CMPT 410/726 Final Exam
Fall 2022

First Name:

Last Name:

Student Number:

Instructions

- Place your student ID on the desk.
- Write down your first name, last name, and student number on the first (this) page.
- Write your first and last names on the top of every page.
- Write your answers legibly.
- If you write your answer on the back of any page, clearly indicate where your answers are written.
- You may not use any electronic devices, including phones and calculators.
- The university policy on academic dishonesty (cheating) will be taken very seriously in this course.

Useful Information

Gradient of a function $f(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$ and Jacobian of a function $\vec{f}(\vec{x}) : \mathbb{R}^n \to \mathbb{R}^m$:

$$\frac{\partial f}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \quad \frac{\partial \vec{f}(\vec{x})}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{1}$$

Derivative rules:

$$\frac{\partial(\vec{f}(\vec{x})^\top \vec{g}(\vec{x}))}{\partial \vec{x}} = \frac{\partial \vec{f}(\vec{x})}{\partial \vec{x}} \vec{g}(\vec{x}) + \frac{\partial \vec{g}(\vec{x})}{\partial \vec{x}} \vec{f}(\vec{x}) \tag{2a}$$

$$\frac{\partial \vec{f}(\vec{y}_1(\vec{x}), \vec{y}_2(\vec{x}))}{\partial \vec{x}} = \frac{\partial \vec{y}_1}{\partial \vec{x}} \frac{\partial \vec{f}}{\partial \vec{y}_1} + \frac{\partial \vec{y}_2}{\partial \vec{x}} \frac{\partial \vec{f}}{\partial \vec{y}_2} \tag{2b}$$

Common derivatives:

$$\frac{\partial(\vec{a}^\top \vec{x})}{\partial \vec{x}} = \vec{a}, \qquad \frac{\partial(A\vec{x})}{\partial \vec{x}} = A^\top, \qquad \frac{\partial(\vec{x}^\top A \vec{x})}{\partial \vec{x}} = (A + A^\top)\vec{x} \tag{3}$$

Common distributions:

$$\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma) \Rightarrow p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp(-1/2(\vec{x} - \vec{\mu})^\top \Sigma^{-1} (\vec{x} - \vec{\mu})) \tag{4a}$$

$$x \sim \text{Bernoulli}(p) \Rightarrow p(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases} = p^x (1-p)^{1-x} \tag{4b}$$

Ridge regression and maximum a posteriori estimation:

$$\vec{w}^* = \arg\min_{\vec{w}} \sum_i^N (y_i - \vec{w}^\top \vec{x}_i)^2 + \lambda \sum_{i=1}^n w_i^2 \qquad = \arg\min_{\vec{w}} \|\vec{y} - X\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2 \tag{5a}$$

$$= \arg\max_{\vec{w}} p\left(\{x_i, y_i\}_{i=1}^N \mid \vec{w}\right) p(\vec{w}) \qquad = (X^\top X + \lambda I)^{-1} X^\top \vec{y} \tag{5b}$$

Support vector machine:

$$\underset{\vec{w}, b}{\text{minimize}} \quad \frac{1}{2}\|\vec{w}\|_2^2 \tag{6a}$$

$$\text{subject to} \quad y_i(\vec{w}^\top \vec{x}_i - b) \geq 1, \quad \forall i \tag{6b}$$

$$\underset{\vec{\lambda}}{\text{maximize}} \quad \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j \vec{x}_j^\top \vec{x}_i \tag{7a}$$

$$\text{subject to} \quad \sum_{i=1}^N \lambda_i y_i = 0 \tag{7b}$$

$$\lambda_i \geq 0 \quad \forall i \tag{7c}$$

1 Linear Regression

Suppose we are given a data set $\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^N$, where $\vec{x}_i \in \mathbb{R}^n, y_i \in \mathbb{R}$. We would like to use the model $\hat{y} = \vec{w}^\top \vec{x}$ to predict, given any input $\vec{x}$, the label y. This can be done by minimizing the square loss function

$$L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 \tag{8}$$

a) Consider a specific data set $\mathcal{D}_1 = \{(10, 11), (10, 9), (-10, -11), (-10, -9)\}$. Compute w^* by solving the normal equations.

b) Consider another data set $\mathcal{D}_{\text{extra}} = \{(k\vec{e}_i, 0)\}_{i=1}^n$, where k is a constant, and $\vec{e}_i$ is the ith standard basis vector – that is, its ith component is 1, and all other components are 0. We create a new data set $\mathcal{D}_{\text{combined}} = \mathcal{D} \cup \mathcal{D}_{\text{extra}}$ that combines $\mathcal{D}$ and $\mathcal{D}_{\text{extra}}$ and contains $n + N$ data points.

   Determine $\vec{w}^*$, where $\vec{w}^* \in \mathbb{R}^n$, the parameters that minimize the sum of squares of the prediction error on every data point in $\mathcal{D}_{\text{combined}}$, in terms of $\vec{x}_i, \vec{e}_i, k$. You may write your answer in terms of matrices containing any or all of $\vec{x}_i, \vec{e}_i, k$.

Extra Space

2 Maximum Likelihood Estimate

Suppose that the probability that it rains on each day is independent and identically distributed, with probability q that it will rain. Let X_i be a random variable that equals to 1 if it rains on day i, and 0 otherwise. We are given a data set $\mathcal{D} = \{x_i\}$ that records whether it rains or not on each day; each x_i is a realization of the random variable X_i. According to the data set, there are r rainy days, and s non-rainy days.

a) Write down the probability mass function for X_i given some q, $p(x_i|q)$.

b) Write down the likelihood function $p(\mathcal{D}|q)$ and log likelihood function $\log p(\mathcal{D}|q)$, as a function of q, r, s.

c) Prove that $\log p(\mathcal{D}|q)$ is a concave function in q. Hints:

- $x \mapsto k\log(x)$ is concave for any non-negative constant k.
- A sum of concave functions is concave.

d) What is the maximum likelihood estimate for q, based on the data?

