Comparison Study of SVM and MLP

Yingxue Su, Qianfan Bai

April 25, 2020

Multilayer Perceptron Structure

- We have in total L cases in the training set. For each case (x_l, y_l), y_l is the true label of the case $x_l \in \mathbf{R}^k$. We have m classes in total.
- The numbers of units in the input layer, hidden layer and output layer are k, N and m, respectively. Note that N is chosen by you.

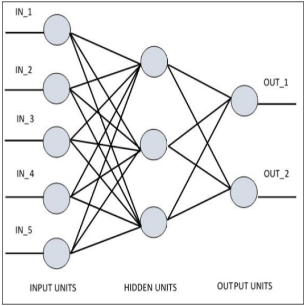
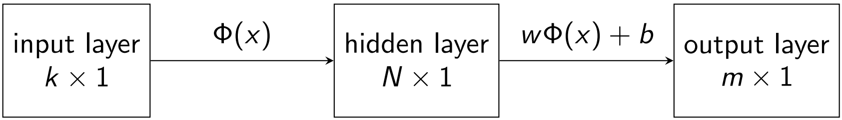


Figure: MLP with one hidden layer

Multilayer Perceptron Structure

- 

- $\Phi(\cdot) = (\Phi_1(\cdot), ..., \Phi_N(\cdot))$ and $\Phi_i(x) = h(v_i x + d_i)$ for $i = 1, ..., N$.
- $f_\theta(x) = w \cdot \Phi(x) + b$ is the decision function.
- h is the activation function which is usually sigmoid function, relu function or hyperbolic tangent function.

Multilayer Perceptron Structure

- MLP is achieved by the minimization of a given loss function using the gradient descent method. Our loss function is defined as

$$loss = \frac{\mu}{2}||\theta||^2 + \frac{1}{L}\sum_{l=1}^{L} Q(f_\theta(x_l), y_l) \tag{1}$$

  $Q(f_\theta(\cdot), \cdot)$ is the criterion function which is usually MSE or Cross-Entropy. If we choose Q to be the Cross-Entropy criterion, we have the following optimization problem:

$$min \quad \frac{\mu}{2}||\theta||^2 + \frac{1}{L}\sum_{l=1}^{L} \log(1 + \exp(-f_\theta(x_l) \cdot y_l)) \tag{2}$$

- $\theta_{t+1} = \theta_t - \epsilon_t \frac{\partial}{\partial\theta} loss_t$. $\epsilon(t)$ is the learning rate and $\epsilon(t) \to 0$ when $t \to \infty$.

Links Between MLP and SVM

- The decision function of SVM also has the form

$$f_\theta(x) = w\Phi(x) + b.$$

The SVM problem is equivalent to the following optimization problem:

$$min \quad \frac{\mu}{2}||w||^2 + \frac{1}{L}|1 - y_l f_\theta(x_l)|_+,$$

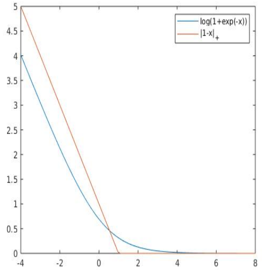
where $|z|_+ = max(0, z)$. The above optimization problem is equivalent to

$$\begin{aligned} \text{min} \quad & \frac{\mu}{2}||w||^2 + \frac{1}{L}\xi_l \\ \text{subject to} \quad & \xi_l \geq 0 \\ & 1 - \xi_l - y_l f_\theta(x_l) \leq 0 \end{aligned} \tag{3}$$

Links Between MLP and SVM

- The margin criterion is a 'hard' version of Cross-entropy criterion.
- Replace the Cross-entropy criterion in optimization problem (2) and rewrite it as:

$$
\begin{aligned}
\min \quad & \frac{\mu}{2}||\theta||^2 + \frac{1}{L}\xi_l \\
\text{subject to} \quad & \xi_l \geq 0 \\
& 1 - \xi_l - y_l f_\theta(x_l) \leq 0
\end{aligned}
\tag{4}
$$



Links Between MLP and SVM

- By comparing the KKT conditions of optimization problem (3) and (4), we notice that (w^*, b^*, Φ^*) which satisfies the KKT of (4) also satisfies the KKT of (3).
- (w^*, b^*) are the optimal weights for SVM using the feature space described by Φ^*,

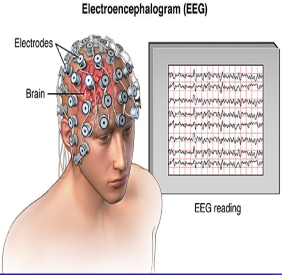
$$\Phi_i^* = h(v_i^* x + d_i^*).$$

- MLP maximize the margin in the hidden layer space.
- For cases x_l such that $|v_i^* x_l + d_i^*| \leq 1$, unites i form a linear SVM. And the standard separation constraints $y_l(v_i^* x_l + d_i^*) \geq 1$ are replaced by

$$y_l(v_i^* x_l + d_i^*) \geq 1 - y_l(b + \sum_{k \neq i} w^* h(v_k^* x_l + d_k^*)).$$

Numerical Task and the Data Set

- Numerical task: We want to train a MLP and a SVM, using the EEG data set and compare their performances on separating the following 3 classes.
- We rearranged the original data set.The classes are:
  Class 1: the EEG signal is related to a tumor. (4600 cases)
  Class 2: the EEG signal is recorded during an eye activity. (4600 cases)
  Class 3: the EEG signal is recorded during a seizure activity.(4600 cases)



PCA Analysis on the Data Set

First, we apply PCA analysis on the whole data set. We can see that the data set is very hard to separate using the linear projection onto 3 dimensions.

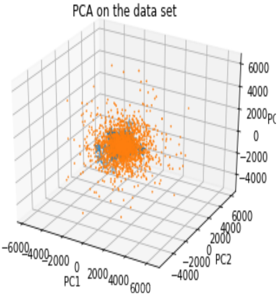


Figure: PCA on the whole data set

Results of SVM and MLP

- We use RBF kernel for this SVM model. The accuracy of this SVM on the training set and test set are 0.93126 and 0.89034, respectively.
- The confusion matrix of this SVM on the test set is

$$C\_M_{test} = \begin{bmatrix} 0.9116 & 0.0667 & 0.0217 \\ 0.2154 & 0.7809 & 0.0037 \\ 0.0197 & 0.0071 & 0.9732 \end{bmatrix}$$

- We select $N = 82$ for MLP by PCA analysis. The accuracy of this MLP on the traning set and test set are 0.8694 and 0.8490, respectively.
- The confusion matrix of this MLP on the test set is

$$C\_M_{test} = \begin{bmatrix} 0.8055 & 0.1529 & 0.0416 \\ 0.1507 & 0.8240 & 0.0253 \\ 0.0298 & 0.0082 & 0.9620 \end{bmatrix}$$

Hidden Layer Activity of MLP

- For case $x_j \in class_i$, let $PROF_i = \frac{1}{4600} \sum_j \Phi(x_j)$. We plot $PROF_i \;\; vs \;\; PROF_j$ for $i \neq j$.

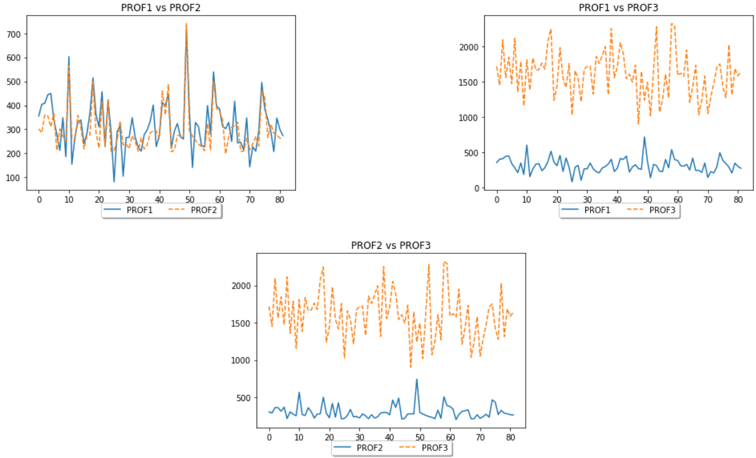


Figure: hidden layer activity

Conclusion of the Numerical Results

- The accuracy on the whole test set of SVM has the accuracy interval $[0.8855, 0.8952]$. Therefore, SVM has a better performance on classifying the 3 classes than MLP.
- By looking at the confusion matrix of SVM on the test set, the accuracy intervals of classifying 3 classes of SVM are $[0.9040, 0.9192]$, $[0.6689, 0.7921]$ and $[0.9689, 0.9775]$, respectively. SVM has better performance in classifying class 1 and class 3. MLP has better performance in classifying class 2.
- By looking at the hidden layer activity of the MLP after training, we can see that the hidden layer units are much more active to the cases belonging to class 3 and they are similarly active to cases from class 1 and 2, which explains why MLP has a higher accuracy in classifying class 3.

Reference

Ronan Collobert, Samy Bengio
Links between Perceptrons, MLPs and SVMs.
Feb 6,2004.