**CS 419M Introduction to Machine Learning** **Spring 2021-22**

# Lecture 11: Some Problems

*Lecturer: Abir De* *Scribe: Group 11*

## 11.1 Regularization of bias

The primary aim of adding a bias in our model is to have clearly demarcated decision boundaries.Theoretically, the bias can be regularized, but in practice it makes little sense to do the same. For the purpose of regularization, the bias can be taken as a hyperparameter along with w i.e.,

$$[w, b] \cdot [x, 1] = w^T x + b$$

But rather than adding separate term of $\lambda b^2$ to our loss function, the performance of the model can instead be analyzed by using cross validation.
Moreover, there might be cases where w is small but the corresponding bias is high. Consider the following scenario shown in figure -

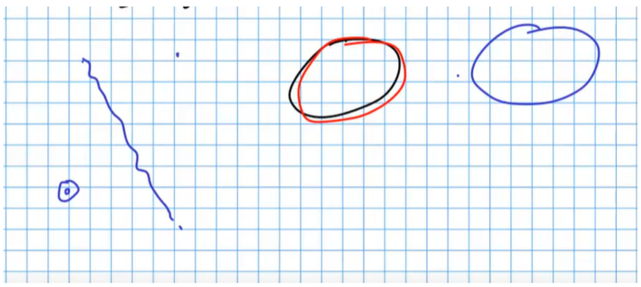

Figure 11.1: If the bias is regularized

Here, unless the origin isn't between the two clusters shown, regularization of the bias has no added advantage for the model. Therefore, regularization of bias generally leads to underfitting and is usually not the method adopted.

## 11.2 Stability of Ranking Loss

We have the ranking loss for all pairs of $x_{bad}$ and $x_{good}$ given by -

$$f(w) = \sum_{\substack{y_{bad}=-1 \\ y_{good}=+1 \\ x_{bad,good}\in S}} (1 + w^T x_{bad} - w^T x_{good})$$

What should be the regularizer $\lambda$ that should be added to the loss to ensure that it is stable?

Suggestion - If we consider $x_{bad} - x_{good}$ as a new proxy variable , then the regulariser will be the number of pairs of $x_{bad}$ and $x_{good}$. So the loss function becomes

$$f(w) = \sum_{\substack{y_{bad}=-1 \\ y_{good}=+1 \\ x_{bad,good}\in S}} (1 + w^T x_{bad} - w^T x_{good}) + \lambda |S_{good}||S_{bad}|||w||^2$$

However proving the stability for this is not the same as proving the stability for a single variable as before because here, when we change the set S by replacing an old point with a new point $x, y$ to get $S^{'}$, all pairs in the summation of which it is a part will change. Thus, this is different from a point-wise loss.

Optimal w can be easily found for this case by differentiating the loss term wrt w to get

$$w = \sum \frac{(x_{good} - x_{bad})}{2\lambda |S_{good}||S_{bad}|}$$

What would it be if the loss was calculated using the hinge function and what is the condition of stability?

$$f(w) = \sum_{\substack{y_{bad}=-1 \\ y_{good}=+1 \\ x_{bad,good}\in S}} (1 + w^T x_{bad} - w^T x_{good})_+ + \lambda |S_{good}||S_{bad}|||w||^2$$

Following the same line of proof as earlier, we will get some bound in terms of $\frac{1}{|S_{good}|} or \frac{1}{|S_{bad}|}$

But can we find some regulariser that is independent of the ratio of $x_{good}$ to $x_{bad}$?

Solution - If we use the following regulariser, though we would be over-regularising w but we can be ensured stability.

$$\frac{|S|(|S|-1)}{2}$$

## 11.3 Formation of Dual

Say we are dealing with the traditional linear regression problem. That is, given a dataset $\mathfrak{D}$,

$$\mathfrak{D} = \{(x_i, y_i) \mid i \in [|\mathfrak{D}|], x_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}$$

where, $[n] := \{1, 2, \dots n\} \forall n \in \mathbb{N}$.

We want to perform the following minimization,

$$\min_w \sum_i (y_i - w^T x_i)^2 + ||w||^2$$

We ask the following question.

Question Can we obtain the dual form of this problem, just like we did in the case of classification?

