Assignment 1 – Pardis Sadatian Moghaddam

Panther ID: 002722641

1. **List at least two loss functions (not mentioned in the lectures) and describe the context in which they will be most useful.**

1. Mean Square Error

The Mean Squared Error **measures how close a regression line is to a set of data points**. To calculate Mean Square Error, we need to take the difference between my predictions and the ground truth, square it, and average it across the whole dataset.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$$

MES = mean squared error

N = number of data points

Yi = observed values

Ŷi = predicted values

2. Binary Cross-Entropy Loss/ Log Loss

This is the most common loss function. The cross-entropy loss decreases as the predicted probability converges to the actual label. It measures the performance of a classification model whose expected output is a probability value between 0 and 1.

$$Logloss = -\frac{1}{N} + \sum_{i=1}^{N}(y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i))$$

Here, pi is the probability of class 1, and (1-pi) is the probability of class 0.

For Multi-Class Classification

$$Logloss = -\frac{1}{N} + \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}\log\left(p_{ij}\right)\right)$$

2. Suppose that X is a random variable taking positive integer values 1,2,3,... Show that E (X) = $\sum_{i=1}^{\infty} \mathrm{P}(\mathrm{X} \geq \mathrm{i})$
Hint: use the relation between P (X=i) , P(X ≥i) and P(X ≥ i+ 1)
We have x1, x2, x3 from the training set X.

E (X) =$\sum_{i=1}^{\infty} \mathrm{P}(\mathrm{X} \geq \mathrm{i}) = P(X=1) + P(X=2) + P(X=3) + \cdots$ **1**
E (X) =$\sum_{i=1}^{\infty} \mathrm{P}(\mathrm{X} \geq \mathrm{i}+1) = P(X=2) + P(X=3) + P(X=4) + \cdots$
E (X) =$\sum_{i=1}^{\infty} \mathrm{P}(\mathrm{X} \geq \mathrm{i}+2) = P(X=3) + P(X=4) + P(X=5) + \cdots$

⇨ If we add all of them E (X) = $P(X=1) + 2 * P(X=2) + 3 * P(X=3) + \cdots$ **2**

E (X) =$\sum_{i=1}^{\infty} x_i \mathrm{P}(x_i) = 1 * P(X=1) + 2 * P(X=2) + 3 * P(X=3) + \cdots$ **3**

We have x1, x2, and x3 from the training set X.

By 1 and 2 and **3**: E (X) =$\sum_{i=1}^{\infty} \mathrm{P}(\mathrm{X} \geq \mathrm{i})$

3. **Let X =R2, Y ={0,1 }, and H be the class of concentric circles with the center at the point (0,0), i.e. each function from H assign the value 1 to all points inside a particular circle and the value 0 – to all points outside that circle. Show that H is PAC-learnable with the sample complexity** $m_{H\,|\varepsilon,\delta|=O(\log\frac{1}{\varepsilon\delta}/\varepsilon)}$

We can try **the fundamental quantitative theorem of Statistical learning**.
We consider H to be a hypothesis class for binary classification concerning 0-1 loss over the domain X; We have VC dimension(H) = d. In here we can shatter the circle in to two parts 0 and 1. H is PAC learnable with the sample

$\mathrm{C1}\frac{d+\log 1/\delta}{\varepsilon} \leq\ m_H(\varepsilon,\delta) \leq \mathrm{C1}\frac{d\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}}{\varepsilon}$

Based on this formula and our problem, we consider the upper bound:

$$m_H(\varepsilon, \delta) \leq \mathrm{C1} \frac{d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}}{\varepsilon}$$

We can make a mathematic operation on the logarithm we have

$$\log a + \log b = \log ab$$

We can rewrite the formula like this:

$$m_H(\varepsilon, \delta) \leq \mathrm{C1} \frac{d \log \frac{1}{\varepsilon\delta}}{\varepsilon}$$

As we discussed, d=1 and C1 are the constant, and we will not consider the constant in big O so we can show:

$$m_{H\,|\varepsilon,\delta|=O(\log\frac{1}{\varepsilon\delta}/\varepsilon)}$$

4. **Suppose that you are designing a machine learning algorithm to predict whether patients have high risk of complications after SARS-CoV-2 . You identified the following relevant features: age.**
(A), body mass index (BMI), time since last vaccination (V), and income (I). You may choose between two hypothesis classes: (a) axis-aligned rectangles in 2-dimensional space spanned by the features BMI and V; (b) axis-aligned hyper rectangles in 4-dimensional space spanned by all features.
What are the pluses and minuses of the alternatives (a) and (b)?
Should the number of training instances influence your choice between them, and if yes, then how?

We are going to use logistic regression for this:

2-dimensional space

Pluses:

If we choose, 2-dimensional space spanned the features BMI and V. The 2-dimensional space is simple and easier to implement. We should only put the elements as the points, and we need to check if the points fit in the rectangles. We can easily visualize it. This means that training the model is more straightforward. On the other hand, we can avoid overfitting. In addition, the decision boundary defined by rectangles is more specific and detailed.

Minuses:

Considering two features, this hypothesis class shows only some of the full complexity of the problems and may lead to lower accuracy, and some of the information may be a loss.

4-dimensional space

Pluses:

Using a 4-dimensional space makes the hypothesis class may better capture the interaction between features. This leads to a more accurate representation of the problem. In addition, it will have high accuracy.

Minuses:

It is more complex and harder to implement. This makes it more to visualize because points should be in 4-dimensional space. Moreover, we need a larger training time, and the risk of overfitting is more. In addition, the risk of training data is limited because we have more data to consider and visualize.

- The number of training instances influences the choice. If the numbers are large, that increases the complexity and might be an overfitting problem. If the numbers are small, then the numbers are limited for computing resources. Overall, it may be more feasible to implement and train.

5. President X uses advisors to make important decisions. For that, he has a pool H of 2000 advisors. Now, X wants to predict the senate votes for an important law. The laws are proposed in a random fashion independently and identically according to some distribution D. After some deliberation, President X selected an expert who correctly predicted outcomes of the previous m = 200 votes. Give a bound on the probability that such an expert incorrectly predicts the next vote. This bound should be correct with 95% confidence.

$\delta = 1 - 0.95 = 0.05$

H = 2000

M = 200

$m \geq \log(H/\delta)/\varepsilon$

$200 \geq \log(2000/0.05)/\varepsilon$

$200 \geq 4.602/\varepsilon$

$\varepsilon = 0.02301$