Problem 1. (15 points) Consider a logistic regression problem where \( \mathcal{X} = \mathbb{R}^d \) and \( \mathcal{Y} = \{-1, +1\} \). Derive the weight update rule that maximizes the conditional likelihood assuming that a data set \( D = \{(x_i, y_i)\}_{i=1}^n \) is given.

Problem 2. (20 points) Consider a logistic regression problem with its initial solution obtained through the OLS regression; i.e., \( w^{(0)} = (X^T X)^{-1} X^T y \), in the context of the code provided in class (week 6). Recall that \( x \) was drawn from a mixture of two Gaussian distributions with \( \text{dim}\{x\} = 2 \) (before adding a column of ones) and that \( y \in \{0, 1\} \). You probably noticed that the initial separation line is consistently closer to the data points of class 0.

   a) (10 points) Why was this the case? Draw a picture (if possible) to support your argument.
   
   b) (5 points) Devise a better initial solution by modifying the standard formula \( w^{(0)} = (X^T X)^{-1} X^T y \).
   
   c) (5 points) Now again consider the case where \( y \in \{-1, +1\} \). What is the form of the modified solution from part (b) in this case?

Problem 3. (30 points) Consider a classification problem where \( \mathcal{X} = \mathbb{R}^d \) and \( \mathcal{Y} = \{0, 1\} \). Use maximum conditional likelihood estimation to develop a linear classifier that models the posterior probability of the positive class as

\[
P(Y = 1|x, w) = \frac{1}{2} \left( 1 + \frac{w^T x}{\sqrt{1 + (w^T x)^2}} \right).
\]

Implement the iterative weight update rule and compare the final decision line between logistic regression and this new linear classifier. Use at least 10 different data sets to draw your conclusions.

Problem 4. (15 points) Let \( A, B, C, \) and \( D \) be binary input variables (features). Give decision trees to represent the following Boolean functions:

   a) (3 points) \( A \land \overline{B} \)
   
   b) (3 points) \( A \lor (B \land \overline{C}) \)
   
   c) (3 points) \( \overline{A} \oplus B \)
   
   d) (3 points) \( (A \land B) \lor (C \land D) \)
   
   e) (3 points) \( (A \lor B) \land (C \lor \overline{D}) \oplus B \)
where $\bar{A}$ is the negation of $A$ and $\oplus$ is an exclusive OR operation.

**Problem 5.** (10 points) Consider the data set from the following table:

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>Enjoy Sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
</tr>
<tr>
<td>2</td>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
</tr>
<tr>
<td>3</td>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
</tr>
<tr>
<td>4</td>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
</tr>
</tbody>
</table>

a) (5 points) Using Enjoy Sport as the target, show the decision tree that would be learned if the splitting criterion was information gain.

b) (5 points) Add the training example from the table below and compute the new decision tree. This time, show the value of the information gain for each candidate attribute at each step in growing the tree.

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>Enjoy Sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Weak</td>
<td>Warm</td>
<td>Same</td>
</tr>
</tbody>
</table>

**Problem 6.** (40 points) Empirically evaluating the claim that neural networks with large number of hidden neurons generalize well. The idea is to take a data set $D$ and split it into training and test partitions. Then, train a single network on the training partition and evaluate it on the test partition, but do not use a validation set for early stopping. Increase the number of hidden neurons from 1 to 2, to 4 and so on in the powers of two as much as you have computing resources. Then plot training and test mean squared error as a function of the number of hidden neurons (repeat several times for the results to stabilize). If you are considering a classification problem, you can also plot classification accuracy. Finally, consider plotting the training and test error as a function of the number of epochs.

Your networks can be deep or shallow or both. You can also experiment with different numbers of epochs to run, different regularization schemes, early stopping using a validation set, different data normalization schemes, or with batch vs. stochastic optimization. Make sure to select data sets that are large enough so that you can obtain meaningful results. Consider 5 different data sets, say from the UCI Machine Learning Repository. You can use any toolbox available in any programming language as long as you feel you have full control of the process.
Homework Directions and Policies

Submit a single package containing all answers, results and code. Your submission package should be compressed and named firstnamelastname.zip (e.g., predragradivojac.zip). In your package there should be a single pdf file named main.pdf that will contain answers to all questions, all figures, and all relevant results. Your solutions and answers must be typed\(^1\) and make sure that you type your name and Northeastern username (email) at the beginning of the file. The rest of the package should contain all code that you used. The code should be properly organized in folders and subfolders, one for each question or problem. All code, if applicable, should be turned in when you submit your assignment as it may be necessary to demo your programs to the teaching assistants. Use Matlab, Python, or C/C++.

Unless there are legitimate circumstances, late assignments will be accepted up to 5 days after the due date and graded using the following rules:

- on time: your score $\times 1$
- 1 day late: your score $\times 0.9$
- 2 days late: your score $\times 0.7$
- 3 days late: your score $\times 0.5$
- 4 days late: your score $\times 0.3$
- 5 days late: your score $\times 0.1$

For example, this means that if you submit 3 days late and get 80 points for your answers, your total number of points will be $80 \times 0.5 = 40$ points.

All assignments are individual, except when collaboration is explicitly allowed. All the sources used for problem solution must be acknowledged; e.g., web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously! For detailed information see Office of Student Conduct and Conflict Resolution.

---

\(^1\)We recommend Latex; in particular, TexShop-Mac\TeX combination for a Mac and TeXnicCenter-MiKTeX combination on Windows. An easy way to start with Latex is to use the freely available Lyx. You can also use Microsoft Word or other programs that can display formulas professionally.