INFR10086 Machine Learning (MLG) Semester 1, 2022/23

Coursework 1

# 1 Submission

- Submit your coursework on gradescope.[1]
- Due: 7 Nov, 2022 at noon, 12:00pm

# 2 Questions

1. Answer the following questions about information theory.

   a) Variance is defined as $\text{Var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2]$. Show that $\text{Var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$.

   *[6 marks]*

   b) Entropy of a random variable $x$ is defined as $\mathbb{E}[-\log p(x)]$. Show that if $x$ is Gaussian, i.e., $x \sim \mathcal{N}(\mu, \sigma^2)$, its entropy is

   $$\mathbb{E}[-\log p(x)] = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}. \tag{1}$$

   (Hint: The derivation is a lot simpler if you use the linearity of expectation and the above fact. In particular, $\sigma^2 = \mathbb{E}[x^2] - \mu^2$.)

   *[6 marks]*

   c) Based on the above, can entropy be negative?

   *[3 marks]*

2. Answer the following questions about optimization.

   a) Is it true that if $f$ is convex in $x$ and $f$ has a minimizer, then $f$ has a unique minimizer? If not, provide a counter example.

   *[5 marks]*

   b) Is it true that if $f$ is $\mu$-strongly convex in $x$ and $f$ has a minimizer, then $f$ has a unique minimizer? If not, provide a counter example. (Hint: To show that the minimizer is unique, assume there are two, say $x_1^*$ and $x_2^*$, and show that $x_1^* = x_2^*$.)

   *[5 marks]*

[1] https://www.gradescope.com/courses/454792/assignments/2375358

c) Suppose we are trying to find the minimum of $f(x) = x^2$ with gradient descent. What is the gradient at $x = 1$? Suppose $x_{t-1} = 1$. Based on the gradient update

$$x_t = x_{t-1} - \eta_t \frac{\partial f}{\partial x}(x_{t-1}), \tag{2}$$

could we get a worse value after a gradient update? In other words, could $f(x_t) \geq f(x_{t-1})$? Based on this result, do gradient updates always reduce the objective?

*[5 marks]*

d) Show that

$$\max(a + b, c + d) \leq \max(a, c) + \max(b, d), \tag{3}$$

for any $a, b, c, d \in \mathbb{R}$. Now, use this fact to show that $h(x) = \max(f(x), g(x))$ is convex in $x$ if $f$ and $g$ are both convex.

*[5 marks]*

3. In this question, we consider the following loss function for binary classification, where $x \in \mathbb{R}^d$ and $y \in \{+1, -1\}$.

$$\ell_{\text{new}}(w; x, y) = \begin{cases} (yw^\top \phi(x) - 1)^2 & \text{if } yw^\top \phi(x) \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

a) Show that $f(s) = s^2$ is convex in $s$.

*[3 marks]*

b) With the fact above, show that $\ell_{\text{new}}(w; x, y)$ is convex in $w$.

*[3 marks]*

c) Is $\ell_{\text{new}}(w; x, y)$ a Lipschitz continuous function? Why or why not?

*[3 marks]*

d) Derive the gradient of $\ell_{\text{new}}(w; x, y)$ with respect to $w$.

*[3 marks]*

e) Show that $\ell_{\text{new}}(w; x, y)$ is an upper bound on the zero-one loss

$$\ell_{01}(w; x, y) = \mathbb{1}_{yw^\top \phi(x) < 0}, \tag{5}$$

for all $w$, $x$, and $y$.

*[8 marks]*

4. Suppose we want to add a few new operations to a neural network library. The neural network library is implemented as a computation graph, and our goal is to implement the backward operations.

a) To allow us to train a regression model, we need to implement the squared loss

$$\ell(y, \hat{y}) = (y - \hat{y})^2, \tag{6}$$

where $y$ is the ground truth and $\hat{y}$ is the prediction. Derive $\frac{\partial \ell}{\partial \hat{y}}$.

*[3 marks]*

b) Instead of logistic function as the activation function, we could use rectified linear units

$$\text{ReLU}(z) = \max(0, z) \tag{7}$$

i) We just showed that $h(x) = \max(f(x), g(x))$ is convex in $x$ if $f$ and $g$ are convex in $x$. Use this fact to show that ReLU is convex in $z$.

*[3 marks]*

ii) Is ReLU a differentiable function? If so, derive the derivative of ReLU. If not, find the point that is not differentiable and a subgradient at that point.

*[6 marks]*

iii) Show that ReLU is 1-Lipschitz continuous.

*[3 marks]*

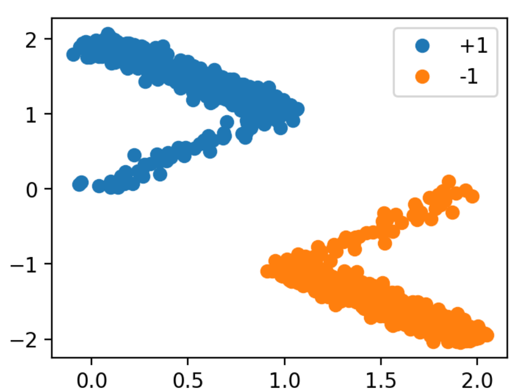
5. There are two common approaches in machine learning that are often compared to each other. One is the **generative approach**, where the distribution of the data points are modeled and hence assumed. The other is the **discriminative approach**, where we do not model the distrubition of data points but only care about achieving a goal, for example, separating data points into two classes.

In this question, we will use an example to illustrate the differences of the two. We are given a 2-dimensional data set.[2] The following piece of code

```
import pickle

f = open('two-L.pkl', 'rb')
pos, neg = pickle.load(f)
f.close()
```

shows how you load the data set. The variable `pos` is a list of points from the positive class (+1), while `neg` is a list of points from the other (−1). The points look like the following in the space.




a) To demonstrate the decision boundary learned from a discriminative approach, write a program to train a linear classifier with log loss to separate the two classes. Show a plot of the line (the classifier) after training together with the points.

[2] Download the data set here `https://homepages.inf.ed.ac.uk/htang2/mlg2022/two-L.pkl`.

*[15 marks]*

b) To demonstrate the decision boundary learned from a generaive approach, we assume points from the positive class is drawn from a Gaussian $\mathcal{N}(\mu_1, \Sigma_1)$, and points from the negative class is drawn from another Gaussian $\mathcal{N}(\mu_2, \Sigma_2)$.

   i) Write a program to compute

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top \tag{9}$$

   and estimate $\mu_1$, $\Sigma_1$, $\mu_2$, and $\Sigma_2$.

   ii) Find the line that passes through $\frac{\mu_1+\mu_2}{2}$ while perpendicular to the vector $\mu_2 - \mu_1$. Show a plot of the line (the classifier) together with the two means and the points.

*[12 marks]*

c) Discuss the pros and cons of both approaches based on the observation.

*[3 marks]*