Final - Math 4803

Dennis Goldenberg

April 2023

1 Theoretical Problems

1. Recall standard logistic regression. The typical method of fitting the coefficient parameters β is via minimizing the negative log likelohood of the observed data (equivalent to maximizing the log likelihood). For this problem, consider logistic regression but with a regularizer/penalty that is the same that used in ridge regression. Recall that the rdige regression penalty term does not include the constant offset parameter β_0. Please do the following:
a) Write down the minimization problem described above whose solution would yield the optimal values for β. The coffiecient of the penalty term in your minimization problem should be λ.
b) Derive a system of equations that one would need to solve in order to find the minimizer for part (a). You do not need to solve this system of equations in any way - just derive them and write them down.
c) For this part of the problem, consider the case where $\lambda \to \infty$. What will be the optimal values for all the β_J with $j > 0$. Find an an explicit formula for the optimal value of β_0 in this case.
Solution
a) - Let $\{x_1, x_2, ..., x_n\}$ be our data, each with p features, and $\{y_0, y_1, ..., y_n\}$ be the response. From equation 4.7 in the textbook, we can deduce:

$$p(y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

Since there are only two classes, we have that:

$$p(y = 0|X) = 1 - p(y = 1|X) = \frac{1}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

We can extend equation 4.5 in the textbook to p features to obtain the likelihood function:

$$\ell(\beta_0, \beta_1, ..., \beta_p) = \prod_{i;y_i=1} p(x_i) \prod_{i':y_{i'}=0} 1 - p(x_i)$$

Note, however, that maximizing this function is equivalent to minimizing the negative log-likelihood function, or $-\ln(\ell(\beta_0, \beta_1, ..., \beta_p))$. Also, note that in

linear regression with ridge regularization, to obtain the optimal β values, we must solve (from equation 6.5 in the textbook):

$$\text{argmin}_{\vec{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

The first part of that is the Residual sum of squares, which is specific to linear regression; however, it's the second term that regularizes by punishing too large coefficients. Therefore, we combine the minimization of the negative log-likelihood and the ridge regularization (replacing the term from linear regression in Equation 6.5 with the negative log likelihood function), we can obtain the following minimization problem:

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) &= \text{argmin}_{\vec{\beta}} \left\{ -\ln(\ell(\beta_0, \beta_1, ..., \beta_p)) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \\ &= \text{argmin}_{\vec{\beta}} \left\{ -\ln \left(\prod_{i;y_i=1} p(x_i) \prod_{i':y_{i'}=0} 1 - p(x_i) \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}\end{aligned}$$

b) - To find the system of equations that need to be solved, we first must simplify the negative log-likelihood function. We start by specifying the likelihood function, from our values for $p(x_i)$ and $1 - p(x_i)$ specified in (a):

$$\begin{aligned}\ell(\beta_0, \beta_1, ..., \beta_p) &= \prod_{i;y_i=1} p(x_i) \prod_{i':y_{i'}=0} 1 - p(x_i) \\ &= \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}} \prod_{i':y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_{i'1}+...+\beta_p x_{i'p}}}\end{aligned}$$

From this, we can use properties of natural logs to simplify:

$$\begin{aligned}&-\ln(\ell(\beta_0, \beta_1, ..., \beta_p)) \\ &= -\ln \left(\prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}} \prod_{i':y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_{i'1}+...+\beta_p x_{i'p}}} \right) \\ &= -\left(\sum_{i:y_i=1} \ln \left(\frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}} \right) + \sum_{i':y_i'=0} \ln \left(\frac{1}{1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}} \right) \right) \\ &= -\left(\sum_{i:y_i=1} \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} - \sum_{i=1}^{n} \ln \left(1 + e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}\right) \right) \\ &= \sum_{i=1}^{n} \ln \left(1 + e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}\right) - \sum_{i:y_i=1} (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})\end{aligned}$$

Now, since the natural log function is concave, we know that the negative of the natural log function is convex. Similarly, as the $f(x) = cx^2$, where $c > 0$, is convex; therefore, the minimization problem is minimizing over the sum of two convex functions, which means the sum of those functions itself is convex; thus, to find the minima, we can take the derivative with respect to each and every β_j and set them all to 0. We start with β_0, since there is no β_0 in the regularization term.

$$
\begin{aligned}
0 &= \frac{d}{d\beta_0}\left(-\ln(\ell(\beta_0, \beta_1, ..., \beta_p)) + \lambda \sum_{j=1}^{p} \beta_j^2\right) \\
&= \frac{d}{d\beta_0}(-\ln(\ell(\beta_0, \beta_1, ..., \beta_p))) + \frac{d}{d\beta_0}\left(\lambda \sum_{j=1}^{p} \beta_j^2\right) \\
&= \frac{d}{d\beta_0}(-\ln(\ell(\beta_0, \beta_1, ..., \beta_p))) \\
&= \frac{d}{d\beta_0}\left(\sum_{i=1}^{n} \ln\left(1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}\right) - \sum_{i:y_i=1} (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})\right) \\
&= \left(\sum_{i=1}^{n} \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}\right) - \sum_{i:y_i=1} 1 \\
&= \left(\sum_{i=1}^{n} \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}\right) - n\overline{y}
\end{aligned}
$$

Now, we calculate for β_k, $k \in \{1, 2, ..., p\}$:

$$
\begin{aligned}
0 &= \frac{d}{d\beta_k}\left(-\ln(\ell(\beta_0, \beta_1, ..., \beta_p)) + \lambda \sum_{j=1}^{p} \beta_j^2\right) \\
&= \frac{d}{d\beta_k}(-\ln(\ell(\beta_0, \beta_1, ..., \beta_p))) + \frac{d}{d\beta_k}\left(\lambda \sum_{j=1}^{p} \beta_j^2\right) \\
&= \left(\sum_{i=1}^{n} \frac{x_{ik} e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}\right) - \sum_{i:y_i=1} x_{ik} + 2\lambda\beta_k
\end{aligned}
$$

