Factorization Machines

山下 滉

2018/05/25

1 Overview

Factorization Machines（FM）は 2010 年に Steffen Rendle によって発表された [1].
回帰，2 値分類，ランキングに利用できる SVM のような一般的な予測モデル．
SVM では無理だったスパースな入力データに対応できる．
Kaggle の CTR 予測コンペの優勝チームが使用してて最近アツい．
Steffen Rendle が kaggle の e-Learning での学習者の正誤予測コンペで優勝してる．
libfm

2 Description of Data

データ $D = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \cdots\}$ があるものとする．

FM では，入力ベクトル $\boldsymbol{x}$ は非常にスパースであるものとする．ここで，$m(\boldsymbol{x})$ をベクトル $\boldsymbol{x}$ の非ゼロ成分の数，$\overline{m}_D$ を全てのベクトル $\boldsymbol{x} \in D$ の非ゼロ成分の数 $m(\boldsymbol{x})$ の平均とする．

実世界には，$\overline{m}_D \ll n$ となるような非常にスパースなデータがよくある．例えば，購買履歴データであったり，テキスト処理における Bag of Words データなどである．また，e-Learning のログデータもスパースなデータのひとつである．このようにデータがスパースになってしまうひとつの理由として，カテゴリ変数[1]を多く用いることが挙げられる．カテゴリ変数は One-Hot Encoding をして使用することが多いため，スパースになりやすい．

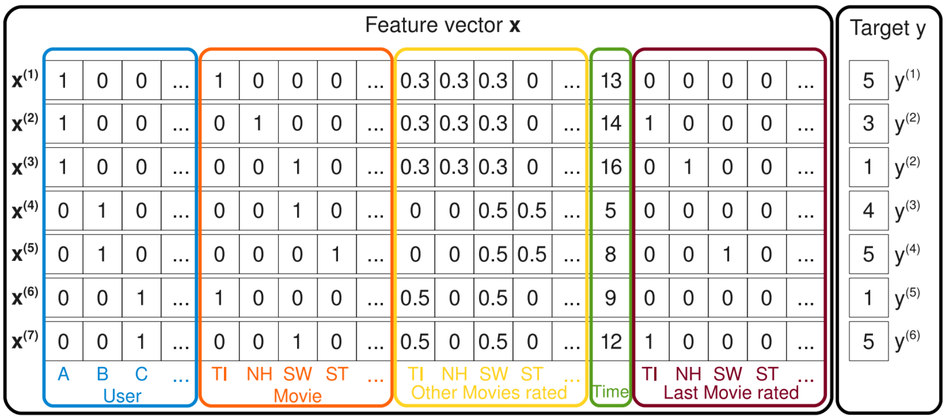

図 1: データ例

[1] ユーザー，アイテム，性別，単語など有限のラベルに分けられるもの．カテゴリ変数以外に，離散変数，連続変数がある．

3 Factorization Machines

3.1 Model Equation

式（3.1）に，FM の次元 $d=2$ のときのモデルの式を示す．

$$\hat{y}(\boldsymbol{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_i x_j \tag{3.1}$$

また，推定すべき FM のパラメーターは以下である．

$$w_0 \in \mathbb{R}, \boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{V} \in \mathbb{R}^{n \times k} \tag{3.2}$$

ここで，$\langle \cdot, \cdot \rangle$ は k 次元の 2 つのベクトルの内積を表す．

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f} \tag{3.3}$$

$\boldsymbol{v}_i$ は V の i 行目の k 次元のベクトルである．$k \in \mathbb{N}^+$ は分解の次元を定義するハイパーパラメーターである．

次元 $d=2$ の FM は 2 つの変数間の関係を捉えることができ，w_0 は全体のバイアス，w_i は i 番目の変数の強さ，$\hat{w}_{i,j} := \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ は変数間の関係を表す．$w_{i,j} \in \mathbb{R}$ を使わず各変数に設定された重みの内積で表現することにより，学習データが少ないスパースなデータでも効率よく学習することができる．

3.2 計算量

式（3.1）の計算量は，$\mathcal{O}(kn^2)$ であるが，式（3.4）による式変換を行うことにより，$\mathcal{O}(kn)$ に削減することができる．

$$\begin{aligned}
& \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_i x_j \\
= \quad & \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_i x_j - \frac{1}{2} \langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle x_i x_i \\
= \quad & \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{f=1}^{k} v_{i,f} v_{j,f} x_i x_j - \sum_{i=1}^{n} \sum_{f=1}^{k} v_{i,f} v_{i,f} x_i x_i \right) \\
= \quad & \frac{1}{2} \sum_{f=1}^{k} \left(\left(\sum_{i=1}^{n} v_{i,f} x_i \right) \left(\sum_{j=1}^{n} v_{j,f} x_j \right) - \sum_{i=1}^{n} v_{i,f}^2 x_i^2 \right) \\
= \quad & \frac{1}{2} \sum_{f=1}^{k} \left(\left(\sum_{i=1}^{n} v_{i,f} x_i \right)^2 - \sum_{i=1}^{n} v_{i,f}^2 x_i^2 \right)
\end{aligned} \tag{3.4}$$

