1. Consider using logistic regression for a two-class classification problem in two dimensions:

\[ p(y = 1|\mathbf{x}) = \sigma(w_0 + w_1 x_1 + w_2 x_2) \]

Here \( \sigma \) denotes the logistic (or sigmoid) function \( \sigma(z) = 1/(1 + \exp(-z)) \), \( y \) is the target which takes on values of 0 or 1, \( \mathbf{x} = (x_1, x_2) \) is a vector in the two-dimensional input space, and \( \mathbf{w} = (w_0, w_1, w_2) \) are the parameters of the logistic regressor.

(a) Consider a weight vector \( \mathbf{w}_A = (-1, 1, 0) \). Sketch the decision boundary in \( \mathbf{x} \) space corresponding to this weight vector, and mark which regions are classified with labels 0 and 1.

(b) Consider a second weight vector \( \mathbf{w}_B = (5, -5, 0) \). Again sketch the decision boundary in \( \mathbf{x} \) space corresponding to this weight vector, and mark which regions are classified with labels 0 and 1.

(c) Plot \( p(y = 1|\mathbf{x}) \) as a function of \( x_1 \) for both \( \mathbf{w}_A \) and \( \mathbf{w}_B \), and comment on any differences between the two.

2. Consider the logistic regression setup in the previous questions, but with the weight vector \( \mathbf{w}_A = (0, -1, 1) \). Consider the following data set: Compute the gradient of the log likelihood of the logistic regression model for this data set. Suppose that we take a single gradient step with \( \eta = 1.0 \); what is the new parameter setting? Do the new parameters do a better job of classifying the training data?

<table>
<thead>
<tr>
<th>Instance</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>-0.35</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>-0.1</td>
<td>0.1</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-1.2</td>
<td>1.0</td>
<td>+</td>
</tr>
</tbody>
</table>

It will help you to remember the following facts:

- The log-likelihood in logistic regression is
  \[
  L(\mathbf{w}) = \sum_{i=1}^{n} \log p(y = y_i | \mathbf{x}_i) \\
  = \sum_{i=1}^{n} [y_i \log p(y = 1|\mathbf{x}_i) + (1 - y_i) \log p(y = 0|\mathbf{x}_i)]
  \]

- The partial derivative of the log-likelihood with respect to a parameter \( w_j \) is
  \[
  \frac{\delta L}{\delta w_j} = \sum_{i=1}^{n} (y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i))x_{ij}
  \]

- To maximize a function \( L(\mathbf{w}) \), we use the gradient ascent rule, which is
  \[
  \mathbf{w}' \leftarrow \mathbf{w} + \eta \nabla L
  \]