1. In class we discussed filter methods and wrapper methods for feature selection. Discuss the pros and cons of each approach. In your answer, please make sure to explain at least one advantage and at least one disadvantage of wrapper methods as compared to filter methods.

2. Consider classification with linear threshold units, where we predict Yes \((y = +1)\) when \(\sum_k w_k x_k \geq 0\) and No \((y = -1)\) otherwise. Now consider a 2-dimensional space where the example \((1, 3)\) is positive and where \((2, -5)\) is negative.

What is the range of possible \(w\) vectors that agree with these labels? Follow the example from class and draw the 2-dimensional space in terms of \(w_1\) and \(w_2\) identifying that space. Please make sure to explain your answer.

3. Professor Polynomial proposed to use neural networks to approximate real valued functions by using a concrete polynomial \(\sigma(s) = 6s^2 - 3s + 2\) as the activation function. In other words a node with inputs \(\vec{x} = (x_1, \ldots, x_n)\) and corresponding weights \(\vec{w} = (w_1, \ldots, w_n)\) will output \(o = \sigma(\vec{w} \cdot \vec{x}) = \sigma(\sum w_i x_i)\).

Recall that in class we derived the update formula of the Perceptron with Margin as a gradient descent step on the “hinge loss” error function. Similarly, online logistic regression update was derived as a gradient ascent step on the log likelihood (so the log likelihood can be thought of as a negated error function).

Develop an online gradient descent algorithm for a single node of this type when using the error function \(E = \frac{1}{2} [y - \sigma(\vec{w} \cdot \vec{x})]^2\). In particular, given a training example \((\vec{x}, y)\) use the error function \(E = \frac{1}{2} [y - \sigma(\vec{w} \cdot \vec{x})]^2\) to derive the gradient descent update for \(\vec{w}\).