1. (10 points) Remember that when using Naive Bayes, we make the assumption that all features are conditionally independent given the target value. Factorize the joint probability $P(\text{Cold, Headache, Cough, Sore Throat})$ using this assumption.

2. (10 points) Using no Laplacian smoothing, what is $P(\text{Cold}|\neg\text{Headache, Cough, Sore Throat})$?

3. (15 points) Now estimate the conditional likelihoods using Laplacian smoothing with $\alpha = 1$. What is $P(\text{Cold}|\text{Headache, } \neg\text{Cough, } \neg\text{Sore Throat})$?

4. (15 points) In our dataset above, $\text{Cough}$ and $\text{Sore Throat}$ are completely correlated, which doesn’t fit with our conditional independence assumption. Ignore the $\text{Sore Throat}$ feature, and compute $P(\text{Cold}|\text{Headache, } \neg\text{Cough})$ (use Laplacian smoothing with $\alpha = 1$ again). Is our prediction the same as when we used the $\text{Sore Throat}$ feature?
2 Perceptrons [50 points]

Recall that a perceptron learns a linear classifier with weight vector $w$. It predicts

$$y^{(i)} = \text{sign}(w^T x^{(i)})$$

(assuming here that $y^{(i)} \in \{+1, -1\}$. Also, note that we are not using a bias weight $w_0$, for simplicity).

When the perceptron makes a mistake, it updates the weights using the formula

$$w = w + y^{(i)} x^{(i)}$$

Imagine that we have $x^{(i)} \in \mathbb{R}^2$, and we encounter the following data points

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-3</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

1. (25 points) Starting with $w = [0 \ 0]^T$, use the perceptron algorithm to learn on the data points in the order from top to bottom. Show the perceptron’s linear decision boundary after observing each data point in the graphs below. Be sure to show which side is classified as positive.
2. (10 points) Does our learned perceptron maximize the margin between the training data and the decision boundary? If not, draw the maximum-margin decision boundary on the graph below.

3. (15 points) Assume that we continue to collect data and train the perceptron. If all data we see (including the points we just trained on) are linearly separable with margin $\gamma = 0.5$ and have maximum norm $\|x^{(i)}\| \leq 5$, what is the maximum number of mistakes we can make on future data?
3  Logistic Regression [100 points]

The programming portion of this assignment can be found at https://github.com/pjreddie/Logistic-SGD.