We will start by converting this unconstrained optimization to a constrained optimization problem. Further, we should check whether the constrained optimization problem is convex. With the idea of minimizing each of the square terms in mind, we define $\zeta_i$ such that

$$\zeta_i \geq (y_i - w^T x_i) \text{ ,with } \zeta_i \geq 0 \ \forall i$$

Thus now we have the problem:

$$\underset{w,\zeta_i}{min} (\sum_i \zeta_i + \lambda ||w||^2)$$

$$\textit{such that } \zeta_i \geq (y_i - w^T x_i) \ \forall i$$

Now, the second equation above gives us the constraint, and then we have

$$\underset{\alpha_i,\beta_i}{max} \ \underset{w,\zeta_i}{min} \left( \lambda ||w||^2 + \sum_i (\zeta_i + \alpha_i((y_i - w^T x_i)^2 - \zeta_i) - \beta_i \zeta_i) \right)$$

Further note that at the optimum value of $\zeta_i = \zeta_i^*$, the following would hold,

$$(y_i - w^T x_i)^2 = \zeta_i^* \ \forall i \text{ (Why?)}$$

## 11.4 Dimension of w

Suppose we are given a data set $\mathfrak{D}$,

$$\mathfrak{D} = \{(x_i, y_i) \mid i \in [|\mathfrak{D}|], x_i \in \mathbb{R}, y_i \in \mathbb{R}\}$$

where, $[n] := \{1, 2, \dots n\} \forall n \in \mathbb{N}$. Denote,

$$\mathfrak{D} \mid_x := \{x_i \mid (x_i, y_i) \in \mathfrak{D}\}$$

$$\mathfrak{D} \mid_y := \{y_i \mid (x_i, y_i) \in \mathfrak{D}\}$$

Further suppose that the data set is distributed around some line given by $y = mx + c$. That is,

$$y_i = mx_i + c + \epsilon_i \ \forall i \in [|\mathfrak{D}|]$$

where $\epsilon_i \sim \mathbb{P}(\cdot)$. Thus we have the following picture,

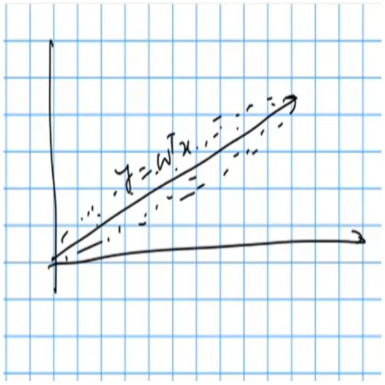

Figure 11.2: A picture of the data

Now, for each $x_i \in \mathfrak{D} \mid_x$, and for all $n \in \mathbb{N}$, define,

$$\mathbf{x}_i := (x_i, x_i^2 \dots x_i^n)^T \in \mathbb{R}^n$$

Further, let $\mathbf{w} \in \mathbb{R}^n$. We ask the following question,

Question What is the $n$, such that we can be *assured* that the following holds:

$$y_i = \mathbf{w}^T \mathbf{x}_i \forall y_i \in \mathfrak{D}_y$$

Informally, in the extreme overfitting case, what is the minimum error that we get – and what is the corresponding dimension of the weights vector, that is $n$?

Note the following lemma.

**Lemma 11.1.** *Given $n+1$ points $(x_i, y_i)$ with $n \geq 1$, such that all the $x_i$ are distinct. There exists a $n$ degree polynomial $p$ with coefficients $(a_n, a_{n-1}, \dots a_0)$ such that*

$$p(x_i) = y_i \forall i \in [n+1]$$

*Proof.* We have the following equations,

$$a_n x_i^n + a_{n-1} x_i^{n-1} + \dots a_0 = y_i \forall i \in [n+1]$$

This can be expressed in the matrix form as,

$$\begin{pmatrix} x_1^n & x_1^{n-1} & \dots & x_1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n+1}^n & x_{n+1}^{n-1} & \dots & x_{n+1} & 1 \end{pmatrix} \begin{pmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n+1} \end{pmatrix}$$

Note that the matrix on the left is a Vandermonde matrix, and has a non zero determinant. Thus the system of equations has a unique solution. □

For our case, since we do not have a bias term (note the definition of $\mathbf{x}_i$), our equations would be of the form,

$$a_n x_i^n + a_{n-1} x_i^{n-1} + \dots + a_1 x_i = y_i$$

Thus (assume non zero $x_i$),

$$a_n x_i^{n-1} + a_{n-1} x^{n-2} + \dots + a_1 = y_i / x_i$$

By the previous lemma, we have,

$$n - 1 = |\mathfrak{D}| - 1$$

Thus,

$$n = |\mathfrak{D}|$$

Thus, we should have,

$$\dim(\mathbf{w}) \leq |\mathfrak{D}|$$

□

But, this says that the number of parameters is dependent on the number of features! This should not be the case!

Interestingly, in deep learning, networks are more often than not extremely overparametrised (look up some typical neural nets) - and still we obtain good generalization errors! So, even with a *lot* of parameters *without* regularization, the deep learning model will “understand” the situation and put certain weights to zero!

## 11.5 Group Details and Individual Contribution

| Problem | Scribe |
|---|---|
| Regularization of Bias | Raavi Gupta (200070064) |
| Stability of Ranking Loss | Latika Patel (180100062) |
| Formation of Dual | Siddhant Midha (200070078) |
| Dimension of **w** | Siddhant Midha (200070078) |