Vol. 1 (2023), pp.˜1–1 ISSN XXX-XXXX-XXXX

Maximizing a Correlation to Identify Features With Linear Predictive Power

Brian Tenneson

August 19, 2023

Tenneson (no affiliation)

Draft of 08/19/23

Abstract

Suppose we are given an $m \times n$ matrix D and interpret D in the following intuitive sense: we seek to find out if the first column of D is predicted as it were from some linear combination of other columns. In fact, we will introduce bias in the remaining $n-1$ entries to the right of the first column and find out which bias vector maximizes the correlation between the entry in the first column and a set of synthetic scores using biases as weights applied to the entries to the right. The goal is to determine which bias vector maximizes the correlation between the first column of D and the ith synthetic scores for $1 \leq i \leq m$. Under the right conditions, the correlation between the first column of D and a linear combination of the remaining columns is guaranteed to attain a maximum, depending on the bias vector $\overrightarrow{b} = (b_2, \ldots, b_n)$. We will use this approach to solve simple classification problems with data taken from Kaggle.com.

keywords: maximizing correlation, classification

1 Correlation Formulation

Let D be a given $m \times n$ matrix with entries a_{ij} where $1 \leq i \leq m$ and $1 \leq j \leq n$. Out of all columns but the first column of D, we create the ith synthetic score:

$$S_i\left(\overrightarrow{b}\right) = \sum_{j=2}^{n} b_j a_{ij}. \tag{1}$$

Let $Y_i = a_{i1}$. For the sake of notation, let $S\left(\overrightarrow{b}\right)$ be the collection $\left\{S_i\left(\overrightarrow{b}\right) : 1 \leq i \leq m\right\}$ and let $Y = \{a_{i1} : 1 \leq i \leq m\}$. Then we can articulate the goal presently. We seek to maximize the following correlation ρ:

$$\rho\left(\overrightarrow{b}\right) = \mathsf{cor}\left(S\left(\overrightarrow{b}\right), Y\right). \tag{2}$$

We know that

$$\mathsf{cor}\left(A, B\right) = \frac{\mathsf{cov}\left(A, B\right)}{\sqrt{\mathsf{var}\left(A\right)\mathsf{var}\left(B\right)}}; \tag{3}$$

cov stands for covariance; var stands for variance. Thus, we seek to find out where ρ has an absolute maximum, depending on the bias introduced, $\overrightarrow{b}$.

$$\rho\left(\overrightarrow{b}\right)=\frac{\mathsf{cov}\left(S\left(\overrightarrow{b}\right),Y\right)}{\sqrt{\mathsf{var}\left(S\left(\overrightarrow{b}\right)\right)\mathsf{var}\left(Y\right)}}. \tag{4}$$

Recall that

$$\mathsf{var}\left(A\right)=\frac{1}{m}\sum_{i=1}^{m}\left(A_i-\overline{A}\right)^2 \tag{5}$$

with

$$\overline{A}=\frac{1}{m}\sum_{i=1}^{m}A_i. \tag{6}$$

The covariance is known to satisfy the following relation:

$$\mathsf{cov}\left(A,B\right)=\overline{(AB)}-\overline{A}\cdot\overline{B} \tag{7}$$

$$\mathsf{cov}\left(A,B\right)=\left(\frac{1}{m}\sum_{i=1}^{m}A_iB_i\right)-\left(\frac{1}{m}\sum_{i=1}^{m}A_i\right)\left(\frac{1}{m}\sum_{i=1}^{m}B_i\right). \tag{8}$$

Then ρ takes the following form:

$$\rho\left(\overrightarrow{b}\right)=\frac{\left(\frac{1}{m}\sum_{i=1}^{m}S_i\left(\overrightarrow{b}\right)Y_i\right)\text{-}\left(\frac{1}{m}\sum_{i=1}^{m}S_i\left(\overrightarrow{b}\right)\right)\left(\frac{1}{m}\sum_{i=1}^{m}Y_i\right)}{\sqrt{\left(\frac{1}{m}\sum_{i=1}^{m}\left(S_i\left(\overrightarrow{b}\right)-\overline{S\left(\overrightarrow{b}\right)}\right)^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2\right)}}. \tag{9}$$

Since $m>0$, this reduces to the following formula for correlation:

$$\rho\left(\overrightarrow{b}\right)=\frac{\left(\sum_{i=1}^{m}S_i\left(\overrightarrow{b}\right)Y_i\right)\text{-}\left(\frac{1}{m}\right)\left(\sum_{i=1}^{m}S_i\left(\overrightarrow{b}\right)\right)\left(\sum_{i=1}^{m}Y_i\right)}{\sqrt{\left(\sum_{i=1}^{m}\left(S_i\left(\overrightarrow{b}\right)-\overline{S\left(\overrightarrow{b}\right)}\right)^2\right)\left(\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2\right)}} \tag{10}$$