Thus, we have arrived at the equations we need to solve; there are $p+1$ equations and $p + 1$ unknowns, and they are as follows (they correspond the derivative with respect to each individual coefficient):

$$
\begin{aligned}
0 &= \left(\sum_{i=1}^{n} \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}\right) - n\overline{y} \text{ (derivative w.r.t } \beta_0) \\
0 &= \left(\sum_{i=1}^{n} \frac{x_{ik} e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}\right) - \sum_{i:y_i=1} x_{ik} + 2\lambda\beta_k \ (\forall k \in \{1, 2, ..., p\})
\end{aligned}
$$

c) - Note that, as $\lambda \to \infty$, $2\lambda\beta_k \to \infty$ if $\beta_k > 0$ (similarly $2\lambda\beta_k \to -\infty$ if $\beta_k < 0$). Due to the fact that we have a finite amount of data (a.k.a n is finite), the values for any given feature are finite, and the odds for any given X are in the interval $(0, 1)$, the penalty term will begin to dominate the minimization problem because it has such a much bigger magnitude as $\lambda \to \infty$. Therefore, to accomplish the 0 derivative, we must have it that $\forall j \in \{1, 2, ..., p\}, \hat{\beta}_j = 0$ (this is the only way for the penalty term in any of the p equations that contains it to not shoot off into either ∞ or $-\infty$). We can plug this in to the first equation to solve get an explicit value for $\hat{\beta}_0$:

$$
\begin{aligned}
0 &= \left(\sum_{i=1}^{n} \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}} \right) - n\overline{y} \\
&\Rightarrow 0 = \left(\sum_{i=1}^{n} \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right) - n\overline{y} \\
&\Rightarrow n\overline{y} = \sum_{i=1}^{n} \frac{e^{\beta_0}}{1 + e^{\beta_0}} \\
&\Rightarrow n\overline{y} = n * \frac{e^{\beta_0}}{1 + e^{\beta_0}} \\
&\Rightarrow \overline{y} = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \\
&\Rightarrow \overline{y}(1 + e^{\beta_0}) = e^{\beta_0} \\
&\Rightarrow \overline{y} + \overline{y}e^{\beta_0} = e^{\beta_0} \\
&\Rightarrow e^{\beta_0} - e^{\beta_0}\overline{y} = \overline{y} \\
&\Rightarrow e^{\beta_0}(1 - \overline{y}) = \overline{y} \\
&\Rightarrow e^{\beta_0} = \frac{\overline{y}}{1 - \overline{y}} \\
&\Rightarrow \beta_0 = \ln\left(\frac{\overline{y}}{1 - \overline{y}} \right) = \ln\left(\frac{\text{Number of 1's}}{\text{Number of 0's}} \right)
\end{aligned}
$$

Our resultant value for β_0 is effectively the natural log of the proportion of the number of data points where $y_i = 1$ to the number of data points where $y_i = 0$ (assuming the typical encoding where $\forall i, y_i \in \{0, 1\}$); therefore, we have come up with the following coefficients:

$$
\begin{aligned}
&\text{argmin}_{\vec{\beta}} \left\{ -\ln\left(\prod_{i; y_i = 1} p(x_i) \prod_{i' : y_{i'} = 0} 1 - p(x_i) \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \lambda \to \infty \\
&= \left(\ln\left(\frac{\overline{y}}{1 - \overline{y}} \right), 0, 0, \ldots, 0 \right)
\end{aligned}
$$

So, $\hat{\beta}_0 = \ln\left(\frac{\overline{y}}{1-\overline{y}}\right)$, and $\forall j > 0, \hat{\beta}_j = 0$.

2. This problem will essentially mirror problem 7.9, Exercise 1 from the textbook, so please refer to it when answering. The only difference is that we will consider a quadratic spline, rather than a cubic one. So, the function we will consider is $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x-\xi)_+^2$, and in parts (a) and (b) you can drop the cubic term from $f_1(x)$ and $f_2(x)$. Then simply follow the 7.9 Exercise 1 with this change, answering parts (a)-(d).

Solution

a) - For $f_1(x)$, our goal is to make it so $\forall x \leq \xi$:

$$f_1(x) = a_1 + b_1 x + c_1 x^2 = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x-\xi)_+^2 = f(x)$$

Note that $\forall x \leq \xi$, $(x-\xi)_+^2 = 0$; therefore:

$$f_1(x) = a_1 + b_1 x + c_1 x^2 = \beta_0 + \beta_1 x + \beta_2 x^2$$

Via inspection, this can be achieved via setting the following equal:

$$a_1 = \beta_0 \text{ and } b_1 = \beta_1 \text{ and } c_1 = \beta_2$$

Thus, $\forall x \leq \xi, f_1(x) = a_1 + b_1 x + c_1 x^2 = \beta_0 + \beta_1 x + \beta_2 x^2 = f(x)$.

b) - Via the definition of the spline, we note that:

$$\forall x > \xi, (x-\xi)_+^2 = (x-\xi)^2$$

Thus, we need to do some algebraic manipulation to get this equation into the correct form for the fitting process;

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x-\xi)_+^2 \\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x-\xi)^2 \\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x^2 - 2x\xi + \xi^2) \\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^2 - 2\beta_3 \xi x + \beta_3 \xi^2 \\ &= (\beta_0 + \beta_3 \xi^2) + (\beta_1 - 2\beta_3 \xi) x + (\beta_2 + \beta_3) x^2 \end{aligned}$$

Via inspection, our fitting can be achieved with the following coefficients:

$$a_2 = \beta_0 + \beta_3 \xi^2 \text{ and } b_2 = \beta_1 - 2\beta_3 \xi \text{ and } c_2 = \beta_2 + \beta_3$$

