EFFICIENT CALCULATION OF POLYNOMIAL FEATURES ON SPARSE MATRICES

Andrew Nystrom
Savvysherpa Inc.
6200 Shingle Creek Pkwy
Suite 400
Minneapolis, MN 55430, USA
awnystrom@gmail.com

John Hughes
Department of Computational Neuroscience
University of the Witwatersrand
Joburg, South Africa
{robot,net}@wits.ac.za

ABSTRACT

We provide an algorithm for polynomial feature expansion that can operate directly on sparse a matrix of size NxD with $O(d^k D^k N)$ where k is the polynomial order and d is the density of each row.

1 INTRODUCTION

Polynomial feature expansion is a tool long used in statistics for approximating nonlinear functions (see Gergonne (1974); Smith (1918)). While their use has long been widespread, the authors are unaware of any improvements made to their calculation efficiency. In this work we provide an algorithm for calculating polynomial features for a matrix of size matrix of size NxD with $O(d^k D^k N)$ where k is the polynomial order and d is the density of each row. The density of a matrix is the percent of its nonzero items, so $0 \leq d \leq 1$. The standard algorithm is $O(D^k N)$, so the added factor of d^k represents a significant reduction in time. The algorithm does not require the densification of the matrix, e.i. the matrix remains in sparse form, so the space complexity is also $O(d^k D^k N)$ as opposed to $O(D^k N)$.

2 ALGORITHM

The standard method of calculating polynomial features for a vector $\vec{x}$ involves augmenting a feature for the product of each combination of features in $\vec{x}$ of orders 2 to k. This method does not exploit the sparsity of a sparse matrix and will yield a product of zero any time one of the features involved in the product is zero. In a sparse matrix, such products will be common, and since the default value of a sparse matrix is zero, these products are entirely unnecessary to compute.

The main idea behind our algorithm is to only compute products that do not involve zeros. In a compressed sparse row matrix, the columns containing nonzero data are the only columns that are stored. We can therefore iterate over products of combinations of orders 2 to k of these columns for each row to calculate k-degree polynomial features.

While the idea is straightforward, there is yet one unaddressed caveat; Given a set of nonzero columns whose product was just calculated to generate a polynomial feature, where in the polynomial matrix does the result of the product belong? To adress this, we give a bijective mapping from the set of possible column index combinations of order 2 to k onto the column index space of the polynomial feature matrix. More precisely, this the mapping is of the form

$$(x_{i_0}, x_{i_1}, \ldots, x_{i_{k-1}}) \mapsto p_{i_0 i_1 \ldots i_{k-1}} \in \{0, 1, \ldots, \binom{D}{k}\} \forall (i_0, i_1, \ldots, i_{k-1}) \mid 0 \leq i_0 \leq i_1 \leq \cdots \leq i_{k-1} < D \tag{1}$$

2.1 Construction of Mappings

We seek a map from matrix indices (i, j) (with $i < j$ and $0 \leq i < D$) to numbers $f(i, j)$ with $0 \leq f(i, j) < \frac{D(D-1)}{2}$, one that follows the pattern indicated by

$$\begin{bmatrix} x & 0 & 1 & 3 \\ x & x & 2 & 4 \\ x & x & x & 5 \\ x & x & x & x \end{bmatrix} \tag{2}$$

where the entry in row i, column j, displays the value $f(i, j)$.

To simplify slightly, we introduce a notation for the nth triangular number,

$$T_2(n) = \frac{n(n+1)}{2} \tag{3}$$

The subscript 2 indicates that these are triangles in two dimensions; we'll use $T_3(n)$ to indicate the nth tetrahedral number, and so on for higher dimensions.

Observe that in Equation 2, each entry in column j (for $j > 0$) lies in the range

$$T_2(j-1) \leq e < T_2(j). \tag{4}$$

And in fact, the entry in the ith row of that column is just $i + T_2(j-1)$. Thus we have

$$\begin{aligned} f(i, j) &= i + T_2(j-1) && (5) \\ &= i + \frac{(j-1)j}{2} && (6) \\ &= \frac{2i + j^2 - j}{2}. && (7) \end{aligned}$$

For instance, in column $j = 2$ in our example (the *third* column), the entries range from 1 to 2, while $T_2(j-1) = T_2(1) = 1$ and $T_2(j) = T_2(2) = 3$, and the entry in column $j = 2$, row $i = 1$ is $i + T_2(j-1) = 1 + 1 = 2$.

