This is a closed-book test. It is marked out of 15 marks. Please answer ALL of the questions. Here is some advice:

- The questions are NOT arranged in order of difficulty, so you should attempt every question.

- Questions that ask you to “briefly explain” something only require short (1-3 sentence) explanations. Don’t write a full page of text. We’re just looking for the main idea.

- None of the questions require long derivations. If you find yourself plugging through lots of equations, consider giving less detail or moving on to the next question.

- Many questions have more than one right answer.
Q1: _____ / 2
Q2: _____ / 1
Q3: _____ / 1
Q4: _____ / 2
Q5: _____ / 1
Q6: _____ / 2
Q7: _____ / 2
Q8: _____ / 2
Q9: _____ / 2

Final mark: _____ / 15
Have you taken CSC321 at UofT? (This question is used for calibration purposes.)

1. As discussed in lecture, when applying K-nearest-neighbors, it is common to normalize each input dimension to unit variance.

   (a) [1pt] Why might it be advantageous to do this?

   (b) [1pt] When might this normalization step not be a good idea? (Hint: You may want to consider the task of classifying images of handwritten digits, where the digit is centered within the image.)

2. [1pt] In random forests, what is the motivation for randomizing the set of attributes considered for each split?
3. [1pt] Suppose you want to evaluate the test error rate of a 1-nearest-neighbors classifier. Assume you implement the algorithm the naïve way, i.e. by explicitly computing all the distances and taking the min, rather than by using a fancy data structure. What is the running time of evaluating the test error? Give your answer in big-O notation, in terms of the number of training examples \( N_{\text{train}} \), the number of test examples \( N_{\text{test}} \), and the input dimension \( D \). Briefly explain your answer.

4. (a) [1pt] Give one advantage of K-nearest-neighbors over linear regression.

(b) [1pt] Give one advantage of linear regression over K-nearest-neighbors.
5. [1pt] Suppose linear regression (with squared error loss) is used as a classification algorithm. TRUE or FALSE: if it correctly classifies every training example, then its cost is zero. (By “cost”, we mean the function minimized during training.) Briefly justify your answer.

6. [2pts] Let $Z$ be a random variable and $t$ be a real number. Show that

$$\mathbb{E}[(Z - t)^2] = (\mathbb{E}[Z] - t)^2 + \text{Var}[Z].$$

(This is a simplified version of the bias-variance decomposition.)
7. [2pts] Suppose binary-valued random variables $X$ and $Y$ have the following joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>1/8</td>
<td>3/8</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>2/8</td>
<td>2/8</td>
</tr>
</tbody>
</table>

Determine the information gain $IG(Y|X)$. You may write your answer as a sum of logarithms.
8. [2pts] Recall that combining the logistic activation function with squared error loss suffers from saturation, whereby the gradient signal is very small when the prediction for a training example is very wrong. Logistic regression (i.e. logistic activation function with cross-entropy loss) doesn’t have this problem. Recall that the logistic function is defined as \( \sigma(z) = 1/(1 + e^{-z}) \). Now suppose we modify the activation function to squash the prediction \( y \) to be in the interval \([0.1, 0.9]\), and then apply cross-entropy loss. I.e.,

\[
\begin{align*}
z &= \mathbf{w}^\top \mathbf{x} + b \\
y &= 0.8\sigma(z) + 0.1 \\
\mathcal{L}(y, t) &= -t \log y - (1 - t) \log(1 - y),
\end{align*}
\]

where \( \sigma \) is the logistic activation function. Does this model have a problem with saturation? You don’t need to give a formal proof, but you should informally justify your answer. Hint: it is possible to answer this question without calculating derivatives. Think qualitatively.
9. **[2pts]** Recall that the soft-margin SVM can be viewed as minimizing the hinge loss with an $L_2$ regularization term. I.e.,

$$z = w^\top x + b$$

$$\mathcal{L}(z, t) = \max(0, 1 - tz)$$

$$\mathcal{J}(w, b) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(z^{(i)}, t^{(i)}).$$

Here, $t \in \{-1, +1\}$. Complete the formulas for the gradient calculations. You don’t need to show your work.

$$\frac{\partial \mathcal{J}}{\partial w} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{L}^{(i)}}{\partial w} + \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{L}^{(i)}}{\partial w}$$  

(fill in the blank)

$$\frac{d\mathcal{L}}{dz} =$$

$$\frac{\partial \mathcal{L}}{\partial w} =$$  

(give in terms of $\frac{d\mathcal{L}}{dz}$)
(Scratch work or continued answers)