Therefore:

$$\forall x > \xi, f_2(x) = a_2 + b_2 x + c_2 x^2 = (\beta_0 + \beta_3 \xi^2) + (\beta_1 - 2\beta_3 \xi) x + (\beta_2 + \beta_3) x^2 = f(x)$$

c) - We must plug ξ into both equations and show equality to show continuity:

$$\begin{aligned} f_2(\xi) &= a_2 + b_2(\xi) + c_2(\xi)^2 \\ &= (\beta_0 + \beta_3 \xi^2) + (\beta_1 - 2\beta_3 \xi)(\xi) + (\beta_2 + \beta_3)(\xi)^2 \\ &= \beta_0 + \beta_3 \xi^2 + \beta_1 \xi - 2\beta_3 \xi^2 + \beta_2 \xi^2 + \beta_3 \xi^2 \\ &= \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^2 - 2\beta_3 \xi^2 + \beta_3 \xi^2 \\ &= \beta_0 + \beta_1(\xi) + \beta_2(\xi)^2 = a_1 + b_1 \xi + c_1 \xi^2 = f_1(\xi) \end{aligned}$$

d) - We must first take the derivative of both equations:

$$\begin{aligned} f_1'(x) &= \frac{d}{dx}\left(a_1 + b_1 x + c_1 x^2\right) \\ &= b_1 + 2c_1 x \\ &= \beta_1 + 2\beta_2 x \end{aligned}$$

Similarly, for $f_2(x)$:

$$\begin{aligned} f_2'(x) &= \frac{d}{dx}\left(a_2 + b_2 x + c_2 x^2\right) \\ &= b_1 + 2c_1 x \\ &= \beta_1 - 2\beta_3\xi + 2(\beta_2 + \beta_3)x \\ &= \beta_1 - 2\beta_3\xi + (2\beta_2 + 2\beta_3)x \end{aligned}$$

Therefore, when we plug in ξ:

$$\begin{aligned} f_2'(\xi) &= \beta_1 - 2\beta_3\xi + (2\beta_2 + 2\beta_3)(\xi) \\ &= \beta_1 - 2\beta_3\xi + 2\beta_2\xi + 2\beta_3\xi \\ &= \beta_1 + 2\beta_2\xi - 2\beta_3\xi + 2\beta_3\xi \\ &= \beta_1 + 2\beta_2(\xi) \\ &= f_1'(\xi) \end{aligned}$$

Thus, $f_2'(\xi) = f_1'(\xi)$, so $f'(x)$ is continuous; since we showed $f(x)$ is continuous in part c), we can deduce that $f(x)$ is indeed a quadratic spline.

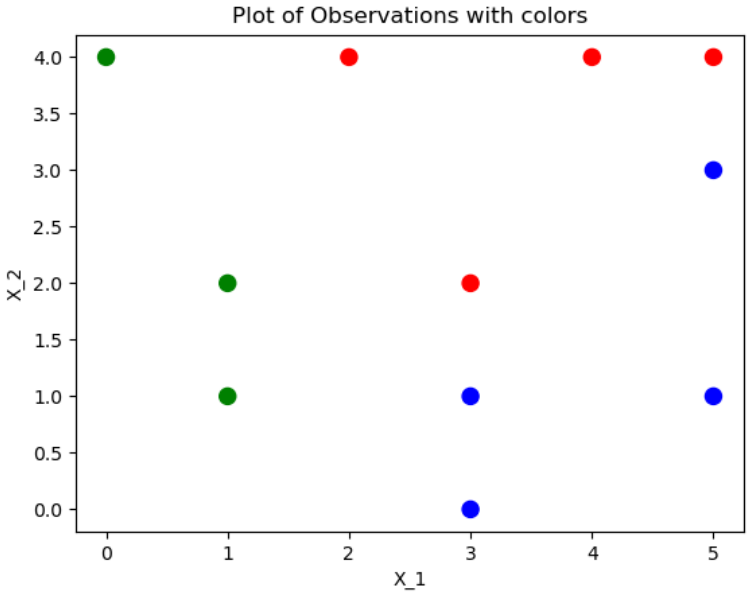
3. You are given $n = 11$ data points in $p = 2$ dimensions, where each point is classified into one of $K = 3$ classes. The data is in the table shown below:

Point	X_1	X_2	Class
1	3	2	Red
2	2	4	Red
3	4	4	Red
4	5	4	Red
5	3	0	Blue
6	3	1	Blue
7	5	1	Blue
8	5	3	Blue
9	1	1	Green
10	1	2	Green
11	0	4	Green

a) Make a neat, scaled plot of the data, using different colors and/or markers for each of the three class types. You can use a computer to construct this plot.
b) Now suppose you are going to use one-versus-one classification on this problem, as described in the lecture and the textbook. Find the optimal separating hyperplane (line) for each pair of classes. Draw these on your plot from (a), and give the formula for each in the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$ for the appropriate values of β in each case. You must show your work and reasoning here to explain how you arrived at these formulas - simply reporting the resutls of a computer implementation will not suffice.
c) Following up from part (b), indicate on your plot (via shading, for example) the regions of feature space where test points should be attributed to each of the 3 classes according to standard one-versus-one classification. If you've done everything correctly, you will find a region where standard one-versus-one classification will fail to assign a class. Indicate this region on your plot as well. We will refer to this region as Z below.
d) Suppose that within Z, we offer a modified rule to assign a class to a test point. For the test point in question, compute how far it is from the closest point within the region within each of the regions of known classification (Red, Green, Blue), then assign it to the class whose region is closest to the point. Using this rule, update your plot to graphically indicate within Z which subregions will be classified as each of the given classes. Be sure to explain your reasoining, but you do not have to provide any specific equations of lines, etc.
e) Find the coordinates of the only point within Z that still could not be attributed to any of the 3 classes. Be sure to show your work.