Simplifying further, we obtain

$$\rho\left(\overrightarrow{b}\right)=\frac{\left(\sum_{i=1}^{m}S_i\left(\overrightarrow{b}\right)Y_i\right)\text{-}\overline{Y}\left(\sum_{i=1}^{m}S_i\left(\overrightarrow{b}\right)\right)}{\sqrt{\left(\sum_{i=1}^{m}\left(S_i\left(\overrightarrow{b}\right)-\overline{S\left(\overrightarrow{b}\right)}\right)^2\right)\left(\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2\right)}}; \tag{11}$$

so, by utilizing equation (1) and (11), we can rewrite ρ as having nested sums in terms of the bias vector $\overrightarrow{b}=(b_2,...,b_n)$:

$$\rho\left(\overrightarrow{b}\right)=\frac{\left(\sum_{i=1}^{m}\left(\sum_{j=2}^{n}b_ja_{ij}\right)Y_i\right)\text{-}\overline{Y}\left(\sum_{i=1}^{m}\left(\sum_{j=2}^{n}b_ja_{ij}\right)\right)}{\sqrt{\left(\sum_{i=1}^{m}\left(\left(\sum_{j=2}^{n}b_ja_{ij}\right)-\overline{S\left(\overrightarrow{b}\right)}\right)^2\right)\left(\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2\right)}}. \tag{12}$$

Recall that $\overline{S\left(\overrightarrow{b}\right)} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=2}^{n} b_j a_{ij}$. Consequently, we have arrived at this formula for the correlation depending on the bias vector $(b_2, ..., b_n)$:

$$\rho\left(\overrightarrow{b}\right) = \frac{\left(\sum_{i=1}^{m}\left(\sum_{j=2}^{n} b_j a_{ij}\right) Y_i\right) - \overline{Y}\left(\sum_{i=1}^{m}\left(\sum_{j=2}^{n} b_j a_{ij}\right)\right)}{\sqrt{\left(\sum_{i=1}^{m}\left(\left(\sum_{j=2}^{n} b_j a_{ij}\right) - \left(\frac{1}{m}\sum_{i=1}^{m}\sum_{j=2}^{n} b_j a_{ij}\right)\right)^2\right)\left(\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2\right)}}. \quad (13)$$

The numerator can be simplified to lead us to the following formula for correlation:

$$\rho\left(\overrightarrow{b}\right) = \frac{\sum_{i=1}^{m}\sum_{j=2}^{n}\left(Y_i - \overline{Y}\right) b_j a_{ij}}{\sqrt{\left(\sum_{i=1}^{m}\left(\left(\sum_{j=2}^{n} b_j a_{ij}\right) - \left(\frac{1}{m}\sum_{i=1}^{m}\sum_{j=2}^{n} b_j a_{ij}\right)\right)^2\right)\left(\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2\right)}}. \quad (14)$$

1.1 Finding the ideal bias vector

We wish to find for what bias vector $\overrightarrow{b}$, having $n-1$ variables b_j, for which ρ maximized. We will discuss the approach for finding the bias vector that maximizes correlation by finding out where ρ has an absolute maximum on $[0,1]^{n-1}$ with the constraint that $\sum_{j=2}^{n} b_j = 1$. Once we find a bias vector that maximizes ρ, we can find out the maximum correlation between the first column of D and the column of synthetic scores S which are based on the $n-1$ remaining columns of D. We have a choice of maximizing ρ directly using formula (10) or maximizing ρ using formula (13). We can also perform an analysis when the biases are allowed to have different (or no) constraints. Hence, $[-1,1]^{n-1}$ could be a viable domain.

1.2 Classification

Suppose column one of D is categorical, such as when $Y \subseteq \{0,1\}$. Once we have found an ideal bias vector $\overrightarrow{B}$, in order to make a "prediction," next calculate the ith synthetic score given by equation (1):

$$S_i\left(\overrightarrow{B}\right) = \sum_{j=2}^{n} B_j a_{ij}. \quad (15)$$

Next, find a linear fit for these data points:

$$\left\{\left(S_i\left(\overrightarrow{B}\right), a_{i1}\right) : 1 \leq i \leq m\right\}. \quad (16)$$

Let $f(s)$ be the best linear fit function f of variable s for those data points.

In order to make a classification, define c_δ to be a function defined as follows for $\delta > 0$:

$$c_\delta(s) := \begin{cases} 1 & \text{if} \quad f(s) > \delta \\ 0 & \text{if} \quad f(s) \leq \delta \end{cases}. \quad (17)$$

To predict which category a new sample might be in, note that if a new sample has $n-1$ measurements given by

$$C := \{C_j : 2 \leq j \leq n\}, \quad (18)$$

then our initial solution to the classification problem is as follows: the category that sample the algorithm says is in will be given by

$$c_{1/2}\left(\sum_{j=2}^{n} B_j C_j\right). \tag{19}$$

If we want to test the suitability of $\delta = 1/2$, we may wish to revise that and see how often $c_\delta\left(\sum_{j=2}^{n} B_j a_{ij}\right) = a_{i1}$. Let

$$E_\delta = \left\{ i \in \mathbb{N} : 1 \leq i \leq m \wedge c_\delta\left(\sum_{j=2}^{n} B_j a_{ij}\right) = a_{i1} \right\}. \tag{20}$$

Note that $0 \leq |E_\delta| \leq m$. Larger $|E_\delta|$ means the model is more accurate on its training data.

1.3 Instances

We will explore datasets from Kaggle.com.

https://www.kaggle.com/datasets/erdemtaha/cancer-data
https://www.kaggle.com/datasets/ninjacoding/breast-cancer-wisconsin-benign-or-malignant
https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification