Machine Learning Homework 2-Saket Vishwasrao

Written Problems

1. The gradient for the given objective function will be

$$\nabla_{\boldsymbol{w}_c} L = \lambda \boldsymbol{w}_c + \sum_{i=1}^{n} \boldsymbol{x}_i \left(\frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} - I(y_i = c) \right) \tag{1}$$

$$\tag{2}$$

At the optimum the gradient is zero.Hence we have

$$\lambda \boldsymbol{w}_c = -\sum_{i=1}^{n} \boldsymbol{x}_i \left(\frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} - I(y_i = c) \right) \tag{3}$$

$$\tag{4}$$

Summing over all classes in each dimension we get,

$$\lambda \sum_{c=1}^{C} w_c[j] = -\sum_{c=1}^{C} \sum_{i=1}^{n} x_i[j] \left(\frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} - I(y_i = c) \right) \tag{5}$$

$$\tag{6}$$

where j=1,2,3,..,d
Rearranging the summation

$$\lambda \sum_{c=1}^{C} w_c[j] = -\sum_{i=1}^{n} x_i[j] \left(\sum_{c=1}^{C} \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} - \sum_{c=1}^{C} I(y_i = c) \right) \tag{7}$$

$$\tag{8}$$

Now summing over all classes, the term $\sum_{c=1}^{C} \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} = 1$,beacause it the sum of all $p_(y = c|\boldsymbol{x}; W)$ over all the classes.
$\sum_{c=1}^{C} I(y_i = c)$ is a $C \times 1$ matrix that has only one out of its C elements equal to one corresponding to the predicted class for that particular sample and rest all are zero.Hence $\sum_{c=1}^{C} I(y_i = c) = 1$
Thus we have

$$\lambda \sum_{c=1}^{C} w_c[j] = 0 \tag{9}$$

which implies

$$\sum_{c=1}^{C} w_c[j] = 0 \tag{10}$$

2. (a) From equations 5 and 6 in the problem statement, we have

$$\boldsymbol{w} = \boldsymbol{w} - \eta(y_i' - y_i)\boldsymbol{x} \tag{11}$$

where y_i' is the predicted label and y_i is the given label.η is a constant.
Hence we have

$$\nabla J = \eta(y_i' - y_i)\boldsymbol{x} \tag{12}$$

Therfore

$$J = \int \eta(y_i' - y_i)\boldsymbol{x}\,\mathrm{d}\boldsymbol{w} \tag{13}$$

$$\implies J = \int \eta(sgn(\boldsymbol{w}^T\boldsymbol{x}) - y_i)\boldsymbol{x}\,\mathrm{d}\boldsymbol{w} \tag{14}$$

$$\implies J = \int \eta(sgn(\boldsymbol{w}^T\boldsymbol{x}) - y_i)\boldsymbol{x}\,\mathrm{d}\boldsymbol{w} \tag{15}$$

$$\implies J = \eta(|\boldsymbol{w}^T\boldsymbol{x}| - y_i\boldsymbol{w}^T\boldsymbol{x}) \tag{16}$$

Since we have $y_i' = \frac{|\boldsymbol{w}^T\boldsymbol{x}|}{\boldsymbol{w}^T\boldsymbol{x}}$

$$J = \eta(y_i'\boldsymbol{w}^T\boldsymbol{x} - y_i\boldsymbol{w}^T\boldsymbol{x}) \tag{17}$$

$$\implies J = \eta(y_i' - y_i)\boldsymbol{w}^T\boldsymbol{x} \tag{18}$$

(b) Using the result from eqaution 18,we have a batch training function as

$$F = \arg\min_{w} \sum_{i=1}^{N} |\eta(y_i' - y_i)\boldsymbol{w}^T\boldsymbol{x}| \tag{19}$$

(c) For the correct classification of all data points,there should be no loss.Hence the value of objective F (eq 19)is zero for correct classification.There will always we a trivial solution to the eq(19) with $\boldsymbol{w} = 0$ besides the optimum non-trivial $\boldsymbol{w}$ that minimizes the loss function.

(d) For a non-linear data set,again we have a trivial $\boldsymbol{w} = 0$ as the solution.Since we cannot have a non trivial $\boldsymbol{w}$ that exactly classifies data,optimizing the loss function will give $\boldsymbol{w} = 0$ which is not what we want.We want a $\boldsymbol{w}$ that correctly classifies the most number of samples i.e increases the accuracy.This is a degeneracy as in practice most datasets are non-linear.

(e) The perceptron update still works because in perceptron we do not try to find the optimum $\boldsymbol{w}$ over the entire batch of data but we update $\boldsymbol{w}$ for each sample.Thus we can stop the optimisation after we have a evaluated a specific number of iterations which may not arrive at the optimum w,but gives sufficient accuracy i.e we do not approach the trivial solution.Hence perceptron update works well despite the degeneracies.