Solution

a) - The original plot of the data with the colors labelled is shown on the page below:



b) - In the textbook, section 9.1.4 explains the optimization problem that constructs the maximal margin classifier:

$$\begin{aligned}&\text{argmax}_{\beta_0,\beta_1,\beta_2,M} M\\ &\text{subject to } \beta_1^2+\beta_2^2=1\\ &\text{where } y_i(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}) \geq M \ \forall i=1,...,n\end{aligned}$$

To this problem, we effectively have to show the support vector machine between each pair of classes and then find the line that serves as the maximal margin classifier.

Pair 1 - Green and Red

First, we must find the points that are in the green class and red class respectively that are closest in euclidean distance to each other. Note that observations 9, 10, and 11 are green and observations 1-4 are red, so, calling the i'th observation x_i, we test out each pair of points with x_9 and a red point:

$$\begin{aligned}||x_9-x_1||_2 &= \sqrt{(1-3)^2+(1-2)^2} = \sqrt{4+1} = \sqrt{5}\\ ||x_9-x_2||_2 &= \sqrt{(1-2)^2+(1-4)^2} = \sqrt{1+9} = \sqrt{10}\\ ||x_9-x_3||_2 &= \sqrt{(1-4)^2+(1-4)^2} = \sqrt{9+9} = \sqrt{18}\\ ||x_9-x_4||_2 &= \sqrt{(1-5)^2+(1-4)^2} = \sqrt{16+9} = 5\end{aligned}$$

Then, we do the same thing with x_{10}:

$$||x_{10} - x_1||_2 = \sqrt{(1-3)^2 + (2-2)^2} = \sqrt{4} = 2$$
$$||x_{10} - x_2||_2 = \sqrt{(1-2)^2 + (2-4)^2} = \sqrt{1+4} = \sqrt{5}$$
$$||x_{10} - x_3||_2 = \sqrt{(1-4)^2 + (2-4)^2} = \sqrt{9+4} = \sqrt{13}$$
$$||x_{10} - x_4||_2 = \sqrt{(1-5)^2 + (2-4)^2} = \sqrt{16+4} = \sqrt{20}$$

Finally, we examine the distances of red points from x_{11}:

$$||x_{11} - x_1||_2 = \sqrt{(0-3)^2 + (4-2)^2} = \sqrt{9+4} = \sqrt{13}$$
$$||x_{11} - x_2||_2 = \sqrt{(0-2)^2 + (4-4)^2} = \sqrt{4} = 2$$
$$||x_{11} - x_3||_2 = \sqrt{(0-4)^2 + (4-4)^2} = \sqrt{16} = 4$$
$$||x_{11} - x_4||_2 = \sqrt{(0-5)^2 + (4-4)^2} = \sqrt{25} = 5$$

So the pairs of points that are closest to each other in the opposite classes are (x_1, x_{10}) and (x_2, x_{11}). These form the support vector machine. Let us first examine (x_1, x_{10}). Note that $x_1 = (3, 2)$ and $x_{10} = (1, 2)$. So, $x_{1,2} = x_{10,2} = 2$. Since they are in opposing classes, any seperating hyperplane has to have an X_1 value in between 1 and 3 when $X_2 = 2$. To maximize the margin, we note that the distance between x_1 and x_{10} is 2; thus $y(x_1) + y(x_{10}) = 2$; to maximize the minimum distance of either x_1 or x_{10}, we note that, if $y(x_1) > 1$, $y(x_2) < 1$ (and visa versa), so $M < 1$ (the minimum distance from a point is less than 1). To overcome this hurdle, can we set $y(x_1) = y(x_{10}) = 1$? In fact, we can, if the optimally separating hyperplane goes through the point $(2, 2)$, which has a euclidean distance of 1 from both x_1 and x_2. Similarly, we have that $x_2 = (2, 4)$ and $x_{11} = (0, 4)$, and $x_{2,2} = x_{11,2} = 4$. These points also have a distance of 2; to make sure that the margin is 1 when $X_2 = 4$, we must have the distance between both points be 1, so the line must go through the point $(1, 4)$, which has a euclidean distance of 1 from both x_2 and x_{11}; thus, we have two points and can create our optimal separating hyperplane. We can think of our line as $X_2 = \beta_0 + \beta_1 X_1$, setting $\beta_2 = -1$ for now. So, given that our line has to pass through $(1, 4)$ and $(2, 2)$:

$$\begin{aligned}
&4 = \beta_0 + 1\beta_1 \text{ and } 2 = \beta_0 + 2\beta_1 \\
&\Rightarrow (4-2) = (\beta_0 - \beta_0) + (1\beta_1 - 2\beta_1) = -\beta_1 \\
&\Rightarrow 2 = -\beta_1 \rightarrow \beta_1 = -2 \\
&\Rightarrow 2 = \beta_0 + 2(-2) \rightarrow 2 = \beta_0 - 4 \\
&\Rightarrow \beta_0 = 6
\end{aligned}$$

So, we have that $6 - 2X_1 - X_2 = 0$. Note that $\beta_1 = 2\beta_2$ and they are both negative. To help satisfy the $\beta_1^2 + \beta_2^2 = 1$ constraint, we can use this:

$$\beta_1^2 + \beta_2^2 = 1 \Rightarrow (2\beta_2)^2 + \beta_2^2 = 1 \Rightarrow 5\beta_2^2 = 1 \rightarrow \beta_2 = -\frac{\sqrt{5}}{5} \rightarrow \beta_1 = -\frac{2\sqrt{5}}{5}$$

