CS 574 Machine Learning
Homework 2 - Linear Classifiers

Parnal Joshi
Bioinformatics and Computational Biology PhD Program
Veterinary Microbiology and Preventive Medicine
Iowa State University
Ames, IA, USA

March 3, 2021

Unless otherwise stated,

- all 1-D vectors, such as x and w below, are column vectors by default.
- the classification problem is binary.
- a lower case letter variable is a scalar or vector, where a upper case letter (in any font) is a matrix.

1. Given a sample with feature vector $x = [1.1, 2.2, 3.3]^T$, what is its augmented feature vector?

   The augmented feature vector is: $x = [1.1, 2.2, 3.3, 1]^T$

2. If the weight vector of a linear classifier is $w = [1, 0, 1, 0]^T$, and we define that a sample belongs to class $+1$ if $w^T x > 0$ and -1 if $w^T x < 0$ where x is the augmented feature vector of the sample, what is the class of the sample?

   $w^T = [1, 0, 1, 0]^T$

   $x = [1.1, 2.2, 3.3, 1]$ (x is augmented)

   $w^T x = 1.1 + 3.3 = 4.4$

   Since $w^T x$ is > 0, the sample of feature value $x = [1.1, 2.2, 3.3]^T$ belongs to class $+1$

3. When discussing the sum of error squares loss function in the class, we used augmented but not normalized augmented (normalized and augmented) feature vectors. Please rewrite that loss function $J(\mathbf{W}) = \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{W} - y_i)^2$ in terms of **normalized augmented** feature vectors. Let x_i'' be the normalized augmented feature vector of the i-th sample, and w be the weight vector of the classifier. A correct prediction shall satisfy $w^T x_i'' > 0$ regardless of the class of the sample because x_i'' has been normalized. You may use a computational algebra system to help – but it is not required. It might be easier by hand on scratch paper.

   $$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{pmatrix}^T \text{ and } \mathbf{W} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ -w_1 w_2 \dots w_n \end{pmatrix}$$

   Each x_i in X is a feature vector

   $\mathbf{x}_i' = (x_{i1}, x_{i2}, \dots, x_{in})^T$

   $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in}, 1)^T$ which is the augmented feature vector

$\mathbf{x}''_i = \mathbf{x}'_i \mathbf{y}'_i$ for $y_i \in \{+1, -1\}$ which is the normalized augmented feature vector

$$J(\mathbf{W}) = \sum_{i=1}^{N} (\mathbf{W}^T \mathbf{x}_i \mathbf{y}_i - y_i)^2 = \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{y}_i \mathbf{W} - y_i)^2 = \sum_{i=1}^{N} (\mathbf{x}''^T_i \mathbf{W} - y_i)^2$$

where $\mathbf{x}''_i$ is the i-th normalized sample (we have N samples here), y_i the corresponding label, $\mathbf{w}$ is the weight vector, $\mathbf{w}^T \mathbf{x}''_i > 0$ is true for correct prediction.

4. Please find the solution for minimizing the new loss function. Keep variables and font style consistent with those in the class notes/slides, except that you can reuse the matrix $\mathbb{X} = \begin{pmatrix} — & \mathbf{x''}^T_1 & — \\ — & \mathbf{x''}^T_2 & — \\ & \vdots & \\ — & \mathbf{x''}^T_N & — \end{pmatrix}$, each **row** of which is re-purposed into a normalized and augmented feature vector. The right most column of the new $\mathbb{X}$ should contain only 1's and -1's.

   Minimizing $J(\mathbf{W})$ means:

   $\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}} = 2 \sum_{i=1}^{N} \mathbf{x}''_i (\mathbf{x}''^T_i \mathbf{W} - y_i) = (0, \ldots, 0)^T$

   Hence, $\sum_{i=1}^{N} \mathbf{x}''_i \mathbf{x}''^T_i \mathbf{W} = \sum_{i=1}^{N} \mathbf{x}''_i y_i$

   The sum of a column vector multiplied with a row vector produces a matrix.

$$\sum_{i=1}^{N} \mathbf{x}''_i \mathbf{x}''^T_i = \begin{pmatrix} | & | & & | \\ \mathbf{x}''_1 & \mathbf{x}''_2 & \cdots & \mathbf{x}''_N \\ | & | & & | \end{pmatrix} \begin{pmatrix} — & \mathbf{x}''^T_1 & — \\ — & \mathbf{x}''^T_2 & — \\ & \vdots & \\ — & \mathbf{x}''^T_N & — \end{pmatrix} = \mathbb{X}^T \mathbb{X}$$

$$\sum_{i=1}^{N} \mathbf{x}''_i y_i = \begin{pmatrix} | & | & & | \\ \mathbf{x}''_1 & \mathbf{x}''_2 & \cdots & \mathbf{x}''_N \\ | & | & & | \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \mathbb{X}^T \mathbf{y}$$

   $\mathbb{X}^T \mathbb{X} \mathbf{W} = \mathbb{X}^T \mathbf{y}$

   $(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{X} \mathbf{W} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{y}$

   $\mathbf{W} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{y}$