CSE 512 – Homework I

By Yasha Singh - 112970310

March 15th 2020

1 Theory Questions

1.1 A "warm up" problem

Since, $Pr(y = 1|x > 0)$ is 0.9 and $Pr(y = -1|x \leq 0)$ is also 0.9 according the question. This would mean that both $\{+1, \quad -1\}$ labels for any sample X drawn randomly from the uniform distribution D are being correctly classified with 0.9 accuracy, therefore the training error for this hypothesis h will be equal to 0.1.

1.2 Bayes Optimal Predictor- ME

Given a classifier g such that, $L_D(f_D) \leq L_D(g)$ we have to show that for every Probability Distribution D, Bayes optimal predictor will be optimal i.e. it will minimize the loss function which is given by :-

$$L_D(h) = \begin{Bmatrix} P_{X\sim D}[y \neq 0|x] & if h(x) = 0 \\ P_{X\sim D}[y \neq 1|x] & if h(x) = 1 \end{Bmatrix}$$

This loss function can be re-written in these terms:-

$$L_D(h) = \begin{Bmatrix} P_{X\sim D}[y = 1|x] & if h(x) = 0 \\ 1 - P_{X\sim D}[y = 1|x] & if h(x) = 1 \end{Bmatrix}$$

If Pr[y=1 — x] ¡ 1 - Pr[y=1—x], then we select h(x) = 0, else we take h(x) = 1.
In both cases, the error $L_D(h)$ will then be min(Pr[y=1—x], 1 - Pr[y=1—x]).
Thus, $L_D(f_D) \leq L_D(g)$.

1.3 Perceptron with a Learning Rate

Given to us is $w^{(t+1)} = w^{(t)} + y_i * x_i$ which is getting modified to $w^{(t+1)} = w^{(t)} + \eta y_i x_i$.
(a) The number iterations is irrespective of the weight update and it is either determined by us by setting some epsilon such that difference between consecutive error is below epilson and we will stop the looping or set some max $iterations = T$ after which we stop. In both the cases, weight updation does not influence the number of epochs of the loop.

(b) Since, $0 < \eta < 1$, it will be positive so it will not affect the direction of gradient and second it is smaller than 1 so updating weight by a constant factor so, the algorithm will eventually converge in the same direction.

1.4 Unidentical Distributions

To solve this, we will assume X and Y as two events:-
$X : \exists h \in \mathcal{H} \quad \text{s.t.} \quad L_{(\overline{\mathcal{C}}_{mf})}(h) > \epsilon$ and $Y : L_{(s,f)}(h) = 0$
For two not independent variables X and Y, their probability together is given by the help of conditional probability as below :-
$Pr(XandY) = Pr(Y|X).Pr(X)$
In case of event X, $D_m(L_{(\overline{\mathcal{D}}_{mf})}(h) > \epsilon) \leq |H|$, will be true as the True error is for the mean of distribution and h is some hypothesis belonging to class H.
In case of event Y, loss of hypothesis L(h) = 0 for function f and so it means that training error is 0 and $h(x_i) = f(x_i)$. It follows that for each sample i :-

$$D(x_i : h_i(x) = f_i) = 1 - L_{D,f}(h)$$

So,

$$D(x_i : h_i(x) \leq f_i) \leq 1 - \epsilon$$

And for all m distributions :-

$$D_m(S|x : L_s(h) = 0) \leq (1 - \epsilon)^m$$

, So,

$$D_m(S|x : L_s(h) = 0) \leq (\epsilon)^{-em}$$

From, the results of above events X and Y, we can say that,

$$\Pr\Big(\exists h \in \mathcal{H} \quad \text{s.t.} \quad L_{(\overline{\mathcal{C}}_{mf})}(h) > \epsilon \ \text{ and } \ L_{(s,f)}(h) = 0 \Big) \leq |\mathcal{H}|\epsilon^{-em}$$

1.5 Vapnik-Chervonenkis (VC) Dimension

(a) Given to us, H^d is the class of axis-aligned rectangles in R^d and we have to prove that $VCdim(H^d) = 2d$. To prove this let us consider 2 cases.

First, we prove that VC Dimension $< 2d + 1$: We will consider points are in 2d+1 dimension and we will try to construct a rectangle R^d. As there are 2d+1 points, one of the points will lie within this rectangle. If we label this point as negative there will not have have a rectangle capable of separating this labelling, so VC Dimension has to be less than 2d.

Second, we prove that VC Dimension $\geq$ 2d : When this happens, the points the 2D space can be shattered by some axis aligned rectangle given that each point has one such dimension set to -1 or +1 and rest all dimension set to 0.

As in both the above cases, shattering can be accomplished by an axis aligned rectangle therefore, combining case 1 and 2, we have that VC Dimension = 2d.

(b)

1.6 Boosting

Updation in Adaboost is done by :-

$$D_i^{(t+1)} = \frac{D_i^{(t)} \exp\left(-w_t y_i h_t\left(\mathbf{x}_i\right)\right)}{\sum_{j=1}^{m} D_j^{(t)} \exp\left(-w_t y_j h_t\left(\mathbf{x}_j\right)\right)}$$

for $i = 1, 2, 3 \ldots$ m

And confidence is given by

$$w_t = \frac{1}{2}\left(\log\frac{1}{\epsilon} - 1\right)$$

Therefore, our numerator effectively will become - $= \varepsilon\sqrt{\frac{1-\varepsilon}{\varepsilon}}$

Where, $\varepsilon = \sum_{i=1}^{\infty} D_i^t \cdot I\left[y_i \neq h_i\left(x_i\right)\right]$

Similarly, for denominator, we will consider all the D_i, when actual label is not equal to predicted label and we will have to add $(\epsilon)\sqrt{(1-\epsilon)/\epsilon}$

But when, the predicted label and true labels are same, we will have to add $(1-\epsilon)\sqrt{(1-\epsilon)/\epsilon}$

For the above two, equally likely cases, upon taking average, denoninator will become -

$$\sum_{i=1}^{m} D_i^{(t+1)} I[ht\left(\mathbf{x}i\right) \neq yi] = \frac{1}{2}$$

1.7 Cross-validation

According to the scenario given to us, label is chosen at random according to $P[y = 1] = P[y = 0] = 0.5$ and our algorithm outputs constant predictor h(x) = 1 if the number of 1s in the training set labels is odd, and h(x) = 0 otherwise. If we consider that the distribution is i.i.d, so validation $loss = 1/2$ according to our h(x). Now, in case of leave-one-out cross validation, we are leaving out some point (x,y) whose label is 1 but our h(x) will label it according to majority so the predicted label for this sample point(x,y) will be 0. Similarly, for the opposite case when label of (x, y) is 0, h(x) will predict it as 1.

Now, if take mean for all the folds, estimation error of h(s) = 1 and validation error as mentioned above will be 0.5 and the therefore the true error would be the difference of the both errors being equal to 0.5.