Natural Language Processing: Self-Supervised Models

Mathematics Background Review

27 January 2023
Adam Byerly (Course TA)
Matrix Operations

Transpose

\[ [A^T]_{ij} = [A]_{ji} \]

\[ A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad A^T = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \]

Addition

\[ A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \]

\[ A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix} \]
Matrix Operations

Scalar Multiplication

\[
A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \alpha = a_\ast
\]

\[
\alpha A = \begin{bmatrix} a_\ast \cdot a_{11} & a_\ast \cdot a_{12} \\ a_\ast \cdot a_{21} & a_\ast \cdot a_{22} \end{bmatrix}
\]

Vector Multiplication

\[
A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}
\]

\[
Ab = \begin{pmatrix} a_{11} b_1 + a_{12} b_2 \\ a_{21} b_1 + a_{22} b_2 \end{pmatrix}
\]
Probability

We denote the probability of an event A occurring as $P(A)$

- $P(A \cap B)$ – probability of A and B both occurring
- $P(A \cup B)$ – probability of A or B both occurring

\[
P(A \cup B) = P(A) + P(B) - P(A \cap B)
\]

If $P(A \cap B) = 0$, we say that A and B are *mutually-exclusive*

If $P(A \cap B) = P(A)P(B)$, we say that A and B are *independent*
Calculus – Simple Derivatives

Constant Rule

\[ \frac{d}{dx} c = 0 \]

Powers

\[ \frac{d}{dx} x^a = ax^{a-1} \]

\[ \frac{d}{dx} x = 1 \quad \frac{d}{dx} x^2 = 2x \]
Calculus – Simple Derivatives

Exponents & Logarithms

\[
\frac{d}{dx} a^x = a^x \ln a \quad \rightarrow \quad \frac{d}{dx} e^x = e^x
\]

Trigonometric Functions

\[
\frac{d}{dx} \sin x = \cos x \quad \frac{d}{dx} \cos x = -\sin x \quad \frac{d}{dx} \tan x = \frac{1}{\cos^2 x}
\]
Calculus – Combined Functions

Addition
\[ \frac{d}{dx}(\alpha f + \beta g) = \alpha \frac{df}{dx} + \beta \frac{dg}{dx} \]

Product
\[ \frac{d}{dx}(fg) = \frac{df}{dx}g + f \frac{dg}{dx} \]

Quotient
\[ \frac{d}{dx}\left(\frac{f}{g}\right) = \frac{\frac{df}{dx}g - f \frac{dg}{dx}}{g^2} \]
Calculus – Chain Rule

Chain Rule

\[
\frac{d}{dx} (f(g(x))) = \frac{df}{dx} (g(x)) \frac{dg}{dx} (x)
\]
Calculus – Multivariate Derivative

Multivariate Functions

\[ f: \mathbb{R}^n \rightarrow \mathbb{R} \]

\[ f(x, y) = x^2 + y^2 \]

Gradient

\[ \nabla f(x) = \left[ \frac{\partial f}{\partial x_1} \ldots \frac{\partial f}{\partial x_n} \right] \]

\[ \nabla f(x, y) = [2x, 2y] \]
Find the largest number in an *unsorted* list of n numbers:
• No additional information
• Need to traverse entire list
• Algorithm scales with list size n
• $O(n)$

Find the largest number in an *sorted* list of n numbers:
• Additional information – list is sorted
• Only need the last element
• $O(1)$