CSCI 416/516 Homework #1

DUE: October 04, 2024 at 11:59 pm on Blackboard

CSCI 416/516: Each Problem begins with an allocation of points, represented as [u pts/g pts]. If you are registered in CSCI 416, you can receive up to u pts on this Problem; if you are registered in CSCI 516, you can receive up to g pts on this Problem. The last Problem is optional for undergraduates (CSCI 416) but required for graduates (CSCI 516). **Write down which session you are in / are you a graduate or undergraduate student**.

Optional/Extra Credit Problem for Undergraduates: Problem 8 is required for graduates but optional to undergraduates. Undergraduates can earn up to 10 credits from Problems 1-7 already, and can earn up to 12 credits from Problems 1-8. On the other hand, graduates can learn up to 10 credits from 1-8, so the maximum credits they can earn from attempting all the questions is 10 out of 10 (while for undergrads this number is 12 out of 10).

Submission: For all the problems excluding the multiple choice problem(s), you need to **show all your works, steps, and calculations** if applicable, or **your justification/expalantion to the answer(s) you provide**. You should submit a PDF to Blackboard with your answers that are recognizable/intelligible. Preferably, you should use LaTeX.

- **Problem 1 [1pt/1pt]: Euclidean Distance.**
  Consider the following 3-dimensional points, $x^{(a)} = [1, -3, 5]$ and $x^{(b)} = [-2, 4, -6]$. Write the formula for the Euclidean distance between two points in a 3-dimensional space. Then, using the formula, calculate the Euclidean distance between $x^{(a)}$ and $x^{(b)}$.

- **Problem 2 [1pt/1pt]: Curse of Dimensionality.**
  Imagine you're working with a dataset of e-commerce product reviews. Each review is represented as a vector, where each dimension corresponds to the frequency of a particular word from a predefined vocabulary. The dataset has 10,000 reviews, and the vocabulary size is 50,000 words. When we talk about the curse of dimensionality, what is the size of the dimensionality in this case?

- **Problem 3 [1pt/1pt]: KNN.**
  KNN typically uses Euclidean distance as its default metric. However, which of the following metrics CANNOT be used as the distance metric in KNN? (Hint: we covered this question in class.)

  - A. Kullback–Leibler (KL) Divergence

- B. Cosine Similarity
- C. Edit Distances (such as Hamming Distance)
- D. All of the above can be used as the distance metric

- **Problem 4 [1pt/1pt]: Linear Regression Cost Function.**
  What is the typical cost function in linear regression (that we covered in class) - how is it defined mathematically?

- **Problem 5 [2pt/2pt]: Regularized Linear Regression.**
  For this problem, we will use the linear regression model from the lecture.

$$y = f(x) = \sum_{j}^{D} w_j x_j + b \tag{1}$$

In the lecture, we saw that regression models with too much capacity can overfit the training data and fail to generalize. We also saw that one way to improve generalization is regularization: adding a term to the cost function that favors some explanations over others. For instance, we might prefer that weights not grow too large in magnitude. We can encourage them to stay small by adding a penalty:

$$\mathcal{R}(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_2^2 = \frac{1}{2}\sum_j w_j^2 \tag{2}$$

  - **5(a) [1pt/1pt]** What is the mathematical definition of the regularized cost function given $\mathcal{R}(\boldsymbol{w})$ and the unregularized cost function $\boldsymbol{\mathcal{J}}(\boldsymbol{w})$, represented by the weight $\boldsymbol{w}$, the sample $\boldsymbol{x}$, the bias b, and the target t?
  - **5(b) [1pt/1pt]** Determine the update rules using gradient descent for the regularized cost function $\boldsymbol{\mathcal{J}}(\boldsymbol{w})_{\text{reg}}$, given the learning rates α_{w_j} and α_b . Your answer should have the form:

$$w_j \leftarrow ... \tag{3}$$

$$b \leftarrow ... \tag{4}$$

- **Problem 6 [2pt/1pt]: Gradients in Logistic Regression.** In the lecture, we show that we can find gradients, which allows us to use gradient descent update to find the weights of logistic regression. The gradient regarding w_j is defined as the following Equation:

$$\frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \left(-\frac{t}{y} + \frac{1-t}{1-y}\right) \cdot y(1-y) \cdot x_j = (y-t)x_j \tag{5}$$

How do we arrive to $\left(-\frac{t}{y} + \frac{1-t}{1-y}\right) \cdot y(1-y) \cdot x_j = (y-t)x_j$ from $\frac{\partial \mathcal{L}_{CE}}{\partial w_j}$? **Show all your works, steps, and calculations**.

- **Problem 7 [2pt/2pt]: When should we use stochastic gradient descent over batch gradient descent?**

- **Problem 8 (Optional to Undergraduates) [2pt/1pt]: Linear Regression Using Gradient Descent.**
  Given $\boldsymbol{x} = \{x_1, x_2, x_3, x_4\} = \{1, 2, 3, 4\}$ and $\boldsymbol{t} = \{t_1, t_2, t_3, t_4\} = \{10, 20, 30, 40\}$, and the initial $\boldsymbol{w}_{iter=0} = \{w_0, w_1\} = \{1, 1\}$ in which the bias b is incorporated as w_0, what is the weight $\boldsymbol{w}_{iter=1}$ after 1 iteration with a learning rate of 0.1? **Show all your works, steps, and calculations**.