CPSC 532M Assignment 1

Lucas Palazzi, Student #79285755

1 Linear Algebra Review

1.1 Basic Operations

1. $\sum_{i=1}^{n} x_i y_i = 14$
2. $\sum_{i=1}^{n} x_i z_i = 0$
3. $\alpha(x + z) = \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix}$
4. $x^T z + \|x\| = \sqrt{5} \approx 2.236$
5. $Ax = \begin{bmatrix} 6 \\ 5 \\ 7 \end{bmatrix}$
6. $x^T A x = 19$
7. $A^T A = \begin{bmatrix} 11 & 10 & 10 \\ 10 & 14 & 10 \\ 10 & 10 & 14 \end{bmatrix}$

1.2 Matrix Algebra Rules

1. True
2. True
3. False
4. True
5. False
6. True
7. True
8. False
9. True
10. False

2 Probability Review

2.1 Rules of probability

1. Fair price = \$8.75
2. $\Pr(B) = 0.55$
3. $\Pr(B) = 0.55 + \Pr(A, B)$

2.2 Bayes Rule and Conditional Probability

1. $\Pr(T = 1) = 0.010096$
2. Most of the positive tests come from false positives.
3. $\Pr(D = 1 \mid T = 1) = 0.009607765$
4. They are not likely to be a drug user.
5. Perform repeated tests.

3 Calculus Review

3.1 One-variable derivatives

1. $f'(x) = 6x - 2$
2. $f'(x) = 1 - 2x$
3. $f'(x) = \frac{e^x}{1+e^{-x}} = 1 - e^{-x}p(x)$

3.2 Multi-variable derivatives

1. $\nabla f(x) = \begin{bmatrix} 2x_1 + e^{x_1+2x_2} & 2e^{x_1+2x_2} \end{bmatrix}$
2. $\nabla f(x) = \begin{bmatrix} \frac{e^{x_1}}{Z} & \frac{e^{x_2}}{Z} & \frac{e^{x_3}}{Z} \end{bmatrix}$
3. $\nabla f(x) = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$
4. $\nabla f(x) = \begin{bmatrix} 2x_1 - x_2 & 2x_2 - x_1 \end{bmatrix}$
5. $\nabla f(x) = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_d \end{bmatrix}$

3.3 Optimization

1. $\min\left(3x^2 - 2x + 5 \mid x \in \mathbb{R}\right) = 14/3$
2. $\max\left(x(1-x) \mid x \in [0,1]\right) = 1/4$
3. $\min\left(x(1-x) \mid x \in [0,1]\right) = 0$
4. $\arg\max\left(x(1-x) \mid x \in [0,1]\right) = 1/2$
5. $\min\left(x_1^2 + e^{x_2} \mid x \in [0,1]^2\right) = 1$
6. $\arg\min\left(x_1^2 + e^{x_2} \mid x \in [0,1]^2\right) = \begin{bmatrix} 0 & 0 \end{bmatrix}$

3.4 Derivatives of code

See `README.md` file.

4 Algorithms and Data Structures Review

4.1 Trees

1. minimum depth = 6
2. minimum depth = 6

4.2 Common Runtimes

1. $O(n)$
2. $O(\log n)$
3. $O(1)$
4. $O(d)$
5. $O(d^2)$

4.3 Running times of code

`func1`: $O(n)$

`func2`: $O(n)$

`func3`: $O(n)$

`func4`: $O(n \log n)$

5 Data Exploration

5.1 Summary Statistics

See `README.md` file for link to code.

1.

minimum	0.352
maximum	4.862
mean	1.3246
median	1.159
mode	0.77

2.

5%	0.46495
25%	0.718
50%	1.159
75%	1.81325
95%	2.62405

3.

Region with minimum mean	Pac
Region with maximum mean	WtdILI
Region with minimum variance	Pac
Region with maximum variance	Mtn

For continuous data such as in this data set, the mode is not a reliable estimate of the most "common" value. It can be unlikely that the exact same value will repeat itself even if multiple values occur that are approximately equal, since the data can have many decimal places. In this case it could be better to consider *ranges* of values to compare, for example using a histogram.

5.2 Data Visualization

1. **D** - frequency of each individual column (region) is included in graph
2. **C** - histogram with only values, all regions included together
3. **B** - boxplot with one for each week, all regions considered
4. **A** - plot shows illness percentage for each region over time
5. **F** - this scatterplot has high correlation, looks to be a strong linear fit, no outliers
6. **E** - scatterplot appears to have lower correlation, handful of outliers

6 Decision Trees

See `README.md` file for links to code.

6.1 Splitting rule

Equality-based splitting should be used instead of threshold-based splitting for features where the numerical order doesn't matter, for example if numbers are representing data like cities or true/false.

6.2 Decision Stump Implementation

Reported error = 0.265

Figure 1 shows the generated figure for 6.2.

6.3 Decision Stump Info Gain Implementation

Reported error = 0.265

Figure 2 shows the generated figure for 6.3.

6.4 Constructing Decision Trees

See `README.md` for link to code.

