Advanced Learning Models
First Homework 2019-2020

Xavier Alameda-Pineda & Julien Mairal

Due date January 10th, 2020

Provide a latex-generated PDF to `xavier.alameda-pineda@inria.fr` AND `julien.mairal@inria.fr` by the due date. Delays will affect the grade.

---

1 Neural Networks

Let $\mathbf{X} = (x_{ij})_{ij},\ i,j \in \{1,\dots,5\}$ denote the input of a convolutional layer with no bias. Let $\mathbf{W} = (w_{ij})_{ij},\ i,j \in \{1,\dots,3\}$ denote the weights of the convolutional filter. Let $\mathbf{Y} = (y_{ij})_{ij},\ i \in \{1,\dots,I\}, j \in \{1,\dots,J\}$ denote the output of the convolution operation.

1. What is the output size (i.e. values of I and J) if:
   (a) the convolution has no padding and no stride?
   (b) the convolution has stride 1 and no padding?
   (c) the convolution has no stride and padding 2?
2. Let us suppose that we are in situation 1.(b) (i.e. stride 1 and no padding). Let us also assume that the output of the convolution goes through a ReLU activation, whose output is denoted by $\mathbf{Z} = (z_{ij})_{ij},\ i \in \{1,\dots,I\}, j \in \{1,\dots,J\}$:
   (a) Derive the expression of the output pixels x_{ij} as a function of the input and the weights.
   (b) How many multiplications and additions are needed to compute the output (the forward pass)?
3. Assume now that we are provided with the derivative of the loss w.r.t. the output of the convolutional layer $\partial\mathcal{L}/\partial z_{ij}, \forall\, i \in \{1,\dots,I\}, j \in \{1,\dots,J\}$:
   (a) Derive the expression of $\partial\mathcal{L}/\partial x_{ij}, \forall\, i,j \in \{1,\dots,5\}$.
   (b) Derive the expression of $\partial\mathcal{L}/\partial w_{ij}, \forall\, i,j \in \{1,\dots,3\}$.

Let us now consider a fully connected layer, with two input and two output neurons, without bias and with a sigmoid activation. Let $x_i, i = 1,2$ denote the inputs, and $z_j, j = 1,2$ the output. Let w_{ij} denote the weight connecting input i to output j. Let us also assume that the gradient of the loss at the output $\partial\mathcal{L}/\partial z_j, j = 1,2$ is provided.

4. Derive the expressions for the following derivatives:
   (a) $\dfrac{\partial\mathcal{L}}{\partial x_i}$
   (b) $\dfrac{\partial\mathcal{L}}{\partial w_{ij}}$
   (c) $\dfrac{\partial^2\mathcal{L}}{\partial w_{ij}^2}$
   (d) $\dfrac{\partial^2\mathcal{L}}{\partial w_{ij} w_{i'j'}}, i \neq i', j \neq j'$
   (e) The elements in (c) and (d) are the entries of the Hessian matrix of $\mathcal{L}$ w.r.t the weight vector. Imagine now that storing the weights of a network requires 40 MB of disk space: how much would it require to store the gradient? And the Hessian?

2 Conditionally positive definite kernels

Let $\mathcal{X}$ be a set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *conditionally positive definite* (c.p.d.) if and only if it is symmetric and satisfies:

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

for any $n \in \mathbb{N}$, $x_1, x_2, \ldots, x_n \in \mathcal{X}^n$ and $a_1, a_2, \ldots, a_n \in \mathbb{R}^n$ with $\sum_{i=1}^{n} a_i = 0$.

1. Show that a positive definite (p.d.) function is c.p.d.

2. Is a constant function p.d.? Is it c.p.d.?

3. If $\mathcal{X}$ is a Hilbert space, then is $k(x, y) = -||x - y||^2$ p.d.? Is it c.p.d.?

4. Let $\mathcal{X}$ be a nonempty set, and $x_0 \in \mathcal{X}$ a point. For any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the function defined by:

$$\tilde{k}(x, y) = k(x, y) - k(x_0, x) - k(x_0, y) + k(x_0, x_0).$$

Show that k is c.p.d. if and only if $\tilde{k}$ is p.d.

5. Let k be a c.p.d. kernel on $\mathcal{X}$ such that $k(x, x) = 0$ for any $x \in \mathcal{X}$. Show that there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that, for any $x, y \in \mathcal{X}$,

$$k(x, y) = -||\Phi(x) - \Phi(y)||^2.$$

6. Show that if k is c.p.d., then the function $\exp(tk(x, y))$ is p.d. for all $t \geq 0$

7. Conversely, show that if the function $\exp(tk(x, y))$ is p.d. for any $t \geq 0$, then k is c.p.d.

8. Show that the shortest-path distance on a tree is c.p.d over the set of vertices (a tree is an undirected graph without loops. The shortest-path distance between two vertices is the number of edges of the unique path that connects them). Is the shortest-path distance over graphs c.p.d. in general?