COMP9417 Machine Learning and Data Mining

Lectures: Learning Theory
Topic: Questions from lecture topics
Last revision: Wed 5 Aug 2020 23:21:05 AEST

Introduction

Some questions and exercises from the course lectures covering theoretical aspects of machine learning independent of particular algorithms. We start by generating a simple concept learning algorithm that is provably correct, then look at some of the ideas used to prove sample complexity in the PAC learning setting, then cover some results for the VC dimension of some concept classes, and finish with an example of the mistake bounds for aan attribute-efficient linear threshold classifier.

This tutorial note, together with the corresponding lecture notes, are intended to contain all the relevant material to enable you to answer all the questions, but you may wish to refer to the cited textbooks, for example [Blum et al., 2019], a draft copy of which is freely available online at `http://www.cs.cornell.edu/jeh/book%20no%20so;utions%20March%202019.pdf` for additional background.

Question 1 ([Blum et al., 2019]) To get a sense of how learning theory characterises sample complexity we start by formulating a simple *consistent learner* to learn *disjunctions*, i.e., Boolean OR functions, of d variables. Recall that a consistent learner is just one that makes no mistake on the training data. The target concept c is assumed to be expressed as a disjunction of literals, where a literal is defined as some feature x_i being true (having value 1).

So an instance is just a set of literals, such as $\{\mathbf{x}|x_1 = 1 \vee x_3 = 1 \vee x_8 = 1\}$. For example, if the target concept was to distinguish between spam and non-spam emails, the presence in an email of any of the features x_1, x_3 or x_8 would be enough to classify it as spam, whereas the absence of all of them would mean non-spam.

Question 1a) The hypothesis space H is the set of all disjunctions of d features. What is the size of this hypothesis space ?

Question 1b) Give an algorithm for a consistent learner for such disjunctive concepts from a set of labelled noise-free training examples S. *HINT*: try adapting the basic approach of the FIND-S algorithm to learn conjunctive concepts shown on slide 44 of the lecture notes.

Question 1c) Outline the steps in a proof that your disjunctive concept learning algorithm will find a consistent hypothesis h, i.e., that the error on sample S $error_S(h) = 0$.

Question 1d) Analyse the sample complexity of the consistent learner for disjunctive concepts in the PAC learning setting, i.e., use the formula from slide 23 in the lecture notes.

Question 2 Consider how you could prove the theorem bounding the probability that a consistent learner will output a hypothesis h with $error(h) \geq \epsilon$ that appears on slides 19–23 of the lecture notes.

Theorem:

If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independently drawn random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that the version space with respect to H and D is not ϵ-exhausted (with respect to c) is less than

$$|H|e^{-\epsilon m}$$

Background Proofs of this theorem are in Chapter 7 of [Mitchell, 1997] and Chapter 5 of [Blum et al., 2019]. These proofs use the following facts:

- For events $A_1, A_2, \ldots, A_n$, the probability of the union of these events $Pr(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} Pr(A_i)$ (this is called the "union bound")
- If $0 \leq \epsilon \leq 1$ then $(1 - \epsilon) \leq e^{-\epsilon}$ (from Taylor series)

Question 3 Refer to the VC dimension example for linear classifiers in the 2-dimensional x, y plane of slides 39–41 of the lecture notes. Answer the following:

1. give an intuitive argument for why the VC dimension must be at least 3;
2. suppose you have a set of 3 points that are collinear – does that change your argument ?
3. can the VC dimension be 4 ?

Question 4 With reference to slides 52–54 of the lecture notes, outline a version of the HALVING ALGORITHM for Boolean functions, and wih reference to it give the worst-case mistake bounds, and an intuitive explanation of why this bound holds. Now repeat this for the best-case performance!

Background Key points about the mistake bounds framework:

- this is intuitively based on a "learning curve" idea
- it is based on an online-learning framework, but can be adapted for batch learning too
- it is closely related to PAC learning, boosting, and other theoretical frameworks

Question 5 Work through applying the WINNOW 2 algorithm on slides 55–57 of the lecture notes to the examples below *in the order in which they appear.*

Use the settings: $\alpha = 2$, $\theta = 2$; and initialise all weights $w_i = 1$. Show all predictions, and whether a mistake has occurred or not. When the algorithm has passed through the examples, do you think the target concept has been learned ?

Example	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	$\mathbf{x_5}$	$\mathbf{x_6}$	$\mathbf{x_7}$	$\mathbf{x_8}$	$\mathbf{x_9}$	$\mathbf{x_{10}}$	**Class**
1)	0	1	0	1	0	1	1	0	1	1	$\oplus$
2)	1	0	1	0	1	0	1	1	1	0	$\ominus$
3)	0	0	0	1	0	0	0	0	1	0	$\oplus$
4)	1	0	0	0	1	1	1	1	1	0	$\ominus$
5)	1	0	1	0	1	1	1	1	0	1	$\ominus$

References

[Blum et al., 2019] Blum, A., Hopcroft, J., and Kannan, R. (2019). *Foundations of Data Science*. Cambridge University Press.

[Mitchell, 1997] Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, New York.