CS 337 (Spring 2019): Class Test 1

Instructor: Shivaram Kalyanakrishnan

2.00 p.m. – 3.15 p.m., February 8, 2019, LH 301

Total marks: 20

Note. Provide brief justifications and/or calculations along with each answer to illustrate how you arrived at the answer.

Question 1. Describe two factors that have contributed to the dramatic growth of AI in the last 5–10 years. Keep your answer brief: ideally 2–3 lines. [2 marks]

Question 2. Answer these questions relating to the Perceptron Learning Algorithm, as discussed in class. Assume that the input data set comprises two classes which are separable using an origin-centred hyperplane.

2a. Recall that the Perceptron Learning Algorithm is free to pick an *arbitrary* point from the currently misclassified set at each iteration, and update the weight vector based on that point. Hence, for a given data set, we can end up with different output weight vectors depending on how we resolve the choice at each iteration. Concretely, let $\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^L$ be the *output* weight vectors produced by L separate runs of the Perceptron Learning Algorithm. As we have already shown, each of these weight vectors guarantees perfect separation of the training data points. Now consider the "average" of the output vectors:

$$\mathbf{w}^{\text{avg}} = \frac{\mathbf{w}^1 + \mathbf{w}^2 + \cdots + \mathbf{w}^L}{L}.$$

Is $\mathbf{w}^{\text{avg}}$ also guaranteed to separate the training points perfectly? Prove that your answer is correct. [2 marks]

2b. Consider a change to the Perceptron Learning Algorithm, wherein we use a "learning rate" $\alpha^k = \frac{1}{k}$ for $k \geq 1$. In other words, the update made by this variant at the k-th iteration is

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \alpha^k y^j \mathbf{x}^j,$$

with $(\mathbf{x}^j, y^j)$ being the arbitrarily-chosen misclassified point. By contrast, in class we had used a constant learning rate of 1.

The use of a harmonically-annealed learning rate is common in algorithms such as gradient descent. How does it affect the Perceptron Learning Algorithm? Does the algorithm still converge; if so, does it still yield a separating hyperplane? Prove that your answer is correct. [5 marks]

Question 3. We consider a 1-dimensional example of gradient descent. For $w \in \mathbb{R}$, let

$$\text{Error}(w) = 3w^4 - 4w^3 - 12w^2 + 50.$$

3a. What is $G(w) = \nabla_w \text{Error}(w)$? [1 mark]

3b. Consider a procedure that begins with some initial guess $w^0 \in \mathbb{R}$, and progressively obtains iterates through gradient descent: for $t \geq 0$,

$$w^{t+1} \leftarrow w^t - \frac{1}{t+1} G(w^t).$$

Draw a plot with w^0 on the x axis and $\lim_{t \to \infty} \text{Error}(w^t)$ on the y axis. [4 marks]

3c. What is $G^2(w) = \nabla_w G(w)$? Based on your answer to 3b, suggest why this function might be useful to compute. [2 marks]

Question 4. Consider a data set containing every possible tuple of three binary variables x_1, x_2, and x_3. The label associated with each tuple is encoded by the decision tree T_1 shown below. Observe that the variables and the labels both take values in $\{0, 1\}$.

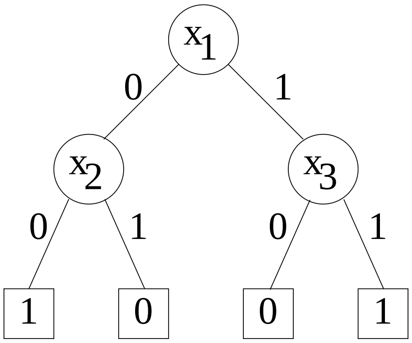

Figure 1: Decision tree T_1. Internal nodes are shown as circles, and leaves as squares.

Draw a decision tree T_2 that assigns each possible (x_1, x_2, x_3)-tuple the same label as assigned by T_1, but which splits on the variable x_2 at its root node. Use the least number of internal nodes possible to construct T_2, and argue why you cannot reduce this number further. [4 marks]

Solutions

1. The growth of the Internet has made it possible to collect large amounts of data, which it is also possible now to store cheaply. Computing and memory have also become orders of magnitude more efficient in the last few years, allowing for algorithms to process stored data. Cameras and other sensors—also ubiquitous and cheap!—have made different types of data available for processing. From a technical standpoint, the maturing of machine learning as a field has led to many "off-the-shelf" solutions and libraries for AI. The resurgence of neural networks as an effective model for tasks in domains such as vision and speech has also been the reason behind many success stories.

