FM $O(dk)$ weight gradient calculation

Andrew Bai

July 2022

Let $X \in \mathbb{R}^d$ and $P \in \mathbb{R}^{k \times d}$. We will ignore the linear terms for now. A factorization machine ϕ is defined as follows

$$\phi(X) := \frac{1}{2}\Big(\|PX\|^2 - \sum_{i=1}^{k} \|P_{i,:} \circ X\|^2\Big) \tag{1}$$

$$= \frac{1}{2}\Big(\sum_{i=1}^{k}\big(\sum_{j=1}^{d} P_{i,j} \cdot X_j\big)^2 - \sum_{j=1}^{d} {X_j}^2 \cdot \big(\sum_{i=1}^{k} {P_{(i,j)}}^2\big)\Big) \tag{2}$$

$$= \frac{1}{2}\Big(\sum_{i=1}^{k}\sum_{j=1}^{d} P_{i,j}X_j \sum_{j'=1}^{d} P_{i,j'}X_{j'} - \sum_{j=1}^{d} {X_j}^2 \cdot \big(\sum_{i=1}^{k} {P_{(i,j)}}^2\big)\Big) \tag{3}$$

Given a binary classification setting where $y \in \{+1, -1\}$ and the model is trained with logistic regression, the loss function is as follows

$$l(X, y) = \log(1 + \exp(-y \cdot \phi(X)))$$

We now derive the derivative of the loss function with respect to one single weight parameter $P_{i,j}$.

$$\frac{\mathrm{d}l(X,y)}{\mathrm{d}P_{i,j}} = \frac{\mathrm{d}l(X,y)}{\mathrm{d}\phi(X)} \cdot \frac{\mathrm{d}\phi(X)}{\mathrm{d}P_{i,j}} = \frac{-y}{1 + \exp(y \cdot \phi(X))} \cdot \frac{\mathrm{d}\phi(X)}{\mathrm{d}P_{i,j}} \tag{4}$$

According to Eq (3) the only terms in $\phi(X)$ involving $P_{i,j}$ is

$$\phi_{i,j}(X) = \frac{1}{2}\big(2 \cdot P_{i,j}X_j \cdot \sum_{j'=1}^{d} P_{i,j'}X_{j'} - {P_{i,j}}^2{X_j}^2 - {X_j}^2 \cdot {P_{i,j}}^2\big) \tag{5}$$

Let us pre-compute the embedding $Z = PX$ and memoize the results. We then proceed to compute the second term of Eq (4)

$$\frac{\mathrm{d}\phi(X)}{\mathrm{d}P_{i,j}} = \frac{\mathrm{d}\phi_{i,j}(X)}{\mathrm{d}P_{i,j}} = X_j \cdot \sum_{j'=1}^{d} P_{i,j'}X_{j'} - 2P_{i,j}{X_j}^2 \tag{6}$$

$$= X_j \cdot Z_i - 2P_{i,j}{X_j}^2 \tag{7}$$

Memoizing Z takes $O(dk)$ time. Eq (7) can be calculated in $O(1)$ by looking up the memoized Z_i. Thus, the total complexity of calculating the weight gradient of FM is $O(dk)$.