Motivation & Problem Statement

- Background: Breast cancer, when abnormal breast cells grow and form tumors, is a prevalent type of cancer affecting many women worldwide. In fact, 670,000 people have died from breast cancer in 2022. Thus, it is important to know if tumors are benign or malignant to determine if an individual should receive treatment.
- Project Goal: Predict whether tumors are benign (label -1) or malignant (label +1) using continuous feature variables extracted from tumors.
- Questions of interest:
  - Can tumors be classified with > 90% testing accuracy using 31 predictor variables?
  - Which 3 feature variables are best able to classify tumors?
  - How fast is the algorithm able to converge?
- Classification method: Single layer perceptron of equation $\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} = 0$ to linearly separate data points as benign or malignant.
- Problem Statement: Let $(\boldsymbol{x}_i, y_i)$ be an observation with predictors $\boldsymbol{x}_i$ and

  corresponding label y_i. Define the classifier function $C_{\boldsymbol{w}}(\boldsymbol{x}) = \begin{cases} -1, \ \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} < 0 \\ \ 1, \ \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} \geq 0 \end{cases}$.

- Denote p as the number of predictor variables and n_{te} the number of testing observations. We seek $\boldsymbol{w}^* \in \mathbb{R}^p$ such that

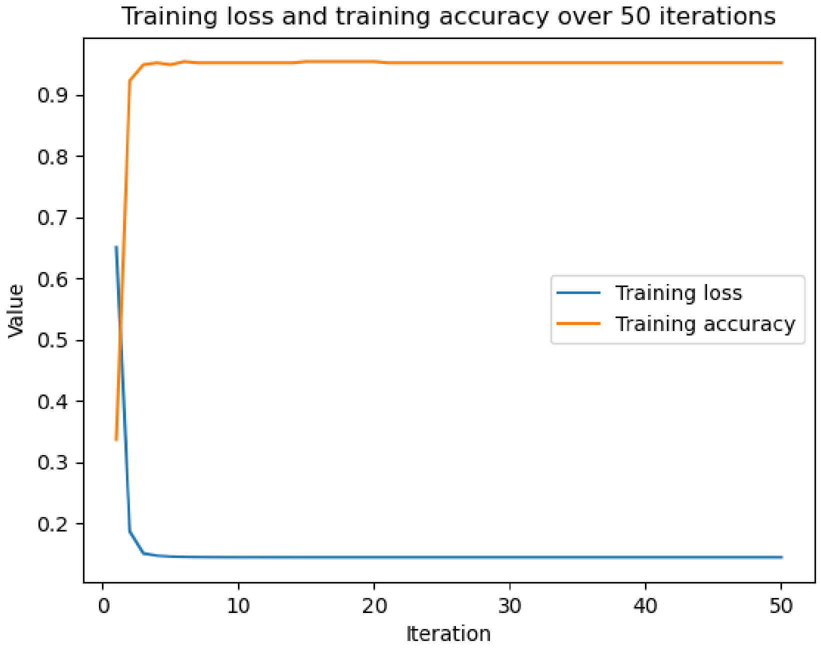
  $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \frac{\sum_{i=1}^{n_{te}} I(C_{\boldsymbol{w}}(\boldsymbol{x}_{\boldsymbol{i}}^{(te)}) = y_i^{(te)})}{n}$, where I is the indicator function. This is the weight with the highest test accuracy, which will best linearly separate the data points.

Methods

- The single layer perceptron will generate a hyperplane to linearly separate the data points by performing the gradient descent algorithm on the mean squared error (MSE) loss function $f(\boldsymbol{w}) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} f_i(\boldsymbol{w}) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} (\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}} - y_i)^2$, where n_{tr} is the number of training observations.
- First, the algorithm will normalize the data matrix $\boldsymbol{X}$. This has been shown to reduce the condition number (ratio of the eigenvalues) of the Hessian, hence speeding up the convergence of the algorithm.
- After selecting an initial weight value of $\boldsymbol{w}^{(0)}$, perform the iterative gradient descent algorithm of $\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \alpha_k \nabla f(\boldsymbol{w}^{(k)})$ to update the weights for iteration k.
- $\nabla f(\boldsymbol{w}^{(k)})$ is chosen as the direction vector because the direction of maximum decrease of a function f at point $\boldsymbol{w}^{(k)}$ has been shown to be $-\nabla f(\boldsymbol{w}^{(k+1)})$ using the Cauchy-Schwarz inequality.
- The learning rate α_k is chosen such that $\alpha_k = \arg\min_{\alpha} f(\boldsymbol{w}^{(k)} - \alpha \nabla f(\boldsymbol{w}^{(k)})$. Since the objective function $f(\boldsymbol{w})$ is a quadratic of the form $f(\boldsymbol{w}) = \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{Q} \boldsymbol{w} - \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{b} + c$, it can be shown that $\alpha_k = \frac{\nabla f(\boldsymbol{w}^{(k)})^T \nabla f(\boldsymbol{w}^{(k)})}{\nabla f(\boldsymbol{w}^{(k)})^T \boldsymbol{Q} \nabla f(\boldsymbol{w}^{(k)})}$ by using the first order necessary conditions for minimizers that $\nabla f(\boldsymbol{w}) = \boldsymbol{0}$.
- The previous conditions guarantee that $f\left(\boldsymbol{w}^{(k+1)}\right) < f(\boldsymbol{w}^{(k)})$, which means the algorithm will converge to the optimal solution.
- Repeat the algorithm for K iterations and select $\boldsymbol{w}^*$ as defined previously.

Results

- Tumors were classified sufficiently well when all 31 predictor variables were used. Using $\boldsymbol{w}^*$ to generate the hyperplane, a testing and training accuracy of 94.7% and 95.4% was achieved, respectively.
- The 3 feature variables best able to classify tumors were “area_mean”, “texture_worst”, and “concave points_worst” with a testing and training accuracy of 96.5% and 92.3%, respectively. Surprisingly, a much smaller feature set had better performance on unseen data rather than the full feature set.
- The algorithm flatlined to a fixed training accuracy values after about $K = 10$ iterations while the training loss decreased very slowly each iteration.
- Since the loss function was reduced after every iteration, the results agree with theory.
- To conclude, tumor data are linear separable and hence able to be classified with high accuracy as benign or malignant via a hyperplane in p-dimensional space. Additionally, the classifications can be performed with few feature variables and the algorithm is able to converge quickly.


Training loss and training accuracy over 50 iterations
Value
Iteration
Training loss
Training accuracy