CSCI 416/516 Practice Final Exam

Name:

Student ID:

Before you start: Check your exam. The exam has 5 pages and 13 questions in total. If your exam is not printed clearly or incomplete, let the instructors know and we will give you a new copy of the exam.

Submission: Write down your name and student ID. You have 3 hours to complete your exam. You are allowed a one-sided (US letter-sized) cheatsheet and a basic or scientific calculator. For all the questions except the binary choice questions, **please show your work/process on how you reach the conclusions to receive full credits assigned to the questions.**

- **Problem 1 [1 pt(s)]: Euclidean Distance.**
  In high-dimensional spaces, does the Euclidean distance metric become more effective at distinguishing between different data points, in the context of KNN?
  - (A) Yes
  - (B) No

- **Problem 2 [1 pt(s)]: Gradient Descent.**
  Is Stochastic Gradient Descent (SGD) more computationally efficient per iteration than Batch Gradient Descent?
  - (A) Yes
  - (B) No

- **Problem 3 [1 pt(s)]: Linear Regression.**
  What is the linearity assumption in linear regression?
  - (A) The linear relationship between features and the predicted variable
  - (B) The linear relationship between features themselves

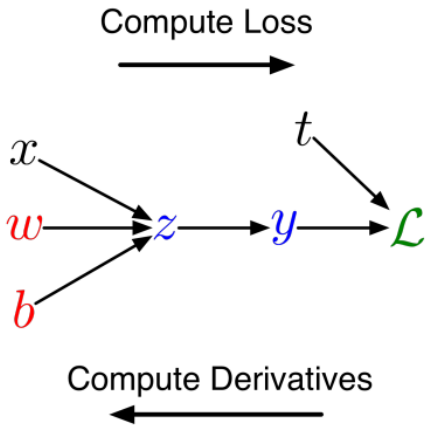
- **Problem 4 [1 pt(s)]: Support Vector Machine.**
  Why do we want to use kernel tricks in SVM?
  - (A) Kernel tricks allow the efficient performance of a non-linear classification without explicitly transforming the data
  - (B) Kernel tricks mitigate the effect of the curse of dimensionality in high-dimensional spaces

- **Problem 5 [2 pt(s)]: 1D Convolution.**
  Given a kernel $k = \{2, 9, 3\}$ and a vector $v = \{1, 2, 3\}$, what is the result of $k * v$? After getting the result $k * v$, what is the new result after applying ReLU on $k * v$? Show the work that leads to your conclusion.

- **Problem 6 [3 pt(s)]: Backpropagation.** What does the back pass look like (in terms of the error signals of parameters/activations, such as but not limited to b, w), given the illustrated forward pass in Figure below? Show your work on how the conclusion is reached.

Computing the loss:

$$z = wx + b$$
$$y = \sigma(z)$$
$$\mathcal{L} = \frac{1}{2}(y - t)^2$$



	Cloudy	Not Cloudy
Raining	50/100	2/100
Not Raining	24/100	24/100

- **Problem 7 [2 pt(s)]: Joint Entropy.**
  Suppose $X = \{\text{Raining, Not raining}\}$, $Y = \{\text{Cloudy, Not Cloudy}\}$, give the above Table. What is the joint entropy, $H(X, Y)$? Show your work on how the conclusion is reached.

- **Problem 8 [2 pt(s)]: Specific Conditional Entropy.**
  Following the setup from the previous question, what is the entropy of cloudiness Y, given that it is raining ($X = \text{Raining}$)? Show your work on how the conclusion is reached.

- **Problem 10 [2 pt(s)]: AdaBoost.**
  Explain what this formula does, in the context of AdaBoost by answering: (1) what is this formula? (2) what does the part $y_i \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + (1 - y_i)\log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$ do? (3) what does the part $\lambda||\boldsymbol{\theta}_{[1:d]}||_2^2$ do? (4) What is the term $w_{i,t}$ and how does it affect $\mathcal{J}_{\text{reg},t}(\boldsymbol{\theta})$?

$$\mathcal{J}_{\text{reg},t}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} w_{i,t}[y_i \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + (1 - y_i)\log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_i))] + \lambda||\boldsymbol{\theta}_{[1:d]}||_2^2 \tag{1}$$

- **Problem 11 [3 pts]: Support Vector Machines.**
  The optimization objective of SVM is given as $\min_{\boldsymbol{\theta}} C \sum_{i=1}^{N} [y_i \text{cost}_1(\boldsymbol{\theta}^\top \boldsymbol{x}_i) + (1-y_i)\text{cost}_0(\boldsymbol{\theta}^\top \boldsymbol{x}_i)] + \frac{1}{2}\sum_{j=1}^{d} \theta_j^2$, where cost_0 and cost_1 are defined using the hinge loss. Explain the difference between the scenario in which the tunable hyperparameter C is large and the scenario in which C is small - what are we favoring, by making C large or small?

- **Problem 12 [2 pts]: Lagrangian Multipliers.**
  Suppose a sample in the training dataset has a Lagrangian multiplier being 0. What does this say about this sample?

- **Problem 13 [3 pts] (bonus): Attention & Transformers.**
  What's the relationship between the Scaled Dot-Product Attention and Multi-Head Attention, in the transformer architecture?