2a. We know from our proof in class that for each $i \in \{1, 2, \ldots, n\}$ and $l \in \{1, 2, \ldots, L\}$,

$$y^i(\mathbf{w}^l \cdot \mathbf{x}^i) > 0.$$

It follows that for each $i \in \{1, 2, \ldots, n\}$,

$$y^i(\mathbf{w}^{\text{avg}} \cdot \mathbf{x}^i) = \frac{y^i(\mathbf{w}^1 \cdot \mathbf{x}^{\mathbf{i}}) + y^i(\mathbf{w}^2 \cdot \mathbf{x}^{\mathbf{i}}) + \cdots + y^i(\mathbf{w}^L \cdot \mathbf{x}^{\mathbf{i}})}{L} > 0,$$

implying that $\mathbf{w}^{\text{avg}}$ also achieves perfect separation of the data.

2b. We follow essentially the same steps as done in our original proof, albeit with the different learning rate. First, we observe

$$\begin{aligned}
\mathbf{w}^{k+1} \cdot \mathbf{w}^\star &= (\mathbf{w}^k + \alpha^k y^j \mathbf{x}^j) \cdot \mathbf{w}^\star \\
&= \mathbf{w}^k \cdot \mathbf{w}^\star + \alpha^k y^j (\mathbf{x}^j \cdot \mathbf{w}^\star) \\
&\geq \mathbf{w}^k \cdot \mathbf{w}^\star + \alpha^k \gamma.
\end{aligned}$$

It follows by induction that $\mathbf{w}^{k+1} \cdot \mathbf{w}^\star \geq (\alpha^1 + \alpha^2 + \cdots + \alpha^k)\gamma$. Since $\mathbf{w}^{k+1} \cdot \mathbf{w}^\star \leq \|\mathbf{w}^{k+1}\| \|\mathbf{w}^\star\| = \|\mathbf{w}^{k+1}\|$, we get

$$\|\mathbf{w}^{k+1}\| \geq (\alpha^1 + \alpha^2 + \cdots + \alpha^k)\gamma. \tag{1}$$

We can also *upper-bound* $\|\mathbf{w}^{k+1}\|$ as follows.

$$\begin{aligned}
\|\mathbf{w}^{k+1}\|^2 &= \|\mathbf{w}^k + \alpha^k y^j \mathbf{x}^j\|^2 \\
&= \|\mathbf{w}^k\|^2 + \|\alpha^k y^j \mathbf{x}^j\|^2 + 2(\mathbf{w}^k \cdot \mathbf{x}^j)\alpha^k y^j \\
&= \|\mathbf{w}^k\|^2 + (\alpha^k)^2 \|\mathbf{x}^j\|^2 + 2(\mathbf{w}^k \cdot \mathbf{x}^j)\alpha^k y^j \\
&\leq \|\mathbf{w}^k\|^2 + (\alpha^k)^2 \|\mathbf{x}^j\|^2 \\
&\leq \|\mathbf{w}^k\|^2 + (\alpha^k)^2 R^2,
\end{aligned}$$

from which it follows by induction that

$$\|\mathbf{w}^{k+1}\|^2 \leq ((\alpha^1)^2 + (\alpha^2)^2 + \cdots + (\alpha^k)^2)R^2. \tag{2}$$

For our particular choice of sequence $\alpha^k = \frac{1}{k}$, we observe that

1. $\alpha^1 + \alpha^2 + \cdots + \alpha^k > \ln(k)$, and
2. there is a constant C such that $(\alpha^1)^2 + (\alpha^2)^2 + \cdots + (\alpha^k)^2 < C$.

It follows that $(\gamma \ln(k))^2 \leq \|\mathbf{w}^{k+1}\|^2 \leq CR^2$, which implies $k \leq \exp(\sqrt{C}\frac{R}{\gamma})$. Hence, the algorithm can only make a finite number of iterations; by construction, termination implies correctness.

3a. $G(w) = \nabla_w(3w^4 - 4w^3 - 12w^2 + 50) = 12w^3 - 12w^2 - 24w$.

