Math behind GloVe

Definitions

- Matrix der **word-word co-occurrence counts**: X (Wie X genau gebildet wird kommt später)
- X_{ij} ist Anzahl wie oft Wort j im Kontext von Word i auftritt.
- $X_i = \sum_k X_{ik}$: Anzahl, wie oft alle Wörter im KOntext von i auftreten.
- $P_{ij} = P(j|i) = X_{ij}/X_i$: Wahrscheinlichkeit, dass Wort j im KOntext von Wort i auftritt.

Man will, dass das Ratio P_{ik}/P_{jk} den Zusammenhang von Wörtern beschreiben können. Dazu nimmt man ein beliebiges Wort k. Hat nun das Wort k einen Bezug zu Wort i, aber nicht zu Wort j, dann soll das Ratio $P_{ik}/P_{jk} > 1$ groß sein. Hat k einen Bezug zu j und nicht zu i wird das Ratio $P_{ik}/P_{jk} < 1$ klein. Hat k zu beiden Wörtern einen bezug erwartet man für das Ratio $P_{ik}/P_{jk} \approx 1$.

Das Modell

Man startet mit dem Ratio, da das Ratio besser in der Lage ist Wörter in einen Kontext zu bringen als die einfachen Wahrscheinlichkeiten P_{ij}. Dazu sind allerdings immer drei Wörter i, j und k nötig.

Das einfachste Modell hat die Form

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}.$$

Dabei ist $w_i \in \mathbb{R}^d$ ein Wortvektor zum Wort i und $\tilde{w}_k \in \mathbb{R}^d$ Wortvektoren welche sich auf einen seperaten Kontext beziehen (dazu später mehr). Die Funktion F kann dabei von noch nicht spezifiezierten Paremtern abhängen. (Die Anzahl an möglichen Werten, die F annehmen kann ist enorm.)

Als nächstes nutzt man aus, dass Wortvektoren in einem linearen Vektorraum leben, kann anstelle von zwei Vektoren w_i und w_j die Differenz $w_i - w_j$ betrachtet werden. Das hat zur folge, dass aus drei Argumenten zwei werden:

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}.$$

Weiter ist F eine Abbildung von $\mathbb{R}^d$ nach $\mathbb{R}$:

$$F : \mathbb{R}^d \to \mathbb{R}$$

Man könnte F z. B. mit einem neuronalen Netz fitten. Dadurch würden wir aber jegliche Information über die lineare Struktur der Wort-Vektoren verwischen.

Eine Möglichkeit hier ist es, dass Skalarprodukt im Argument zu verwenden:

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}.$$

Daduch werden die Vektor Dimensionen nicht willkürlich durchgemischt.

Im nächsten Schritt wollen wir, dass es egal ist ob wir w als Wort-Vektor oder Kontext-Wort-Vektor verwenden (exchange symmetrie). Wir können also $w \leftrightarrow \tilde{w}$ austauschen. Dementsprechend muss $X \leftrightarrow X^T$ ausgetauscht werden.

Die letzte Gleichung erfüllt das allerdings nicht. Um diese Eigenschaft dennoch zu erhalten werden zwei Schritte benötigt:

1. Schritt:

F wird vorausgesetzt ein Homomophismus zwischen $(\mathbb{R}, f) = (\mathbb{R}, +)$ und $(\mathbb{R}_+, g) = (\mathbb{R}_+, \cdot)$ zu sein. D. h.:

$$F(f(a,b)) = F(a+b) = F(a)F(b) = g(F(a), F(b))$$

Dabei ist $a = w_i^T \tilde{w}_k$ und $b = -w_j^T \tilde{w}_k$. Es ist allgemein bekannt, dass $F(a+b) = F(a)F(b)$ dies eine Funktionalgleichung ist, welche durch $F = \exp_a$ gelöst wird.