Advanced Machine Learning: Project 1

1 Loss Functions Not Covered in Lecture

1. Mean Squared Error (MSE)

The Mean Squared Error (MSE) is commonly used for regression tasks. It measures the average squared difference between the actual values and the predicted values. It is defined as:

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

Where:

- N = number of data points
- Y_i = observed (ground truth) value
- $\hat{Y}_i$ = predicted value

Best Use: Regression problems where larger errors should be penalized more heavily due to the squaring term.

2. Binary Cross-Entropy Loss (Log Loss)

Binary Cross-Entropy is widely used in binary classification problems. It measures how well a probabilistic prediction model performs.

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right]$$

Where:

- p_i is the predicted probability of class 1
- $1 - p_i$ is the predicted probability of class 0

Multi-Class Case:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

Best Use: Classification problems with probabilistic outputs.

2 Expected Value Representation via Cumulative Probabilities

Let X be a random variable that takes on positive integer values: $1, 2, 3, \ldots$. We aim to show:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} P(X \geq i)$$

Proof:

Recall:

$$P(X = i) = P(X \geq i) - P(X \geq i + 1)$$

Therefore:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i \cdot P(X = i) = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} P(X = j) = \sum_{j=1}^{\infty} P(X \geq j)$$

Hence, we conclude:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} P(X \geq i)$$

3 PAC Learnability of Concentric Circles

Let $X = \mathbb{R}^2$, $Y = \{0, 1\}$, and $\mathcal{H}$ be the class of all concentric circles centered at the origin. Each hypothesis $h \in \mathcal{H}$ assigns label 1 to points inside the circle and 0 outside.

Goal: Show that $\mathcal{H}$ is PAC-learnable with sample complexity:

$$m_{\mathcal{H}}(\varepsilon, \delta) = O\left(\frac{\log\left(\frac{1}{\varepsilon\delta}\right)}{\varepsilon}\right)$$

Solution

The VC-dimension of $\mathcal{H}$ is $d = 1$ because we can shatter only one point in $\mathbb{R}^2$ with concentric circles.

From the sample complexity bounds for PAC learnability:

$$C_1 \cdot \frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_1 \cdot \frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}$$

Apply logarithmic identity: $\log a + \log b = \log(ab)$

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq C_1 \cdot \frac{\log(1/(\varepsilon\delta))}{\varepsilon}$$

Ignoring constant C_1 in big-O notation, we obtain:

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

4 Modeling Patient Risk After SARS-CoV-2

We want to predict whether patients have high risk of complications using:

- Age (A)
- Body Mass Index (BMI)
- Time since last vaccination (V)
- Income (I)

Hypothesis Class Options:

(a) Axis-aligned rectangles in 2D space spanned by BMI and V.

(b) Axis-aligned hyperrectangles in 4D space spanned by A, BMI, V, and I.

Pros and Cons

2D Model (BMI, V):

- **Pros:** Simple, interpretable, low risk of overfitting, easier to visualize.
- **Cons:** Ignores age and income, potentially reducing predictive accuracy.

4D Model (A, BMI, V, I):

- **Pros:** Captures more feature interactions, potentially higher accuracy.
- **Cons:** Harder to visualize, higher risk of overfitting, requires more training data.

Impact of Training Set Size

Yes, the number of training examples should influence the choice:

- If the training set is small, prefer the 2D model to avoid overfitting.
- If the training set is large, the 4D model could generalize better and yield higher accuracy.

5 Expert Prediction in Voting

President X selects an expert from a pool of $|\mathcal{H}| = 2000$ who correctly predicted $m = 200$ votes. We want to bound the probability ε that this expert will make a mistake on the next vote with confidence $1 - \delta = 0.95$.

Calculation

Given the PAC bound:

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$$

Substitute:

$$200 \geq \frac{\log(2000/0.05)}{\varepsilon} = \frac{\log(40000)}{\varepsilon} \approx \frac{10.6}{\varepsilon}$$

Solving:

$$\varepsilon \leq \frac{10.6}{200} = 0.053$$

Hence, with 95% confidence, the probability that this expert will incorrectly predict the next vote is at most:

$$\boxed{0.053}$$