2.1.1 Other indices

With one-based indexing in both the domain and codomain, the formula above becomes

$$\begin{aligned} f_1(i, j) &= 1 + f(i-1, j-1) && (8) \\ &= 1 + f(i-1, j-1) && (9) \\ &= \frac{2 + 2(i-1) + (j-1)^2 - (j-1)}{2} && (10) \\ &= \frac{2i + j^2 - 3j + 2}{2} && (11) \end{aligned}$$

2.1.2 Polynomial Features

For polynommial features, we seek a map from matrix indices (i, j) (with $i \leq j$ and $0 \leq i < D$) to numbers $g(i, j)$ with $0 \leq f(i, j) < \frac{D(D+1)}{2}$, one that follows the pattern indicated by

$$\begin{bmatrix} 0 & 1 & 3 & 6 \\ x & 2 & 4 & 7 \\ x & x & 5 & 8 \\ x & x & x & 9 \end{bmatrix} \tag{12}$$

i.e., essentially the same task as before, except that the diagonal is included. One can regard all but the last column of entries in Equation 12 as corresponding to the entries in Equation 2, but shifted to the left. Thus the formula for $g(i, j)$ is simply the formula for f, shifted by 1, i.e.,

$$\begin{aligned} g(i, j) &= f(i, j+1) && (13) \\ &= \frac{2i + (j+1)^2 - (j+1)}{2} && (14) \\ &= \frac{2i + j^2 + j + 1)}{2}. && (15) \end{aligned}$$

Alternatively, we can write this as

$$g(i, j) = i + T_2(j), \tag{16}$$

and get the same result.

2.1.3 Higher dimensions

To handle three-way interactions, we need to map triples of indices in a 3-index array to a flat list, and similarly for higher-order interactions.

For three indices, i, j, k, with $i < j < k$ and $0 \leq i, j, k < D$, we have a similar recurrence. Calling the mapping h, we have

$$h(i, j, k) = i + T_2(j - 1) + T_3(k - 2); \tag{17}$$

if we define $T_1(i) = i$, then this has the very regular form

$$h(i, j, k) = T_1(i) + T_2(j - 1) + T_3(k - 2); \tag{18}$$

and from this, the generalization to higher dimensions is straightforward. The formulas for "higher triangular numbers", i.e., those defined by

$$T_k(n) = \sum_{i=1}^{n} T_{k-1}(n) \tag{19}$$

for $k > 1$ can be determined inductively. For $k = 3$, the result is

$$T_3(n) = \sum_{i=1}^{n} T_2(n) \tag{20}$$

$$= \frac{n^3 + 3n^2 + 2n}{6}, \tag{21}$$

so that the formula for 3-way interactions, with zero-based indexing, becomes

$$h(i, j, k) = 1 + (i - 1) + \frac{(j - 1)j}{2} + \tag{22}$$

$$\frac{(k - 2)^3 + 3(k - 2)^2 + 2(k - 2)}{6}. \tag{23}$$

2.1.4 Higher-dimension polynomial features

For the case where we include the diagonal in higher dimensions, we must shift j by 1, k by 2, and so on, and the formula becomes

$$\ell(i, j, k) = T_1(i) + T_2(j) + T_3(k), \tag{24}$$

with analogous formulas for higher degree polynomial interactions.

3 Analysis

3.1 Analytical

3.2 Empirical

4 Conclusion

5 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

5.1 CITATIONS WITHIN THE TEXT

Citations within the text should be based on the `natbib` package and include the authors' last names and year (with the "et al." construct for more than two authors). When the authors or the publication are included in the sentence, the citation should not be in parenthesis (as in "See **?** for more information."). Otherwise, the citation should be in parenthesis (as in "Deep learning shows promise to make progress towards AI (**?**).").

The corresponding references are to be listed in alphabetical order of authors, in the REFERENCES section. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

5.2 FOOTNOTES

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

5.3 FIGURES

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

Figure 1: Sample figure caption.

5.4 TABLES

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

[1]Sample of the first footnote
[2]Sample of the second footnote

Table 1: Sample table title

PART	**DESCRIPTION**
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the References section; see below). Please note that pages should be numbered.

7 Preparing PostScript or PDF files

Please prepare PostScript or PDF files with paper size "US Letter", and not, for example, "A4". The -t letter option on dvips will produce US Letter files.

Consider directly generating PDF files using `pdflatex` (especially if you are a MiKTeX user). PDF figures must be substituted for EPS figures, however.

Otherwise, please generate your PostScript and PDF files with the following commands:

```
dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
ps2pdf mypaper.ps mypaper.pdf
```

7.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the graphicx package. Always specify the figure width as a multiple of the line width as in the example below using .eps graphics

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.eps}
```

or

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

for .pdf graphics. See section 4.4 in the graphics bundle documentation (`http://www.ctan.org/tex-archive/macros/latex/required/graphics/grfguide.ps`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

JD Gergonne. The application of the method of least squares to the interpolation of sequences. *Historia Mathematica*, 1(4):439–447, 1974.

Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.