

Homework 1: Learning Theory and Linear Predictors

CS 6316 Machine Learning

Due on September 22, 2022

Submission Instruction

- For the writing part: please only submit pdf file with name `[ComputingID]-hw01.pdf`. We recommend to use LaTeX and you can find a template on the course webpage. If you are not familiar with LaTeX, using hand writing and scanning it to pdf will also work.
- For the coding part: by default, we will use Python. Your submission should be an iPython notebook file with the name `[ComputingID]-hw01.ipynb`.

Questions (20 points)

1. **The Bayes Predictor** (4 points) For a binary classification problem, if we know the data distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{+1, -1\}$, we can define the Bayes predictor as

$$f_{\mathcal{D}}(\mathbf{x}) = \begin{cases} +1 & \text{if } \mathbb{P}[y = +1 \mid \mathbf{x}] > \frac{1}{2} \\ -1 & \text{if } \mathbb{P}[y = -1 \mid \mathbf{x}] > \frac{1}{2} \end{cases} \quad (1)$$

Note that $\mathbb{P}[y = +1 \mid \mathbf{x}] + \mathbb{P}[y = -1 \mid \mathbf{x}] = 1$. Please show that this is the optimal predictor. In other words, for any predictor h , we have

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(h) \quad (2)$$

2. **Selection of Hypothesis Spaces** (8 points) In lectures, we talked about how to identify the decision boundary using a mixture of Gaussian distributions. As an exercise, please replace the distribution with the following mixture of Gaussian distributions

$$\mathcal{D} = \underbrace{\frac{1}{2}\mathcal{N}(\mathbf{x}; 0, 1)}_{y=-1} + \underbrace{\frac{1}{2}\mathcal{N}(\mathbf{x}; \frac{2}{3}\pi, 0.5)}_{y=+1} \quad (3)$$

Please answer the following questions with the new data distribution

- (a) (2 point) What is the decision boundary of the Bayes predictor b_{Bayes} ? Such as the Bayes predictor can be defined as

$$f_{\mathcal{D}}(x) = \begin{cases} +1 & x > b_{\text{Bayes}} \\ -1 & x < b_{\text{Bayes}} \end{cases} \quad (4)$$

- (b) (1 point) What is the true error of the Bayes predictor, $L_{\mathcal{D}}(f_{\mathcal{D}})$?
- (c) (2 point) With the following hypothesis space \mathcal{H} and the data distribution in equation 3, please find out the best hypothesis $h^* \in \mathcal{H}$ and report the corresponding decision boundary b^*

$$\mathcal{H} = \left\{ \frac{i}{400} : i \in [1200] \right\} \quad (5)$$

- (d) (1 point) What is the true error of h^* , $L_{\mathcal{D}}(h^*)$?
- (e) (1 point) Follow a similar data generation procedure as in the demo code, sample 100 data points from *each* component and label them correspondly. Then, with the same hypothesis space \mathcal{H} in equation 5 and these 200 training examples, please find out the best hypothesis h_S that minimize the empirical error and report the corresponding decision boundary b_S .
- (f) (1 point) What is the true error of h_S , $L_{\mathcal{D}}(h_S)$?
3. **Perceptron algorithm** (3 points) Implementing the Perceptron algorithm with a simple example. The data you need for the implementation is in the file `data.txt`, which is released together with the assignment. Comparing to the pseudocode in our lecture, T was removed from line 3. That is because in practice we do not know the actual value of T . But we can monitor the predictions on all data points and stop the algorithm when the classifier makes correct predictions on all examples.
- 1: **Input:** $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$
 - 2: Initialize $\mathbf{w}^{(0)} = (0, \dots, 0)$
 - 3: **for** $t = 1, 2, \dots$ **do**
 - 4: $i \leftarrow t \bmod m$
 - 5: **if** $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$ **then**
 - 6: $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
 - 7: **end if**
 - 8: **end for**
 - 9: **Output:** the final $\mathbf{w}^{(t)}$
4. **Logistic Regression** (2 points) Given a training set $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, the loss function of logistic regression is defined as

$$L(h_w, S) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)). \quad (6)$$

Please show that the gradient of $L(h_w, S)$ with respect to \mathbf{w} is

$$\frac{dL(h_w, S)}{d\mathbf{w}} = \frac{1}{m} \sum_{i=1}^m \frac{\exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)}{1 + \exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)} \cdot (-y_i \mathbf{x}_i) \quad (7)$$

5. **Linear Regression** (3 points) The loss function of linear regression with ℓ_2 regularization is defined as

$$L_{\ell_2}(h_w, S) = \sum_{i=1}^m (h_w(\mathbf{x}_i) - y_i)^2 + \lambda \|\mathbf{w}\|_2^2 \quad (8)$$

Please show that the solution of this problem, when $\mathbf{A} + \lambda \mathbf{I}$ is invertible, is

$$\mathbf{w} = (\mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{b} \quad (9)$$

where \mathbf{I} is the identify matrix, \mathbf{A} and \mathbf{b} is defined as

$$\mathbf{A} = \sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T \quad \mathbf{b} = \sum_{i=1}^m y_i \mathbf{x}_i \quad (10)$$

Note that $\{\mathbf{x}_i\}$ are column vectors.

Please report your answer in the homework submission and also submit your code with file name as [ComputingID]-hw02.py or [ComputingID]-hw02.ipynb. Without code submission, you will get a 50% deduction on the total points you have on this problem.