Machine Learning, Spring 2019

Homework 1

Wang Yuzhen
2018E8018461008

1 Preliminaries

The weight update rule $\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$ has the nice interpretation that it move in the direction of classifying $\mathbf{x}(t)$ correctly.

(a) Show that $y(t)\mathbf{w}^T(t)\mathbf{x}(t) < 0$. [*Hint*: $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$.] (5 points)

(b) Show that $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^T(t)\mathbf{x}(t)$. [*Hint*: Use update rule.] (5 points)

(c) As far as classifying $\mathbf{x}(t)$ is concerned, argue that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move 'in the right direction'. (5 points)

Solution==========

(a) When $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$
y=-1 but $\mathbf{w}^T(t)\mathbf{x}(t) > 0$;
y=1 but $\mathbf{w}^T(t)\mathbf{x}(t) < 0$;
one equation can written as $y(t)\mathbf{w}^T(t)\mathbf{x}(t) < 0$.

(b) $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t)$
$= y(t)[\mathbf{w}^T(t) + y(t)\mathbf{x}(t)]\mathbf{x}(t)$
$= y(t)\mathbf{w}^T(t)\mathbf{x}(t) + y^2||x||^2$
$> y(t)\mathbf{w}^T(t)\mathbf{x}(t)$.

(c) $||\mathbf{w}(t+1)||^2$
$= ||\mathbf{w}(t) + y^t\mathbf{x}(t)||^2$
$= ||\mathbf{w}(t)||^2 + y^2||x||^2 + 2\mathbf{w}(t)y^t\mathbf{x}(t)$
$< ||\mathbf{w}(t)||^2 + ||x(t)||^2$
If $y(t)\mathbf{w}^T(t)\mathbf{x}(t) < 0$, use the weight update rule $\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$.
After finte steps, the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move 'in the right direction'.
==========

2 Understanding the law of large numbers

The Hoeffding Inequality is one form of the *law of large numbers*. One of the simple forms of that law is the *Chebyshev Inequality*, which you will prove here.

(a) If t is a non-negative random variable, prove that for any $\alpha > 0$, $\mathbb{P}[t \geq \alpha] \leq \mathbb{E}(t)/\alpha$. (5 points)

(b) If u is any random variable with mean μ and variance σ^2, prove that for any $\alpha > 0$, $\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$. (5 points)

(c) If $u_1, \ldots, u_N$ are iid random variables, each with mean μ and variance σ^2, and $u = \frac{1}{N}\sum_{n=1}^{N} u_n$, prove that for any $\alpha > 0$,

$$\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}\,.$$

(10 points)

Solution==========

(a) $\mathbb{P}[t > \alpha] = \mathbb{E}[\mathbb{L}[t > \alpha]] = \mathbb{E}[t/\alpha] = \mathbb{E}(t)/\alpha$

(b) $\mathbb{P}[t \geq \alpha] \leq \mathbb{E}(t)/\alpha$,
$\mathbb{E}((u-\mu)^2) = \sigma^2$
$\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$.

(c) $Var[\frac{1}{N}\sum_{n=1}^{N} u_n] = \frac{1}{N^2}[Var[u_1] + ... + Var[u_n]]$
$= \frac{\sigma^2}{N}$
And, $\mathbb{P}[t \geq \alpha] \leq \mathbb{E}(t)/\alpha$,
Then

$$\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}\,.$$

==========

3 Probability and independence

Consider a sample of 10 marbles drawn independently from a bin that holds red and green marbles. The probability of a red marble is μ. For $\mu = 0.05, \mu = 0.5$, and $\mu = 0.8$, compute the probability of getting no red marbles ($v = 0$) in the followig cases.

(a) We draw only one such sample. Compute the probability that $v = 0$. (5 points)

(b) We draw 1000 independent samples. Compute the probability that (at least) one of the samples has $v = 0$. (5 points)

(c) Repeat (b) for 1000000 independent samples. (10 points)

Solution==========

(a) $P_0 = (1-u)^{10}$
[0.05] $(1-0.05)^{10} = 0.5987$
[0.5] $(1-0.5)^{10} = 9.7656e-04$
[0.8] $(1-0.8)^{10} = .0240e-07$

(b) $P_1 = 1-(1-P_0)^{1000}$

(c) $P_1 = 1-(1-P_0)^{1000000}$
==========

4 Two learning algorithms

We are given a data set $\mathcal{D}$ of 25 training examples from an unknown target target function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{-1, +1\}$. To learn f, we use a simple hypothesis set $\mathcal{H} = \{h_1, h_2\}$ where h_1 is the constant $+1$ function and h_2 is the constant -1.

We consider two learning algorithms, S (smart) and C (crazy). S chooses the hypothesis that agrees the most with $\mathcal{D}$ and C chooses the other hypothesis deliberately. Let us see how these algorithms perform out of sample from the deterministic and probabilistic points of view. Assume in the probabilistic view that there is a probability distribution on $\mathcal{X}$, and let $\mathbb{P}[f(\mathbf{x}) = +1] = p$

(a) Can S produce a hypothesis that is guaranteed to perform better than random on any point outside $\mathcal{D}$? (5 points)

(b) Assume for the rest of the exercise that all the examples in $\mathcal{D}$ have $y_n = +1$. Is it possible that the hypothesis that C produces turns out to be better than the hypothesis that S produces? (5 points)

(c) If $p = 0.9$, what is the probability that S will produce a better hypothesis than C? (5 points)

(d) Is there any value of p for which it is more likely than not that C will produce a better hypothesis than S? (5 points)

Solution==========

(a) No, it can't.

(b) Yes. It is possible that the hypothesis that C produces turns out to be better than the hypothesis that S produces. We don't know performance on any point outside.

(c) $P(0.9 > 0.1) = 1$

(d) $If \quad P(p > 1-p) \quad is \quad , possible, p < 0.5$ ==========

5 Learning a target function

Consider a Boolean target function over a three-dimensional input space $\mathcal{X} = \{0, 1\}^3$. We are given a data set $\mathcal{D}$ of five examples represented in the table below. We denote the binary output by o/• for visual clarity, where $y_n = f(\mathbf{x}_n)$ for $n = 1, 2, 3, 4, 5$. (Please refer to $P28 - 29$ of "Learning from data" for more details of the problem)

Solution==========

x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
000	o	o	o	o	o	o	o	o	o	o
001	•	•	•	•	•	•	•	•	•	•
010	•	•	•	•	•	•	•	•	•	•
011	o	o	o	o	o	o	o	o	o	o
100	•	•	•	•	•	•	•	•	•	•
101		?	o	o	o	o	•	•	•	•
110		?	o	o	•	•	o	o	•	•
111		?	o	•	o	•	o	•	o	•

Figure 1: Learning a target function

(a) a all three points f8 two of them f4,f6,f7 one of them f2,f3,f5

(b) b all three points f1 two of them f2,f3,f5 one of them f4,f6,f7

(c) c all three points f2 two of them f1,f4,f6 one of them f3,f5,f8

(d) d all three points f7 two of them f3,f5,f8 one of them f1,f4,f6

==========