3b. It is easy to see that $G(w)$ factorises as $12w(w-2)(w+1)$, implying that $Error(w)$ has its local optima (maxima or minima) at -1, 0, and 2. By plotting $Error(w)$, we observe that indeed -1 and 2 are local minima, while 0 is a local maximum.

It follows that if we perform gradient descent with a "small enough" learning rate,

1. starting with $w^0 > 0$ should eventually converge to $w^\infty = 2$ (and $Error(-2) = 34$);
2. starting with $w^0 < 0$ should eventually converge to $w^\infty = -1$ (and $Error(-2) = 45$);
3. starting with $w^0 = 0$ will lead to $w^\infty = 0$ (and $Error(0) = 50$).

Unfortunately, there is a **bug** in the question, wherein the learning rate used is not small enough compared to the gradient.[1] Hence, although starting points in the vicinity of the local minima will converge to these minima (and starting at the local maximum will keep the process there for ever), starting from other points could take the process through hard-to-characterise sequences, and in fact even to divergence.

3c. $G^2(w) = \nabla_w(12w^3 - 12w^2 - 24w) = 36w^2 - 24w - 24$. This second derivative lets us determine whether a given local optimum is a local maximum or a local minimum. Observe that $G^2(-1)$ and $G^2(2)$ are positive, while $G^2(0)$ is negative: implying that $Error$ achieves local minima at -1 and 2, and a local maximum at 0. In the unusual but plausible event that we initialise the procedure at a local maximum, note that the gradient will be 0, and thus we would have converged. Knowing the second derivative would inform us whether we are indeed at a local minimum, or whether we can do better by starting with a small perturbation of the initial point.

4. The labels assigned by T_1 to data points are as follows.

x_1	x_2	x_3	y
0	0	0	1
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

We are to replicate the same labelling with T_2, which has x_2 at its root. The left subtree of T_2, containing all points with $x_2 = 0$, cannot be represented with just one node since the labels of the four points cannot all be predicted correctly based on either x_1 or x_3 alone. Hence, the left subtree needs at least two nodes. For the same reason, the right subtree also needs at least two nodes. It can be seen that all four trees shown below use exactly two nodes in each of the left and right subtrees, and they classify all the data points exactly as done by T_1. Any one of them can be provided as the answer.

[1] We acknowledge Utkarsh Gupta for pointing out this bug.

Possibility 1 for T_2 (all paths of tree shown).

- $x_2 \xrightarrow{0} x_1 \xrightarrow{0} 1$.
- $x_2 \xrightarrow{0} x_1 \xrightarrow{1} x_3 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{0} x_1 \xrightarrow{1} x_3 \xrightarrow{1} 1$.
- $x_2 \xrightarrow{1} x_1 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_1 \xrightarrow{1} x_3 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_1 \xrightarrow{1} x_3 \xrightarrow{1} 1$.

Possibility 2.

- $x_2 \xrightarrow{0} x_1 \xrightarrow{0} 1$.
- $x_2 \xrightarrow{0} x_1 \xrightarrow{1} x_3 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{0} x_1 \xrightarrow{1} x_3 \xrightarrow{1} 1$.
- $x_2 \xrightarrow{1} x_3 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_3 \xrightarrow{1} x_1 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_3 \xrightarrow{1} x_1 \xrightarrow{1} 1$.

Possibility 3.

- $x_2 \xrightarrow{0} x_3 \xrightarrow{0} x_1 \xrightarrow{0} 1$
- $x_2 \xrightarrow{0} x_3 \xrightarrow{1} x_1 \xrightarrow{1} 0$.
- $x_2 \xrightarrow{0} x_3 \xrightarrow{1} 1$.
- $x_2 \xrightarrow{1} x_1 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_1 \xrightarrow{1} x_3 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_1 \xrightarrow{1} x_3 \xrightarrow{1} 1$.

Possibility 4.

- $x_2 \xrightarrow{0} x_3 \xrightarrow{0} x_1 \xrightarrow{0} 1$
- $x_2 \xrightarrow{0} x_3 \xrightarrow{1} x_1 \xrightarrow{1} 0$.
- $x_2 \xrightarrow{0} x_3 \xrightarrow{1} 1$.
- $x_2 \xrightarrow{1} x_3 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_3 \xrightarrow{1} x_1 \xrightarrow{0} 0$.
- $x_2 \xrightarrow{1} x_3 \xrightarrow{1} x_1 \xrightarrow{1} 1$.