Midterm Study Guide

September 27, 2022

1 Foundations of Learning

1. Understanding Machine Learning p.20, Ex.2.1.
   Overfitting of polynomial matching: We have shown that the predictor defined in Equation (2.3) leads to overfitting. While this predictor seems to be very unnatural, the goal of this exercise is to show that it can be described as a thresholded polynomial. That is, show that given a training set $S = (x_i, f(x_i))_{i=1}^m (\mathbb{R}^d \times 0, 1)^m$, there exists a polynomial p_S such that $h_S(x) = 1$ if and only if $p_S(x) \geq 0$, where h_S is as defined in Equation (2.3). It follows that learning the class of all thresholded polynomials using the ERM rule may lead to overfitting.

2. Understanding Machine Learning p.20, Ex.2.2.
   For a fixed classifier h in the class of binary classifier $\mathcal{H}$ that operate on domain generated $\mathcal{X}$ generated according to an unknown distribution $\mathcal{D}$ and labeled by f, show that the expected value of $L_s(h)$ over the choice of training sequences $S\,|_x$ equals to $L_{(\mathcal{D},f)}(h)$, specifically:

$$\mathop{\mathbb{E}}_{S|_x \sim \mathcal{D}^m}[L_s(h)] = L_{(\mathcal{D},f)}(h)$$

3. Understanding Machine Learning, p.22, Ex. 2.3.
   Axis aligned rectangles: An axis aligned rectangle classifier in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain rectangle. Formally, given real numbers $a_1 \leq b_1, a_2 \leq b_2$, define the classifier $h(a_1, b_1, a_2, b_2)$ by

$$h(a_1, b_1, a_2, b_2)(x_1, x_2) = \begin{cases} 1 \text{ if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 \text{ otherwise} \end{cases}$$

The class of all axis aligned rectangles in the plane is defined as $H^2_{rec} = h(a_1, b_1, a_2, b_2) : a_1 \leq b_1,$ and $a_2 \leq b_2$. Note that this is an infinite size hypothesis class. Throughout this exercise we rely on the realizability assumption.

(a) Let A be the algorithm that returns the smallest rectangle enclosing all positive examples in the training set. Show that A is an ERM.

(b) Show that if A receives a training set of size $\leq 4 \log \frac{4}{\delta}$ then, with probability of ϵ at least $1 - \delta$ it returns a hypothesis with error of at most ϵ.

Hint: Fix some distribution $\mathcal{D}$ over $\mathcal{X}$, let $R^\star = R(a_1^\star, b_1^\star, a_2^\star, b_2^\star)$ be the rectangle that generates the labels, and let f be the corresponding hypothesis. Let $a_1 \leq a_1^\star$ be a number such that the probability mass (with respect to $\mathcal{D}$) of the rectangle $R1 = R(a_1^\star, b_1^\star, a_2^\star, b_2^\star)$ is exactly $\frac{\epsilon}{4}$. Similarly, let b_1, a_2, b_2 be numbers such that the probability masses of the rectangles $R2 = R(b_1, b_1^\star, a_2^\star, b_2^\star)$, $R3 = R(a_1^\star, b_1^\star, a_2^\star, a_2)$, $R4 = R(a_1^\star, b_1^\star, b_2, b_2^\star)$ are all exactly $\frac{\epsilon}{4}$. Let $R(S)$ be the rectangle returned by A. See illustration in Figure 2.2.

- Show that $R(S) \subseteq R^\star$.
- Show that if S contains (positive) examples in all of the rectangles $R1, R2, R3, R4$, then the hypothesis returned by A has error of at most ϵ.
- For each $i \in 1, \ldots, 4$, upper bound the probability that S does not contain an example from R_i .
- Use the union bound to conclude the argument.

(c) Repeat the previous question for the class of axis aligned rectangles in .

(d) Show that the runtime of applying the algorithm A mentioned earlier is polynomial in $d, \frac{1}{\epsilon}$, and in $\log \frac{1}{\delta}$.

2 PAC Learnability

1. Understanding Machine Learning p.28, Ex.3.1
   Monotonicity of Sample Complexity:
2. Understanding Machine Learning p.29, Ex.3.2
3. Understanding Machine Learning p.29, Ex.3.3
4. Understanding Machine Learning p.29, Ex.3.4
5. Understanding Machine Learning p.29, Ex.3.5
6. Understanding Machine Learning p.30, Ex.3.6
7. Understanding Machine Learning p.30, Ex.3.7
   (*)The Bayes optimal predictor:
8. Understanding Machine Learning p.30, Ex.3.8
9. Understanding Machine Learning p.30, Ex.3.9

3 Linear Algebra

1. Linear Algebra with Applications, pp.192-194
   Let $L(\mathbf{x}) = (2x_1, x_1 + x_2)^T$ be a linear transformation in $\mathbb{R}^2 \to \mathbb{R}^2$, $\mathbf{u_1} = [1, 1]^T$ and $\mathbf{u_2} = [-1, 1]^T$.
   (a) Find the matrix representation, A, of L with respect to the standard basis in $\mathbb{R}^2$.
   (b) Compute $L(\mathbf{u_1})$ and $L(\mathbf{u_2})$ using A.
   (c) Find the transition matrix, U, from the standard basis to basis $\{\mathbf{u_1}, \mathbf{u_2}\}$.

(d) Find the transition matrix from basis $\{\mathbf{u_1}, \mathbf{u_2}\}$ back to the standard basis.

(e) Find the coordinates of $L(\mathbf{u_1})$ and $L(\mathbf{u_2})$ in the basis $\{\mathbf{u_1}, \mathbf{u_2}\}$.

(f) Express $L(\mathbf{u_1})$ and $L(\mathbf{u_2})$ in the basis $\{\mathbf{u_1}, \mathbf{u_2}\}$ as linear combinations of $\mathbf{u_1}$ and $\mathbf{u_2}$.

(g) Give the matrix representation, B, of L with respect to the basis $\{\mathbf{u_1}, \mathbf{u_2}\}$.

(h) Write B using the expression in part e. Use this to formulate the similarity equation between matrices A and B.

4 Optimization Theory

5 Linear Learning Models

6 Principal Component Analysis

7 Curse of Dimensionality

8 Bayesian Decision Theory

9 Parameter Estimation: MLE

10 Parameter Estimation: MAP & Naive Bayes

11 Logistic Regression

12 Kernel Density Estimation

13 Support Vector Machines