Extra Space

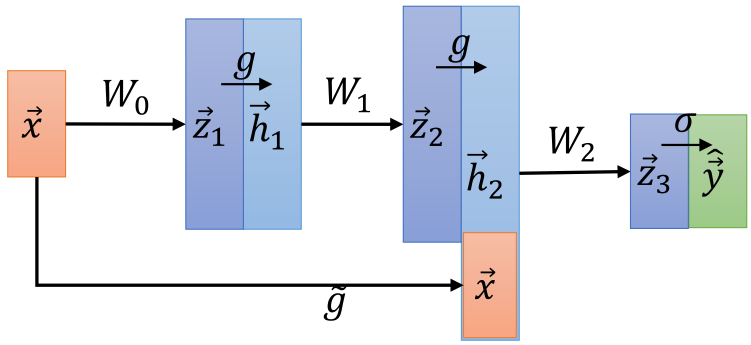

Figure 1: A depiction of a simple neural network with a residual connection.

3 Neural Networks

Consider a neural network shown in Fig. 1, which has four layers. The network parameters are the weight matrices W_0, W_1, W_2. In the two hidden layers, the pre-activation $\vec{z}_i$ goes through the activation function $g(\cdot)$ (element-wise) to form the post-activations $\vec{h}_i$. The output layer involves a different activation function $\sigma(\cdot)$. In this network, we do not concatenate "1"s when going from pre- to post-activations, so $\vec{h}_i$ has the same dimension as $\vec{z}_i$.

In addition to the fully connected layers, the neural network also contains a residual connection: The input $\vec{x}$ is concatenated to the result of applying the activation function on $\vec{z}_2$ to produce $\vec{h}_2$.

Mathematically, each layer of the model can be written as follows:

$$\hat{\vec{y}}(\vec{h}_2) = \sigma\left(\vec{z}_3\right) = \sigma\left(W_2 \vec{h}_2\right), \qquad \vec{h}_2(\vec{h}_1, \tilde{g}) = \begin{bmatrix} g(\vec{z}_2) \\ \vec{x} \end{bmatrix} = \begin{bmatrix} g(W_1 \vec{h}_1) \\ \tilde{g}(\vec{x}) \end{bmatrix},$$
$$\vec{h}_1(\vec{x}) = g(z_1) = g(W_0 \vec{x}) \tag{9}$$

The functions σ, g, and $\tilde{g}$ can be written as follows:

$$\sigma(\vec{z}_i) = \begin{bmatrix} \sigma(z_{i,1}) \\ \vdots \\ \sigma(z_{i,n_i}) \end{bmatrix}, \qquad g(\vec{z}_i) = \begin{bmatrix} g(z_{i,1}) \\ \vdots \\ g(z_{i,n_i}) \end{bmatrix}, \qquad \tilde{g}(\vec{x}) = \vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_{n_0} \end{bmatrix} \tag{10}$$

where $z_{i,j}$ denotes the jth component of $\vec{z}_i$, n_i is the number of elements of z_i, and n_0 is the number of components of $\vec{x}$.

a) Suppose the data set only contains one data point, $(\vec{x}, \vec{y})$. Let $L(W_0, W_1, W_2) = \|\vec{y} - \hat{\vec{y}}\|_2^2$. Compute $\frac{\partial L}{\partial \vec{h}_2}$. What are the dimensions of your answer in terms of $\{n_i\}$?

b) Compute $\frac{\partial \vec{h}_2}{\partial \vec{h}_1}$. What are the dimensions of your answer in terms of $\{n_i\}$?

c) Compute $\frac{\partial \vec{h}_2}{\partial \tilde{g}}$. What are the dimensions of your answer in terms of $\{n_i\}$?

d) Compute $\frac{\partial \vec{h}_1}{\partial \vec{x}}$. What are the dimensions of your answer in terms of $\{n_i\}$?

e) Compute $\frac{\partial L}{\partial \vec{x}}$ in terms of $\frac{\partial L}{\partial \vec{h}_2}, \frac{\partial \vec{h}_2}{\partial \vec{h}_1}, \frac{\partial \vec{h}_2}{\partial \tilde{g}}, \frac{\partial \vec{h}_1}{\partial \vec{x}}$.

f) Suppose that $\vec{x} \in \mathbb{R}^2$, $\vec{z}_1 \in \mathbb{R}^4$, $\vec{z}_2 \in \mathbb{R}^3$, $\vec{z}_3 \in \mathbb{R}^2$. What are the dimensions of W_0, W_1, W_2? How many parameters does the neural network have in this case?

Extra Space

4 Support Vector Machine

Consider a data set containing the following four data points which are linearly separable:

$$\mathcal{D} = \{\big((2,2),1\big), \big((3,2),1\big), \big((-2,-2),-1\big), \big((-3,-2),-1\big)\} \tag{11}$$

a) One linear decision boundary that separates the above data points is the x axis. For this decision boundary, what is(are) the support vector(s)?

b) The above decision boundary can be written as $\vec{w}^\top \vec{x} - b = 0$. Determine the values of $\vec{w}$ and b that take into account your answer in part a), and the corresponding objective value of the primal SVM problem (6).

c) Plot the data points on a graph and draw the optimal decision boundary. What is(are) the support vector(s)?

d) What is the optimal solution $(\vec{w}^*, b^*)$ to the primal SVM problem (6)? You may use a geometric argument. What is the corresponding objective value?

Extra Space