また，入力ベクトル $\boldsymbol{x}$ がスパースであるとすると，非ゼロ成分の数だけ計算をすればいいので，計算量は $\mathcal{O}(k\overline{m}_D)$ まで削減できる．

3.3 d-way FM

d 次元の FM の式を式（3.5）に示す．

$$\hat{y}(\boldsymbol{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{l=2}^{d} \sum_{i_1=1}^{n} \cdots \sum_{i_l=i_{l-1}+1}^{n} \left(\prod_{j=1}^{l} x_{i_j} \right) \left(\sum_{f=1}^{k_l} \prod_{j=1}^{l} v_{i_j,f}^{(l)} \right) \tag{3.5}$$

また，パラメータは以下である．

$$w_0 \in \mathbb{R}, \boldsymbol{w} \in \mathbb{R}^n, \forall l \in \{2, \cdots, d\} : \boldsymbol{V}^{(l)} \in \mathbb{R}^{n \times k_l}, k_l \in \mathbb{N}^+ \tag{3.6}$$

d 次元の FM も，単純に計算すれば $\mathcal{O}(k_d n^d)$ だが，式（3.4）と同様の式変形を行うことにより，線形時間で計算することが可能である．

3.4 パラメータの学習

FM のパラメータは，SGD（Stochastic Gradient Descent）をはじめとした勾配法，ALS（Alternating least-squares）[2]，MCMC（Markov Chain Monte Carlo）による推定の 3 種類が Rendle によるライブラリに実装されている [2]．ここでは，正則化項なしの単純な SGD によるパラメータ推定を紹介する．

まず，問題設定として以下の最適化問題を定義する．

$$\underset{\Theta}{\operatorname{argmin}} \sum_{(\boldsymbol{x},y)\in D} \ell(\hat{y}(\boldsymbol{x} \mid \Theta), y) \tag{3.7}$$

Θ はモデルのパラメータをまとめたものであり，FM の次元 $d = 2$ のとき $\Theta = \{w_0, \boldsymbol{w}, V\}$ である．また，$\hat{y}(\boldsymbol{x} \mid \Theta)$ はパラメータ Θ のもとで入力 $\boldsymbol{x}$ のときの予測値を表す．

この最適化問題では，誤差関数 ℓ を選択することができる．回帰を行う際には式（3.8）に示す最小二乗誤差関数（Least Square Loss）を，分類を行う際には式（3.9）に示すロジスティック誤差関数（Logistic Loss）または hinge loss を用いる．

$$\begin{aligned}
\ell_{\mathrm{LS}}(t, \hat{y}) &= (\hat{y} - t)^2, \qquad t \in \mathbb{R} && (3.8)\\
\ell_{\mathrm{C}}(t, \hat{y}) &= -\ln(\sigma(t \cdot \hat{y})) \\
&= -\ln\left(\frac{1}{1 + \exp(-t\hat{y})}\right) \\
&= \ln(1 + \exp(-t\hat{y})), \qquad t \in \{-1, 1\} && (3.9)
\end{aligned}$$

ここで，t は教師データ，$\hat{y} \in \mathbb{R}$ はモデルの予測値を表す．また，$\sigma(x) = \dfrac{1}{1 + \exp(-x)}$ はシグモイド関数である．

誤差関数を直接最適化するために，それぞれの導関数を求める．

$$\frac{\partial}{\partial\theta}\ell_{\mathrm{LS}}(\hat{y}(\boldsymbol{x} \mid \Theta), y) = \frac{\partial}{\partial\theta}(\hat{y}(\boldsymbol{x} \mid \Theta) - y)^2 = 2(\hat{y}(\boldsymbol{x} \mid \Theta) - y)\frac{\partial}{\partial\theta}\hat{y}(\boldsymbol{x} \mid \Theta) \tag{3.10}$$

