Latent factor model & Learning algorithm

2013-5-30

Seung-hwan baek

FM model

factorize Matrix V: use non-direct information

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j'=j+1}^{p} x_j x_{j'} \sum_{f=1}^{k} \boxed{v_{j,f}\, v_{j',f}},$$

$$w_{j,j} \approx \langle \mathbf{v}_j, \mathbf{v}_{j'} \rangle = \sum_{f=1}^{k} v_{j,f}\, v_{j',f}$$

Lemma: $W = VV^T$ *if* k *is chosen large enough*

$$w_0 \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^p, \quad \boxed{V \in \mathbb{R}^{p \times k}}$$

Learning

$$\hat{y}(\mathbf{x}) = g_{\theta}(\mathbf{x}) + \theta\, h_{\theta}(\mathbf{x}) \quad \forall \theta \in \Theta,$$

$$h_{\theta}(\mathbf{x}) = \frac{\partial \hat{y}(\mathbf{x})}{\partial \theta} = \begin{cases} 1, & \text{if } \theta \text{ is } w_0 \\ x_l, & \text{if } \theta \text{ is } w_l \\ x_l \sum_{j \neq l} v_{j,f}\, x_j, & \text{if } \theta \text{ is } v_{l,f} \end{cases}$$

Gradient of predicted value

Object-Optimization

- Minimize the gap between true value and predicted value
- Loss function
  - Regression
    - Least square loss $l^{\mathrm{LS}}(y_1, y_2) := (y_1 - y_2)^2$
  - Binary Classification
    - Sigmoid/logistic function $l^{\mathrm{C}}(y_1, y_2) := -\ln \sigma(y_1 y_2)$

$$\boxed{\mathrm{OPT}(S) := \operatorname*{argmin}_{\Theta} \sum_{(\mathbf{x},y)\in S} l(\hat{y}(\mathbf{x}|\Theta), y)}$$

$$\Downarrow$$

$$\boxed{\mathrm{OPTREG}(S, \lambda) := \operatorname*{argmin}_{\Theta} \left(\sum_{(\mathbf{x},y)\in S} l(\hat{y}(\mathbf{x}|\Theta), y) + \sum_{\theta\in\Theta} \lambda_\theta \theta^2 \right)}$$

- To avoid overfitting
  - Regularization value per group

$$\lambda^0, \quad \lambda^w_\pi, \quad \lambda^v_{f,\pi}, \quad \forall \pi \in \{1, \dots, \Pi\}, \forall f \in \{1, \dots, k\},$$

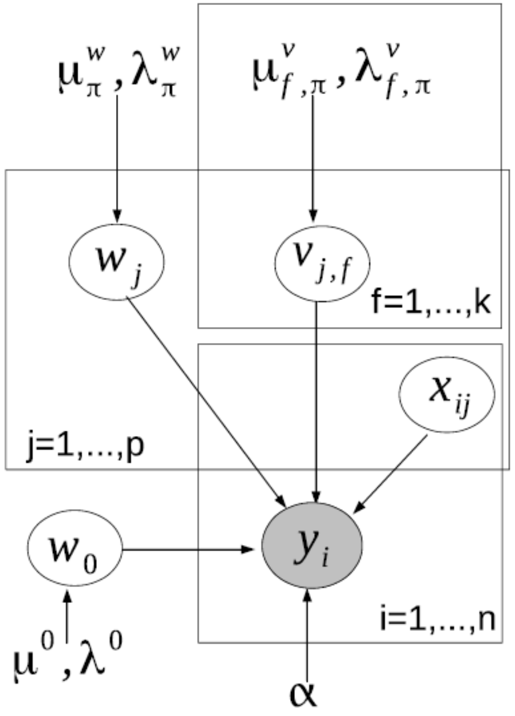
Probabilistic interpretation

- Regression

$$y|\mathbf{x}, \Theta \sim \mathcal{N}(\hat{y}(\mathbf{x}, \Theta), 1/\alpha).$$

- Binary classification
  - b: link function like logistic function

$$y|\mathbf{x}, \Theta \sim \mathrm{Bernoulli}(b(\hat{y}(\mathbf{x}, \Theta))),$$



(a) Factorization machine (FM).

