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

Lecture 8: Multi-Class Classification and Stability

Lecturer: Abir De *Scribe: Group 8*

8.1 Recap

8.1.1 Binary Classification

Binary Classification Problem: Given an input $D = \{(x_i, y_i)|y_i = \{-1, +1\}\}$ where x_i corresponds to the labels and y_i is a binary output, we need to find a function (output) $m : x \to y$.
Mathematically, we let $P_m(y|x) = \frac{1}{1+e^{-w^T xy}} \Rightarrow \max_w \Pi P_m(y_i|x_i) \Rightarrow \min_w \sum \log(1 + e^{-w^T xy})$
If $w^T x$ is high or low, the with high confidence we can conclude a positive or negative classification respectively. We also looked into the overlapping and non overlapping cases (with emphasis on the non overlapping case).

8.1.2 Probabilistic Model for Binary Classification

The fundamental random variable that we need to generate or predict is y. Therefore, we model y as a probability of some event. Since it is binary, the probability distribution we can use for our purpose is a Bernoulli Distribution where,

$$P(y = +1) = p \tag{8.1}$$
$$P(y = -1) = 1 - p \tag{8.2}$$

Here p should be high if $w^T x$ is high. However $P(y = +1)$ depends on x. Therefore:

$$P(y = +1) = P(w^T x) \tag{8.3}$$
$$P(y = -1) = 1 - P(w^T x) \tag{8.4}$$

8.1.2.1 Conditions for the Probabilistic Model

Using the properties of probability $0 \leq p \leq 1$, we reach the following **first condition**:

$$P(w^T x) \geq 0 \tag{8.5}$$
$$P(w^T x) \leq 1 \tag{8.6}$$

The **second condition** follows from convexity i.e, the probability function should be convex.

$$P(y = +1) = \frac{1}{1 + e^{-w^T xy}} \Rightarrow 1 - P(y = +1) = P(y = -1) = \frac{1}{1 + e^{w^T xy}} \tag{8.7}$$

From the above we get:

$$P_w(y|x) = \frac{1}{1 + e^{-w^T xy}} \tag{8.8}$$

8.1.2.2 Drawbacks of the Approach

- Data could be complicated
- For applications, we need to take concrete decisions and report probabilities

8.2 Multi-Class Classification

Multi-Class Classification Problem: Instead of y taking only 2 discrete values, we allow y to take K discrete values, i.e, $y = \{1, 2...K\}$. Here, it is important to note that the difference between the two classes is not represented by the difference between the ordinal separation.

8.2.1 Approach for Analysis

We use a multinomial probability distribution, where:

$$P(y = i) = P_i = f(w_i^T x) \tag{8.9}$$

$$P_i = \frac{e^{w_i^T x}}{\sum_{j=1}^{K} e^{w_j^T x}} \tag{8.10}$$

Suppose $P(y = +1) = \frac{1}{1+e^{-w_1^T xy}}$ and $P(y = -1) = \frac{1}{1+e^{-w_{-1}^T xy}}$. We show that $w_1 = -w_{-1}$:

$$w_i = y.w \quad (K = 2) \tag{8.11}$$

$$w_1^T x = -w_{-1}^T x \tag{8.12}$$

$$\Rightarrow w_1 = -w_{-1} \tag{8.13}$$

Therefore we can write the following with precision:

$$P(y = +1) = \frac{1}{1 + e^{-w_1^T xy}} \tag{8.14}$$

$$P(y = -1) = \frac{1}{1 + e^{w_1^T xy}} \tag{8.15}$$

8.3 Stability

8.3.1 Potential Drawbacks of Classification Problems

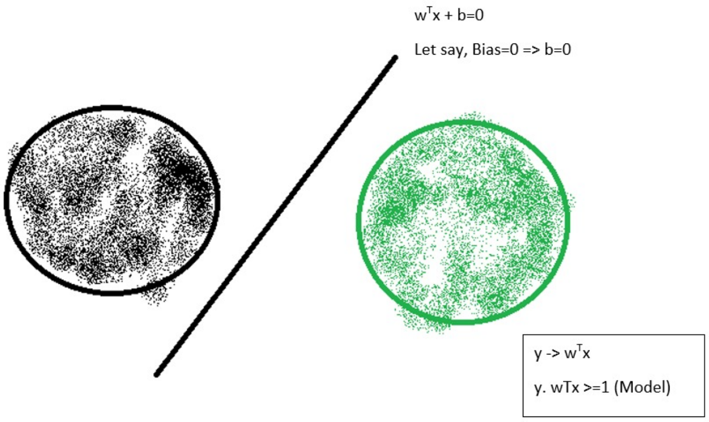

Figure 8.1:

We are forcing the plane to pass through the origin, but these regions could be shifted.

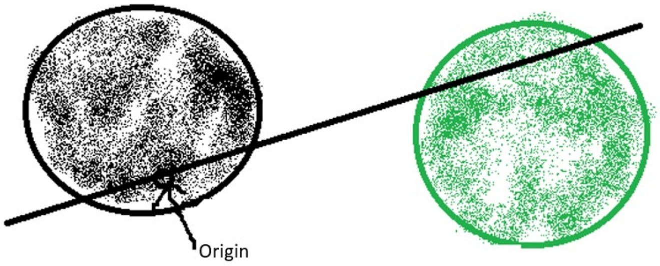

Figure 8.2:

8.3.1.1 Advantages of taking b=0

Let us assume that bias is present

$$w^T x + b = 0 \tag{8.16}$$
$$b \neq 0 \tag{8.17}$$

But now, we add a new point to Class 2 as shown below (figure 8.3).