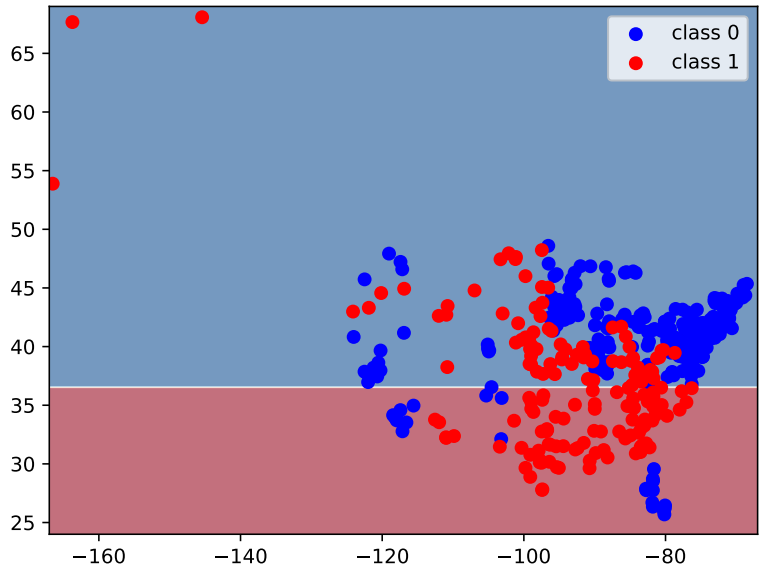

Figure 1: Generated figure for question 6.2

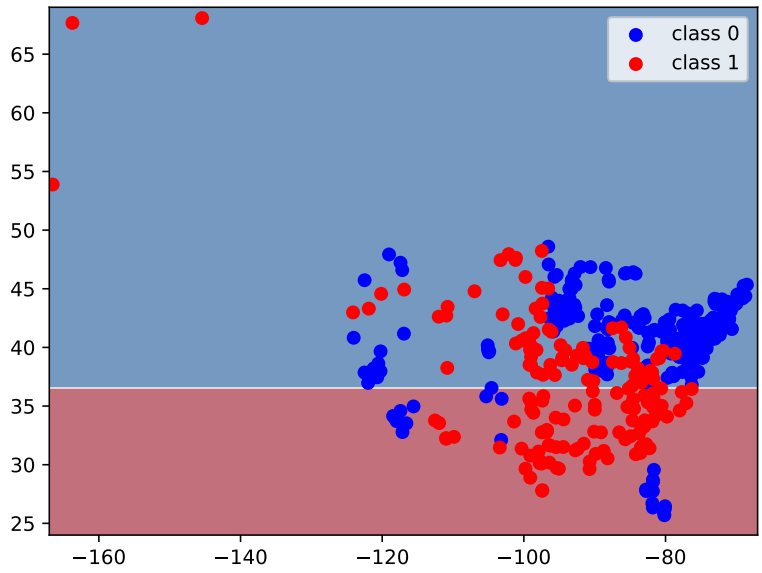

Figure 2: Generated figure for question 6.3

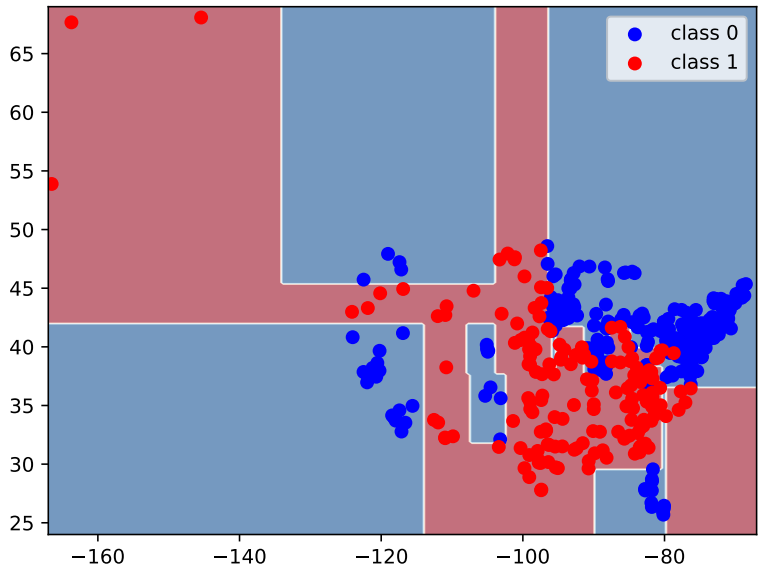

Figure 3: Generated figure for question 6.5

6.5 Decision Tree Training Error

At higher tree depths, approach (1) is the least accurate, approach (2) is the second least accurate, and approach (3) is the most accurate. Option (1) is likely worse than (2) because it uses error rate to pick the best tree while (2) uses information gain. This makes (1) worse because there are likely parts of the map where drawing a vertical or horizontal line doesn't necessarily improve accuracy, but *does* improve information gain.

Figure 3 shows the classification boundary plot for approach (3), which has the lowest training error.

6.6 Comparing implementations

This does not conclusively demonstrate that the two implementations are the same. In order to build confidence that the two implementations are equivalent one should test them on data outside of the training data and compare the predictions.

6.7 Cost of Fitting Decision Trees

$O(nd \log n \log m)$