Thus, after normalizing through multiplication by $\frac{\sqrt{5}}{5}$ our maximally separating hyperplane for the red and green classes is:

$$\frac{6\sqrt{5}}{5} - \frac{2\sqrt{5}}{5}X_1 - \frac{\sqrt{5}}{5}X_2 = 0$$

Pair 2 - Green and Blue

Note that observations 9, 10, and 11 are green and observations 5-8 are blue, so, calling the i'th observation x_i, we test out each pair of points with x_9 and a blue point:

$$||x_9 - x_5||_2 = \sqrt{(1-3)^2 + (1-0)^2} = \sqrt{4+1} = \sqrt{5}$$
$$||x_9 - x_6||_2 = \sqrt{(1-3)^2 + (1-1)^2} = \sqrt{4} = 2$$
$$||x_9 - x_7||_2 = \sqrt{(1-5)^2 + (1-1)^2} = \sqrt{16} = 4$$
$$||x_9 - x_8||_2 = \sqrt{(1-5)^2 + (1-3)^2} = \sqrt{16+4} = \sqrt{20}$$

Then, we do the same thing with x_{10}:

$$||x_{10} - x_5||_2 = \sqrt{(1-3)^2 + (2-0)^2} = \sqrt{4+4} = \sqrt{8}$$
$$||x_{10} - x_6||_2 = \sqrt{(1-3)^2 + (2-1)^2} = \sqrt{4+1} = \sqrt{5}$$
$$||x_{10} - x_7||_2 = \sqrt{(1-5)^2 + (2-1)^2} = \sqrt{16+1} = \sqrt{17}$$
$$||x_{10} - x_8||_2 = \sqrt{(1-5)^2 + (2-3)^2} = \sqrt{16+1} = \sqrt{17}$$

Finally, we examine the distances of blue points from x_{11}:

$$||x_{11} - x_5||_2 = \sqrt{(0-3)^2 + (4-0)^2} = \sqrt{9+16} = 5$$
$$||x_{11} - x_6||_2 = \sqrt{(0-3)^2 + (4-1)^2} = \sqrt{9+9} = \sqrt{18}$$
$$||x_{11} - x_7||_2 = \sqrt{(0-5)^2 + (4-1)^2} = \sqrt{25+9} = \sqrt{34}$$
$$||x_{11} - x_8||_2 = \sqrt{(0-5)^2 + (4-3)^2} = \sqrt{25+1} = \sqrt{26}$$

So the points that are closest to each other in the opposing classes are (x_6, x_9); these form the support vector. Note that $x_{6,2} = x_{9,2} = 1$. Therefore, when $X_2 = 1$, the hyperplane has to have an X_1 value in between $x_{6,1} = 3$ and $x_{9,1} = 1$; as in the green-red seperating hyperplane, the distance between these two closest points is 2; therefore, to maximize the margin (setting it to 1, as in the green-red separating hyperplane), we can necessitate that the hyperplane pass through the midpoint, or $(2, 1)$, which has a euclidean distance of 1 from both x_6 and x_9. Note that, since x_6 and x_9 are the only support vectors, these are the only points that matter in construction of the hyperplane. In order to make the hyperplane never get closer to either point, we must have its slope be perpendicular to the line containing x_6 and x_9, which is just a horizontal line. Thus, our hyperplane is a vertical line passing through $(2, 1)$, which can be summed up by $X_1 = 2$; or, in other words, our hyperplane is the following:

$$2 - X_1 = 0$$

Pair 3 - Red and Blue

Observations 1-4 are red, and Observations 5-8 are Blue; we must calculate the distance between each pair. We start with the distances from x_1:

$$||x_1 - x_5||_2 = \sqrt{(3-3)^2 + (2-0)^2} = \sqrt{4} = 2$$
$$||x_1 - x_6||_2 = \sqrt{(3-3)^2 + (2-1)^2} = \sqrt{1} = 1$$
$$||x_1 - x_7||_2 = \sqrt{(3-5)^2 + (2-1)^2} = \sqrt{4+1} = \sqrt{5}$$
$$||x_1 - x_8||_2 = \sqrt{(3-5)^2 + (2-3)^2} = \sqrt{4+1} = \sqrt{5}$$

Then, x_2:

$$||x_2 - x_5||_2 = \sqrt{(2-3)^2 + (4-0)^2} = \sqrt{1+16} = \sqrt{17}$$
$$||x_2 - x_6||_2 = \sqrt{(2-3)^2 + (4-1)^2} = \sqrt{1+9} = \sqrt{10}$$
$$||x_2 - x_7||_2 = \sqrt{(2-5)^2 + (4-1)^2} = \sqrt{9+9} = \sqrt{18}$$
$$||x_2 - x_8||_2 = \sqrt{(2-5)^2 + (4-3)^2} = \sqrt{9+1} = \sqrt{10}$$

Now, x_3:

$$||x_3 - x_5||_2 = \sqrt{(4-3)^2 + (4-0)^2} = \sqrt{1+16} = \sqrt{17}$$
$$||x_3 - x_6||_2 = \sqrt{(4-3)^2 + (4-1)^2} = \sqrt{1+9} = \sqrt{10}$$
$$||x_3 - x_7||_2 = \sqrt{(4-5)^2 + (4-1)^2} = \sqrt{1+9} = \sqrt{10}$$
$$||x_3 - x_8||_2 = \sqrt{(4-5)^2 + (4-3)^2} = \sqrt{1+1} = \sqrt{2}$$

Finally, x_4:

$$||x_4 - x_5||_2 = \sqrt{(5-3)^2 + (4-0)^2} = \sqrt{4+16} = \sqrt{20}$$
$$||x_4 - x_6||_2 = \sqrt{(5-3)^2 + (4-1)^2} = \sqrt{4+9} = \sqrt{13}$$
$$||x_4 - x_7||_2 = \sqrt{(5-5)^2 + (4-1)^2} = \sqrt{9} = 3$$
$$||x_4 - x_8||_2 = \sqrt{(5-5)^2 + (4-3)^2} = \sqrt{1} = 1$$

The closest pair of points are (x_1, x_6) and (x_4, x_8), both with a distance of 1. These are the support vectors. Note that $x_{1,1} = x_{6,1} = 3$; thus, when $X_1 = 3$, the X_2 value must be between $x_{1,2} = 2$ and $x_{6,2} = 1$. The distance between these two points is 1; to ensure maximal margins, we can get, guaranteed, up to a $\frac{1}{2}$ distance if have the hyperplane pass through the midpoint of the vertical line passing through both of these points, or $(3, 1.5)$. Similarly, $x_{4,1} = x_{8,1} = 5$; thus, to separate these two points, the corresponding X_2 value on the hyperplane must be between $x_{4,2} = 4$ and $x_{8,2} = 3$. The distance between these two points is 1; to ensure maximal margins, we can get, guaranteed, up to a $\frac{1}{2}$ distance if have the hyperplane pass through the midpoint of the vertical line going through both of these points, or $(5, 3.5)$. As in the Green-red separator, we can think of

our line as $X_2 = \beta_0 + \beta_1 X_1$, and set $\beta_2 = -1$ for now. So, given that our line has to pass through $(3, 1.5)$ and $(5, 3.5)$:

$$\begin{aligned}
&1.5 = \beta_0 + 3\beta_1 \text{ and } 3.5 = \beta_0 + 5\beta_1 \\
&\Rightarrow (3.5 - 1.5) = (\beta_0 - \beta_0) + (5\beta_1 - 3\beta_1) = 2\beta_1 \\
&\Rightarrow 2 = 2\beta_1 \rightarrow \beta_1 = 1 \\
&\Rightarrow 1.5 = \beta_0 + 3(1) \rightarrow 1.5 = \beta_0 + 3 \Rightarrow \beta_0 = -1.5
\end{aligned}$$

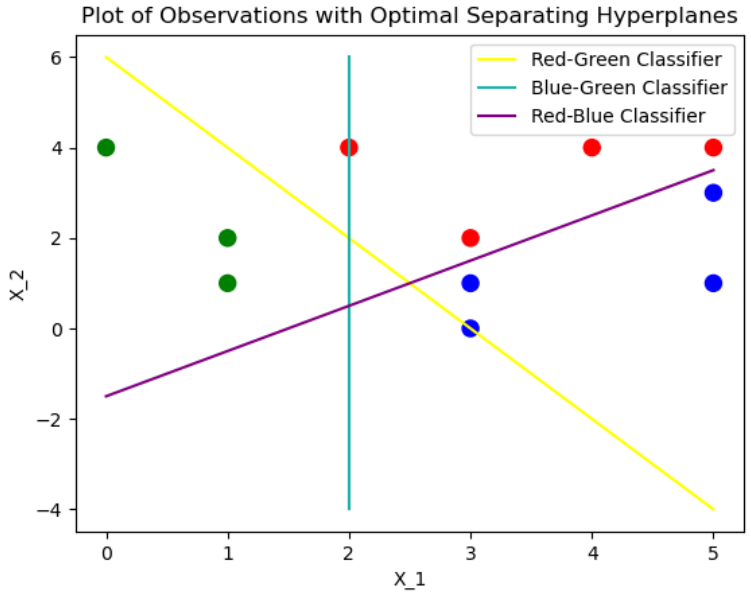
So, we have that $-1.5 + X_1 - X_2 = 0$. Note that $\beta_1 = -\beta_2$. To help satisfy the $\beta_1^2 + \beta_2^2 = 1$ constraint, we can use this and the fact that we set β_2 to be negative:

$$\beta_1^2 + \beta_2^2 = 1 \Rightarrow (-\beta_2)^2 + \beta_2^2 = 1 \Rightarrow 2\beta_2^2 = 1 \rightarrow \beta_2 = -\frac{\sqrt{2}}{2} \wedge \beta_1 = \frac{\sqrt{2}}{2}$$

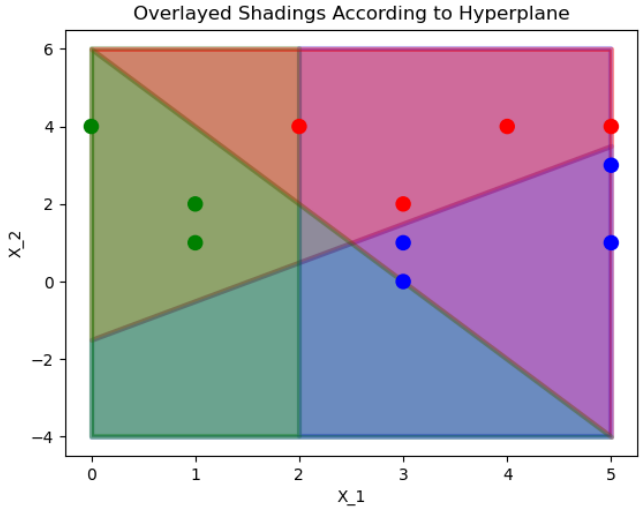
So, in other words, normalizing gets us the following hyperplane:

$$\frac{-1.5\sqrt{2}}{2} + \frac{\sqrt{2}}{2} X_1 - \frac{\sqrt{2}}{2} X_2 = 0$$

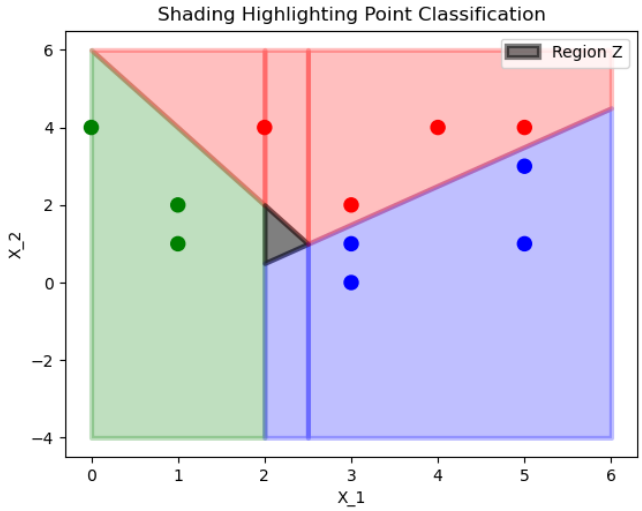
We can plot all three lines on our original plot:



c) - Since we have three classifying lines, they each make a decision on what to classify a point as at each point. Here are those three classification maps overlayed on top of each other as shadings of the color they classify a given point in the feature space as:



Note that, for any potential test point, no color is selected by all 3 classifiers because each classifier is only accounting for 2 of the 3 colors. Therefore, our final classifier will classify a given test point based on what the majority - that is, 2 out of 3 - of classifying hyper planes say that that point should be. Here is the graph highlighting the classifications with this rule:



Note the black region in the middle that we have classsified as region Z. This region is colored in this manner because any test point in this region is classified as a different color by each of the three hyperplane classifiers.

d) - Note that region Z is a triangle whose bounds are dictated by our separating

hyper planes. To identify, within the triangle, what side a given test point is closest to, and therefore its classification, we need to first identify the point equidistant from all sides, or the incenter. To do this, we must compute the coordinates of the vertices of this triangle. We have the vertices $\{v_{rg}, v_{gb}, v_{rb}$ where the subscript represents the color classifications converging at that point.
Vertex 1 - v_{rg}
The two lines that are converging at this point are $X_1 = 2$ and $X_2 = 6 - 2X_1$. So, we can solve this system of two equations and two unknowns to get the exact coordinates:

$$X_1 = 2 \Rightarrow X_2 = 6 - 2(2) = 2 \Rightarrow v_{rg} = (2, 2)$$

Vertex 2 - v_{gb}
The two lines that are converging at this point are $X_1 = 2$ and $X_2 = -1.5 + X_1$. So, solving this system of two equations and two unknowns:

$$X_1 = 2 \Rightarrow X_2 = -1.5 + 2 = 0.5 \Rightarrow v_{gb} = (2, 0.5)$$

Vertex 3 - v_{rb}
The two lines that are converging at this point are $X_2 = 6 - 2X_1$ and $X_2 = -1.5 + X_1$. So, solving this system of two equations and two unknowns:

$$\begin{aligned}
&X_2 = -1.5 + X_1 \text{ and } X_2 = 6 - 2X_1 \\
&\Rightarrow (X_2 - X_2) = (-1.5 - 6) + (X_1 - -2X_1) \\
&\Rightarrow 0 = -7.5 + 3X_1 \\
&\Rightarrow 7.5 = 3X_1 \rightarrow X_1 = 2.5 \rightarrow X_2 = -1.5 + 2.5 = 1 \\
&\Rightarrow v_{rb} = (2.5, 1)
\end{aligned}$$

We also note that we have three edges; to calculate the incenter, we need the lengths of each edge. Call the edges $\{e_g, e_r, e_b\}$ to represent which color of classification borders said edge. We have that $e_g = v_{gb}v_{rg}$, so the endpoints are $(2, 0.5)$ and $(2, 2)$. Thus, using the formula for euclidean distance between two points:

$$|e_g| = \sqrt{(2-2)^2 + (2-0.5)^2} = \sqrt{(1,5)^2} = 1.5 = \frac{3}{2}$$

Next, we calculate the length of the edge on the section of the feature space classified as red. Here $e_r = v_{rb}v_{rg}$. so the endpoints are $(2.5, 1)$ and $(2, 2)$. Thus, using euclidean distance:

$$|e_r| = \sqrt{(2.5-2)^2 + (2-1)^2} = \sqrt{.5^2 + 1^2} = \sqrt{1.25} = \frac{\sqrt{5}}{2}$$

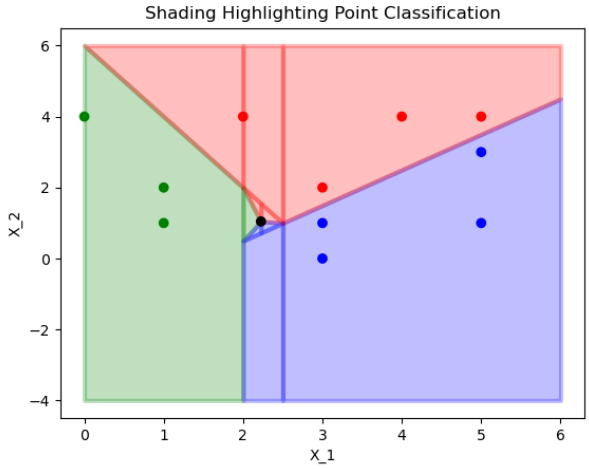
Finally, we calculate the length of the edge on the section of the feature space classified as blue. Here $e_b = v_{rb}v_{gb}$, so the endpoints are $(2.5, 1)$ and $(2, 0.5)$, meaning that the length of the edge is:

$$|e_b| = \sqrt{(2.5-2)^2 + (1-.5)^2} = \sqrt{.5^2 + .5^2} = \sqrt{.5} = \frac{\sqrt{2}}{2}$$

