(\mathbf{x}_i))_+ + \frac{\lambda}{2} \sum_{j=1}^{k} (\|\mathbf{w}_j\|_2^2 + \alpha_j^2).$$

(a) As a warmup, we discuss problem 3c) from the midterm exam. Consider classifying two data points $\mathbf{x}_1 = (a, b)$, $\mathbf{x}_2 = (-a, -b)$, with labels $y_1 = +1$ and $y_2 = -1$, respectively. For this data, calculate the form of the maximum margin separating hyperplane which goes through the origin. Make sure you justify your answer mathematically. Recall that for linear classifiers, the maximum margin is defined as:

$$\max_{\mathbf{w} \in \mathbb{R}^d} \min_{1 \le i \le n} \left(\frac{\mathbf{w}^\top \mathbf{x}_i}{\|\mathbf{w}\|_2} y_i \right)$$

(b) Now, let us consider a data set in $\mathbb{R}^2$ consisting of the 9 data points with integer coefficients (a, b), where $0 \leq a, b \leq 2$. A data point (a, b) is labelled $+1$ if $\max\{a, b\} \geq 2$ and is labelled -1 otherwise. Draw a picture of this data set and draw a maximum margin piecewise linear decision boundary for it. You do not need to be formal about proving it is maximum margin, but provide a discussion about why your choice maximizes the margin.

(c) Find a choice of k and a choice of coefficients $\mathbf{w}_1, \ldots, \mathbf{w}_k, b_1, \ldots, b_k, \alpha_1, \ldots \alpha_k$ such that the ReLU NN introduced in equation (1) achieves zero cost in terms of the loss function

$$\frac{1}{n} \sum_{i=1}^{n} \max\{1 - y_i \cdot \hat{y}_{w,b,\alpha}(\mathbf{x}_i), 0\}. \tag{3}$$

Discuss how the decision boundary of the ReLU NN you chose approximates the decision boundary found in the previous part (here, the decision boundary of the ReLU NN means the set $\{\mathbf{x} | \hat{y}_{\mathbf{w},b,\alpha}(\mathbf{x}) = 0\}$).

(d) Compute the update rule for stochastic gradient descent on the objective F, defined in equation (2), for optimizing the coefficients $\mathbf{w}_j$, b_j and α_j.

(e) Given an arbitrary training set in $\mathbb{R}^2$ of distinct data points, intuitively discuss why a ReLU NN can perfectly fit the data if the number of hidden units k is large enough. Does it immediately follow that running SGD during training would find such a ReLU NN?