[2] 交互最小二乗法．NMF のパラメータ求めるときみたいにあるパラメータを更新するときはそれ以外のパラメータを固定してってやるやつ

$$\begin{aligned}\frac{\partial}{\partial\theta}\ell_{\mathrm{C}}(\hat{y}(\boldsymbol{x}\mid\Theta),y) &= \frac{\partial}{\partial\theta}-\ln(\sigma(\hat{y}(\boldsymbol{x}\mid\Theta)\cdot y)) \\ &= -\frac{1}{\sigma(\hat{y}(\boldsymbol{x}\mid\Theta)\cdot y)}\sigma(\hat{y}(\boldsymbol{x}\mid\Theta)\cdot y)(1-\sigma(\hat{y}(\boldsymbol{x}\mid\Theta)\cdot y)\frac{\partial}{\partial\theta}\hat{y}(\boldsymbol{x}\mid\Theta) \\ &= (\sigma(\hat{y}(\boldsymbol{x}\mid\Theta)\cdot y)-1)\frac{\partial}{\partial\theta}\hat{y}(\boldsymbol{x}\mid\Theta) \qquad (3.11)\end{aligned}$$

また，FM の各パラメータにおける勾配を式（3.12）〜式（3.14）に示す．

$$\frac{\partial}{\partial w_0}\hat{y}(\boldsymbol{x}) = 1 \qquad (3.12)$$

$$\frac{\partial}{\partial w_i}\hat{y}(\boldsymbol{x}) = x_i \qquad (3.13)$$

$$\begin{aligned}\frac{\partial}{\partial v_{i,f}}\hat{y}(\boldsymbol{x}) &= \frac{\partial}{\partial v_{i,f}}\frac{1}{2}\sum_{f'=1}^{k}\left(\left(\sum_{j=1}^{n}v_{j,f'}x_j\right)^2-\sum_{j=1}^{n}v_{j,f'}^2x_j^2\right) \\ &= \frac{\partial}{\partial v_{i,f}}\frac{1}{2}\left(\left(\sum_{j=1}^{n}v_{j,f}x_j\right)^2-v_{i,f}^2x_i^2\right) \\ &= \frac{1}{2}\left(2\left(\sum_{j=1}^{n}v_{j,f}x_j\right)\cdot x_i-2v_{i,f}x_i^2\right) \\ &= x_i\sum_{j=1}^{n}v_{j,f}x_j-v_{i,f}x_i^2 \qquad (3.14)\end{aligned}$$

式（3.14）に含まれる $\sum_{j=1}^{n}v_{j,f}x_j$ は，$\hat{y}(\boldsymbol{x})$ を計算する際に予め計算しておくことが可能であるため，全てのパラメータの勾配は $\mathcal{O}(1)$ で求めることができる．

SGD では，以下の更新式に従ってパラメータを更新していく．

$$\theta \leftarrow \theta-\eta\left(\frac{\partial}{\partial\theta}\ell(\hat{y}(\boldsymbol{x}),y)\right) \qquad (3.15)$$

ここで，データ $(\boldsymbol{x},y)\in D$ は，ランダムに選択される（または 1 周する毎にデータ D をシャッフルする）．また，$\eta\in\mathbb{R}^+$ は学習率を決めるパラメータである．

アルゴリズム 1 に，正則化項なしの SGD による FM のパラメータ推定を示す．

Algorithm 1 Stochastic Gradient Descent(SGD) of Factorization Machines(FM)

Input: Training data D, Learning rate η, Initialization σ

Output: Model parameters $\Theta = \{w_0, \boldsymbol{w}, V\}$

$w_0 \leftarrow 0; \boldsymbol{w} \leftarrow (0, \cdots, 0); V \sim \mathcal{N}(0, \sigma);$

repeat

  for $(\boldsymbol{x}, y) \in D$ **do**

    $w_0 \leftarrow w_0 - \eta \left(\frac{\partial}{\partial w_0} \ell(\hat{y}(\boldsymbol{x} \mid \Theta), y) \right)$

    for $i \in \{1, \cdots, n\} \wedge x_i \neq 0$ **do**

      $w_i \leftarrow w_i - \eta \left(\frac{\partial}{\partial w_i} \ell(\hat{y}(\boldsymbol{x} \mid \Theta), y) \right)$

      for $f \in \{1, \cdots, k\}$ **do**

        $v_{i,f} \leftarrow v_{i,f} - \eta \left(\frac{\partial}{\partial v_{i,f}} \ell(\hat{y}(\boldsymbol{x} \mid \Theta), y) \right)$

      end for

    end for

  end for

until stopping criterion is met;

参考文献

[1] Steffen Rendle. "factorization machines". In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pp. 995–1000, Washington, DC, USA, 2010. IEEE Computer Society.

[2] Steffen Rendle. Factorization machines with libfm. *ACM Trans. Intell. Syst. Technol.*, Vol. 3, No. 3, pp. 57:1–57:22, May 2012.