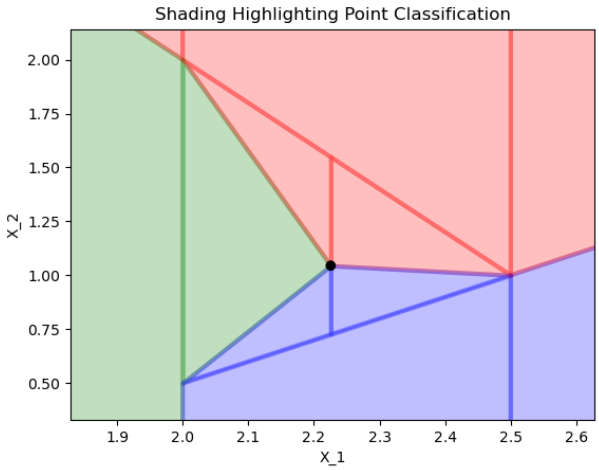
Now that we have the coordinates of the vertices, and the lengths of the edges, we can use the formula for the incenter to calculate (note that, in the numerator of each coordinate, each vertex coordinate is multiplied by the edge length of the edge it is not adjacent to):

$$
\begin{aligned}
&\text{Incenter} \\
&= \left(\frac{|e_b| v_{rg}(x) + |e_r| v_{gb}(x) + |e_g| v_{rb}(x)}{|e_g| + |e_r| + |e_b|}, \frac{|e_b| v_{rg}(y) + |e_r| v_{gb}(y) + |e_g| v_{rb}(y)}{|e_g| + |e_r| + |e_b|} \right) \\
&= \left(\frac{\frac{\sqrt{2}}{2} * 2 + \frac{\sqrt{5}}{2} * 2 + \frac{3}{2} * 2.5}{\frac{3}{2} + \frac{\sqrt{5}}{2} + \frac{\sqrt{2}}{2}}, \frac{\frac{\sqrt{2}}{2} * 2 + \frac{\sqrt{5}}{2} * 0.5 + \frac{3}{2} * 1}{\frac{3}{2} + \frac{\sqrt{5}}{2} + \frac{\sqrt{2}}{2}} \right) \\
&= \left(\frac{\sqrt{2} + \sqrt{5} + \frac{15}{4}}{\frac{3+\sqrt{5}+\sqrt{2}}{2}}, \frac{\sqrt{2} + \frac{\sqrt{5}}{4} + \frac{3}{2}}{\frac{3+\sqrt{5}+\sqrt{2}}{2}} \right) \\
&= \left(\frac{\frac{4\sqrt{2}+4\sqrt{5}+15}{4}}{\frac{3+\sqrt{5}+\sqrt{2}}{2}}, \frac{\frac{4\sqrt{2}+\sqrt{5}+6}{4}}{\frac{3+\sqrt{5}+\sqrt{2}}{2}} \right) \\
&= \left(\frac{\frac{4\sqrt{2}+4\sqrt{5}+15}{2}}{3+\sqrt{5}+\sqrt{2}}, \frac{\frac{4\sqrt{2}+\sqrt{5}+6}{2}}{3+\sqrt{5}+\sqrt{2}} \right) \\
&= \left(\frac{4\sqrt{2}+4\sqrt{5}+15}{6+2\sqrt{5}+2\sqrt{2}}, \frac{4\sqrt{2}+\sqrt{5}+6}{6+2\sqrt{5}+2\sqrt{2}} \right) \approx (2.22555, 1.04454)
\end{aligned}
$$

Since this point is equidistant to all edges, we can draw three lines, one starting at each vertex of the triangle, and the other end being the incenter. This creates three smaller triangles within that triangle, and each triangle corresponds to the region of points within the larger triangle that is closest to the color classification region whose edge it shares. If these lines are drawn and the space in the smaller triangles shaded according to proximity to already shaded areas, the following graph is created:



The incenter is colored black. Zooming into the filled in triangle, we see the new classification:



e) - Note that, within Z, there is exactly one point that is not closer to any of the sides of the triangle: the incenter. We already calculated the coordinates of the incenter, so the coordinates of the only still unclassified point are:

$$\text{Unclassified Point} = \left(\frac{4\sqrt{2} + 4\sqrt{5} + 15}{6 + 2\sqrt{5} + 2\sqrt{2}}, \frac{4\sqrt{2} + \sqrt{5} + 6}{6 + 2\sqrt{5} + 2\sqrt{2}} \right) \approx (2.22555, 1.04454)$$

2 Programming Problems

4.
a) What movie has the lowest average rating in the Ratings data, and what is that rating? Which has the highest, and what is it? Which movie was rated by the largest number of users, and how many? Which was rated by the lowest number, and how many? Which user ID rated the most movies, and how many did they rate? Which user ID rated the fewest movies, and how many did they rate?
b) Make a plot of the value of the objective as a funciton of iteration number and display it. What value of the objective does the algorithm seem to settle down to?
c) Given your final obtained $\tilde{X}$, what movie has the highest average rating, and what is the rating? What movie has the lowest average rating, and what is the rating?
d) Make a plot that displays the average rating of each movie based of $\tilde{X}$ versus its average rating based off of X. Considering this, which movie is the most over-rated: that movie whose difference in average rating between X and $\tilde{X}$ is the largest? Which movie is the most underrated: that mvoie whose difference in average rating from X to $\tilde{X}$ is the smallest (most negative)?
e) Make a scatter plot of the data in $\hat{B}$. including the cluster membership of each point graphically (different colors or plot markers for the different clusters).
f) Consider the cluster you found with the fewest members. List the movies in this cluster. Then do a little online searching, and speculate as to what these movies might have in common.
Solution
a) - This was done in `final_coding.py`. The following results were generated:

```
Name of Lowest Average Rated Movie: The Avengers
Rating of Lowest Average Rated Movie: 2.3242574257425743

Name of Highest Average Rated Movie: Inu-Yasha
Rating of Highest Average Rated Movie: 4.533980582524272

Name of Most Rated Movie: American Beauty
Number of times this Movie was rated: 10197

Name of Least Rated Movie: Inu-Yasha
Number of times this Movie was rated: 103

User ID that Rated the Most Movies: 11468
Number of Movies this user Rated: 131

User ID(s) that Rated the Least Movies: [2399 4168 6547]
Number of Movies this user rated: 0
```

b) -