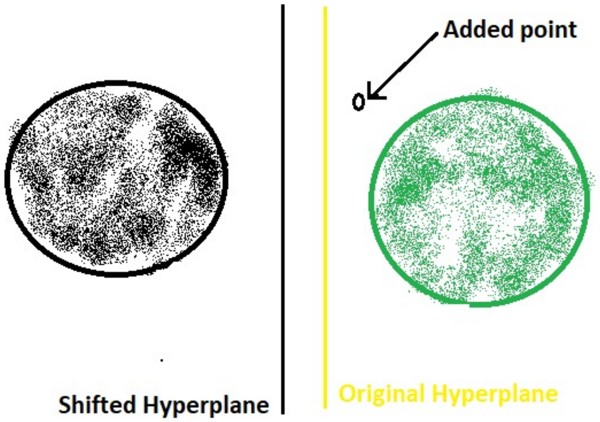

Figure 8.3:

It can be observed that even addition of a single point can lead to shifting of plane by a huge margin.

Explanation: The class 2 shifts more due to addition of the newly added point and thus ultimately the hyper-plane gets shifted more and the following problems can arise:

- Not robust to outliers.
- Have generalisation issues like over-fitting.
- Model is super sensitive to addition of one point. Thus, privacy is at risk. Examples:
  - When a patient data is added, model changes by a huge amount. We can reverse engineer the patient's data.
  - Discussion of privacy issues with respect to data and tech companies like Google, Amazon and Facebook (Meta).

Due to the above issues, the model is NOT stable.

8.3.2 Definition of Stability

Let there be a dataset D, and algorithm A and a singular point a. When a model is stable, the introduction of an additional point must result in a new model that only has a minute difference from the original model. In particular for a the following difference is small ($\emptyset(\frac{1}{|D|})$): $D \rightarrow A \rightarrow A(D)$ and $D \cup a \rightarrow A \rightarrow A(D \cup a)$.

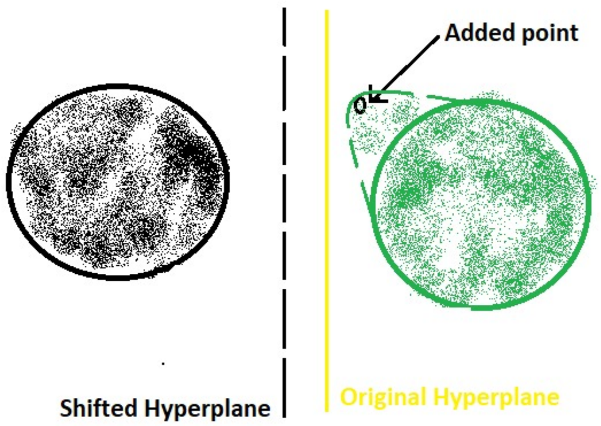

Figure 8.4:

8.3.3 Effect of Bias

If we look at figure 8.2, we notice a considerable amount of incorrect classification. However, there is a bigger problem: the test set error is high. In figure 8.1, however, the test set error is low. Therefore the following inferences can be drawn for $y = sgn(w^T x + b)$:

- $b \neq 0$ improves the training accuracy but not necessarily the test accuracy. It is not stable (not robust to outliers and can result in privacy issues).
- $b = 0$ does not improve training accuracy but may improve test accuracy. With the help of regularization, we can ensure stability and also generalise for real data (avoid over fitting).

In figure 8.5, let the training set be A, B, C and the test set be A', B', C'. Then,

$$\text{error(A', B', C' } |b = 0) \leq \text{error(A', B', C' } |b \neq 0) \text{ - error(A, B, C } |b \neq 0)$$

8.3.4 Summary of the Above Discussion

- If the testing set (A', B', C') is similar to the test set (A, B, C), then $b \neq 0$ is better.
- If the test and training sets are different then $b = 0$ will perform better as it will be more stable. However if this difference is large, then $b = 0$ will NOT help.

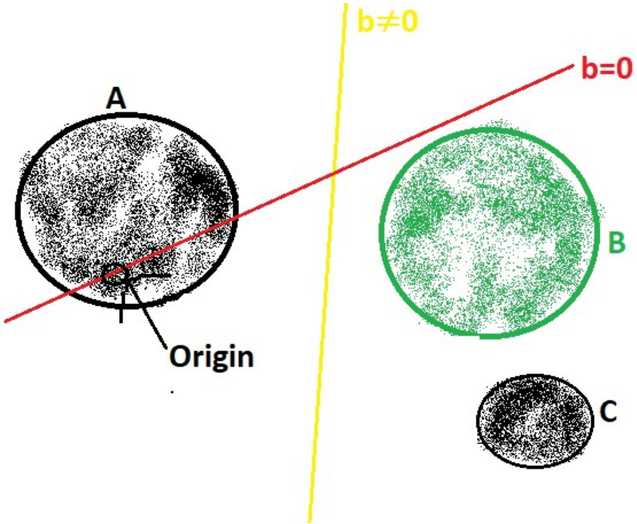

Figure 8.5:

8.4 Summary

In this note we look into the case of multi-class classification based on probabilistic models. We later discuss the concepts of stability and the effect of bias on the stability of the model in terms of shifting hyperplanes, training datasets and test datasets.

8.5 Group Details

- 190100074 M Arvind
- 20D110002 Aarya Ajit Chaudhari
- 18D100007 Bhavini Jeloka
- 190100088 Parag Bajaj
- 200070061 Priyanshi Gupta