Chapter 7 Review Notes

1 Stability or Robustness of Machine Learning

Main focus of this topic will be on Regression or classification. Consider a graph with positive and negative classes plotted. Now assume that the equation to minimize is

$$min_w \Big\{ ||w||^2 + C \sum_{i=1}^{D} max(0, 1 - w^T x_i) \Big\} \tag{1}$$

We assume that the bias is 0. Then the equation of optimal hyperplane is

$$w^T x = 0 \tag{2}$$

Assume that a new point is added to the dataset. Now due to the addition of that point how should the line or hyperplane should change?.

- The line will significantly shift.
- The line will not shift.
- The line will shift to a small extent.

The correct answer is The line will shift to a small extent. Robustness means stability. It means that additional point does not change learned parameters to a large extent. Consider the optimization problem with bias.

$$min_w \Big\{ ||w||^2 + C \sum_{i=1}^{D} max(0, y_i(1 - w^T x_i + b)) \Big\} \tag{3}$$

If we want this equation to be robust then an additional parameter is to be added

$$min_w \Big\{ ||w||^2 + ||b||^2 + C \sum_{i=1}^{D} max(0, y_i(1 - w^T x_i + b)) \Big\} \tag{4}$$

If the line or hyperplane is not robust then it leads to 2 problems.

- Problem of overfitting. It means that it models the noise as well.
- Differential Privacy will be breached. If our model changes with each additional input to a large extent then we can reverse engineer the data.

1.1 Hinge Loss

$$a_+ = max(0, a) = ReLU(a) \tag{5}$$

$$a_+ = \text{Hinge loss on a} \tag{6}$$

Let

$$D = x_i, y_i$$

then the optimization problem can be written as

$$L_D(w) = min_w \Big\{ \lambda ||w||^2 + \sum_{i=1}^{D} (1 - y_i w^T x_i)_+ \Big\} \tag{7}$$

The optimal solution can be expressed as

$$w^*(D, \lambda) = argmin_w L_D(w) \tag{8}$$

Now a new point (x,y)=e is added. Then the optimal solution becomes

$$w^*(D \cup e, \lambda) = argmin_w L_{D \cup e}(w) \tag{9}$$

If the learner or SVM is robust or stable then we expect that

$$||w^*(D, \lambda) - w^*(D \cup e, \lambda)|| \tag{10}$$

is small. Let the loss function be expressed as

$$l(i, w) = max(0, 1 - y_i w^T x_i) \text{For Classification} \tag{11}$$

$$l(i, w) = (y_i - w^T x_i)^2 \text{For Regression} \tag{12}$$

The optimization problem can be expressed as

$$\lambda ||w||^2 + \sum_{i=1}^{D} l(i, w) \tag{13}$$

If l is convex function then

$$l(i, w) = l(i, w^{'}) + \left(\frac{\partial l}{\partial w^{'}} \right)^T (w - w^{'}) + (w - w^{'})^T \left[\frac{\partial^2 l}{\partial w^{'2}} \right] (w - w^{'}) \tag{14}$$

$$l(i, w) \geq l(i, w^{'}) + \left(\frac{\partial l}{\partial w^{'}} \right)^T (w - w^{'}) \tag{15}$$

We should always note that λ should also change when our dataset changes. In theory

$$min_w \left(\lambda ||w||^2 + \frac{1}{|D|} \sum_{i=1}^{D} l(i, w) \right) \tag{16}$$

But in practice

$$min_w \left(\lambda ||w||^2 + \sum_{i=1}^{D} l(i, w) \right) \tag{17}$$

such that

$$\lambda = \lambda_c |D| \tag{18}$$

It means that we put λ of high value. For additional reading you can go through

- Stability and Generalization by olvier bousquet 2002
- Understanding ML (Pages 141-144)