Instructions: read these first!

Do not open the exam, turn it over, or look inside until you are told to begin.

Switch off cell phones and other potentially noisy devices.

Write your full name on the line at the top of this page. Do not separate pages.

You may refer to a calculator and your cheat-sheet. You may not refer to any other printed material, and any other computational device (such as laptops, phones, iPads, friends, enemies, pets, lovers).

Read questions carefully. Show all work you can in the space provided.

Where limits are given, write no more than the amount specified. The rest will be ignored.

Avoid seeing anyone else’s work or allowing yours to be seen.

Do not communicate with anyone but an exam proctor.

If you have a question, raise your hand.

When time is up, stop writing.

You can see your graded final in the student affairs office starting Mon Mar 27.

<table>
<thead>
<tr>
<th>Question</th>
<th>Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td>86</td>
<td></td>
</tr>
</tbody>
</table>
1. [12 points] In this question, we will again look at how $k$-nearest neighbor classifiers can be robust to noise. Suppose we have two labels 0 and 1. We are given a test point $x$, and its $k$ nearest neighbors $z_1,\ldots,z_k$, where $z_i$ is the $i$-th closest neighbor of $x$ (so $z_1$ is the closest neighbor, $z_2$ is the second closest neighbor and so on).

Suppose that the probability that the label of $z_i$ is not equal to the label of $x$ is $p_i$. Also assume that all distances are unique, so the $i$-th closest neighbor of $x$ is unique for all $i$.

Answer the following questions:

(a) [3 points] Now, suppose, for this question that $p_1 = 0.1$, and $p_i = 0.2$ for $i > 1$. What is the probability that the 1-nearest neighbor classifier makes a mistake on $x$?

(b) [6 points] In the setting of part (a), what is the probability that the 3-nearest neighbor classifier makes a mistake on $x$?

(c) [3 points] Based on your calculation, what can you conclude about the relative robustness of 1 and 3-nearest neighbor classifiers in this case?
2. [10] Solve the following optimization problem by the substitution method. Write down the KKT conditions for it, and use the solutions to derive the values of the dual variables.

\[
\begin{align*}
\text{min} & \quad x + y \\
\text{subject to:} & \quad x^2 + y^2 = 4 \\
& \quad x \geq 0
\end{align*}
\]
3. [12 points] For each of the following statements, say whether they are correct or incorrect. Justify your answer.

(a) [4 points] For every training data set \((x_1, y_1), \ldots, (x_n, y_n)\), the training error of the 3-nearest neighbor classifier is always zero. You may assume that each \(x_i\) is unique.

(b) [4 points] As we increase \(k\), the training error of the \(k\)-nearest neighbor classifier always increases.

(c) [4 points] Recall that two classifiers \(c_1\) and \(c_2\) are equal, if \(c_1(x) = c_2(x)\) for all \(x\) in the domain. If two classifiers \(c_1\) and \(c_2\) have the same training error, then they are equal.
4. [10 points] Calculate the gradient and the Hessian for the following functions, and use them to determine if they are convex or not. Justify your answer.
   (a) [4 points] \( f(x) = \sum_{i=1}^{d} x_i \log x_i \).

(b) [4 points] \( f(x) = 20x^T x - 5x^T 1 \). where 1 is the all-ones vector.

(c) [4 points] Now suppose that \( f(x) \) is convex, and let \( g(x) = (f(x))^2 \). Write down the gradient and Hessian of \( g(x) \) in terms of those of \( f \). If \( f \) is convex, is \( g(x) \) always convex?
5. [25 points] State whether each of the following statements is true or false. If it is true, provide a brief justification or proof; if it is false, provide a counterexample or a justification.

(a) [5 points] Suppose we are given a data set \((x_1, y_1), \ldots, (x_n, y_n)\) that is linearly separable. If the feature vectors have zero mean (that is, if \(\frac{1}{n} \sum_{i=1}^{n} x_i = 0\)), then, the data set \((x_1, y_1), \ldots, (x_n, y_n)\) is also linearly separable through the origin.

(b) [5 points] Let \(X\) and \(Z\) be two random variables such that \(Z = X - 2\). Then \(H(Z) = H(X)\).

(c) [5 points] If the training data is linearly separable, then running perceptron on one pass of the data is guaranteed to learn a classifier with zero training error.
(d) [5 points] Two decision trees that have the same training error on a training dataset $S$ also have the same test error on the same test dataset $T$, no matter what $S$ and $T$ are.

(e) [5 points] Let the data domain be the set $X = \{0, 1\}^d$ and let $\mathcal{H}$ be the class of all functions on $X$. Then, $|\mathcal{H}| = 2^d$. 
6. [25 points] State whether each of the following statements is true or false. If it is true, provide a brief justification or proof; if it is false, provide a counterexample or a justification.

(a) [5 points] $x, z$ are real $d$-dimensional vectors. Then the function $K(x, z) = \frac{\langle x, z \rangle}{\|x\|^2 \|z\|^2}$ is a kernel.

(b) [5 points] $x, z$ are real $d$-dimensional vectors. Then the function $K(x, z) = \|x - z\|^2$ is a kernel.

(c) [5 points] If $K(x, z)$ and $L(x, z)$ are kernels, then $M(x, z) = 2K(x, z) - 3L(x, z)$ is a kernel, no matter what $K$ and $L$ are.
(d) [5 points] If \( K(x, z) \) and \( L(x, z) \) are kernels, then \( M(x, z) = K(x, z)L(x, z)^2 \) is also a kernel, no matter what \( K \) and \( L \) are.

(e) [5 points] The matrix \( B \) is a valid kernel matrix:

\[
B = \begin{bmatrix}
1 & 1 \\
1 & 0
\end{bmatrix}
\] (1)