Gradients

- Loss function of gradient
  - Regression

$$\frac{\partial}{\partial\theta} l^{\mathrm{LS}}(\hat{y}(\mathbf{x}|\Theta), y) = \frac{\partial}{\partial\theta}\left(\hat{y}(\mathbf{x}|\Theta) - y\right)^2 = 2\left(\hat{y}(\mathbf{x}|\Theta) - y\right) \boxed{\frac{\partial}{\partial\theta}\hat{y}(\mathbf{x}|\Theta),}$$

  - Binary classification

$$\frac{\partial}{\partial\theta} l^{\mathrm{C}}(\hat{y}(\mathbf{x}|\Theta), y) = \frac{\partial}{\partial\theta} - \ln\sigma\left(\hat{y}(\mathbf{x}|\Theta)\, y\right) = \left(\sigma\left(\hat{y}(\mathbf{x}|\Theta)\, y\right) - 1\right) y \boxed{\frac{\partial}{\partial\theta}\hat{y}(\mathbf{x}|\Theta).}$$

$$\boxed{\frac{\partial}{\partial\theta}\hat{y}(\mathbf{x}|\Theta)} = \boxed{h_\theta(\mathbf{x}).}$$

Stochastic Gradient Descent(SGD)

- Object: Find Loss function's global minimum, but we can't
- At least local minimum
- Iterates over $(\boldsymbol{x}, y)$ in S

$$\theta \leftarrow \theta - \eta \left(\frac{\partial}{\partial \theta} l(\hat{y}(\mathbf{x}), y) + 2 \lambda_\theta \theta \right),$$

ALGORITHM 1: Stochastic Gradient Descent (SGD)

Input: Training data S, regularization parameters λ, learning rate η, initialization σ
Output: Model parameters $\Theta = (w_0, \mathbf{w}, \mathbf{V})$
$w_0 \leftarrow 0$; $\mathbf{w} \leftarrow (0, \ldots, 0)$; $\mathbf{V} \sim \mathcal{N}(0, \sigma)$;
repeat
    for $(\mathbf{x}, y) \in S$ **do**
        $w_0 \leftarrow w_0 - \eta \left(\frac{\partial}{\partial w_0} l(\hat{y}(\mathbf{x}|\Theta), y) + 2 \lambda^0 w_0 \right)$;
        for $i \in \{1, \ldots, p\} \wedge x_i \neq 0$ **do**
            $w_i \leftarrow w_i - \eta \left(\frac{\partial}{\partial w_i} l(\hat{y}(\mathbf{x}|\Theta), y) + 2 \lambda^w_{\pi(i)} w_i \right)$;
            for $f \in \{1, \ldots, k\}$ **do**
                $v_{i,f} \leftarrow v_{i,f} - \eta \left(\frac{\partial}{\partial v_{i,f}} l(\hat{y}(\mathbf{x}|\Theta), y) + 2 \lambda^v_{f,\pi(i)} v_{i,f} \right)$;
            end
        end
    end
until *stopping criterion is met*;

SGD hyperparameters

Learning rate η

Regularization λ

Initialization σ

ALGORITHM 1: Stochastic Gradient Descent (SGD)

Input: Training data S, regularization parameters λ, learning rate η, initialization σ
Output: Model parameters $\Theta = (w_0, \mathbf{w}, \mathbf{V})$
$w_0 \leftarrow 0$; $\mathbf{w} \leftarrow (0, \ldots, 0)$; $\mathbf{V} \sim \mathcal{N}(0, \sigma)$;
repeat
    for $(\mathbf{x}, y) \in S$ **do**
        $w_0 \leftarrow w_0 - \eta \left(\frac{\partial}{\partial w_0} l(\hat{y}(\mathbf{x}|\Theta), y) + 2\,\lambda^0\, w_0\right)$;
        for $i \in \{1, \ldots, p\} \wedge x_i \neq 0$ **do**
            $w_i \leftarrow w_i - \eta \left(\frac{\partial}{\partial w_i} l(\hat{y}(\mathbf{x}|\Theta), y) + 2\,\lambda^w_{\pi(i)}\, w_i\right)$;
            for $f \in \{1, \ldots, k\}$ **do**
                $v_{i,f} \leftarrow v_{i,f} - \eta \left(\frac{\partial}{\partial v_{i,f}} l(\hat{y}(\mathbf{x}|\Theta), y) + 2\,\lambda^v_{f,\pi(i)}\, v_{i,f}\right)$;
            end
        end
    end
until *stopping criterion is met*;

$$\theta \leftarrow \theta - \eta \left(\frac{\partial}{\partial \theta} l(\hat{y}(\mathbf{x}), y) + 2\,\lambda_\theta\, \theta\right),$$

Alternating Least-Squares/Coordinate Descent

- SGD: Minimizing the loss per training data
- ALS: Minimizing the loss per model parameter

$$\theta^* = \operatorname*{argmin}_{\theta} \left(\sum_{(\mathbf{x},y)\in S} \left(\hat{y}(\mathbf{x}|\Theta) - y\right)^2 + \sum_{\theta\in\Theta} \lambda_\theta \theta^2 \right)$$

$$= \operatorname*{argmin}_{\theta} \left(\sum_{(\mathbf{x},y)\in S} \left(g_\theta(\mathbf{x}|\Theta \setminus \{\theta\}) + \theta\, h_\theta(\mathbf{x}|\Theta \setminus \{\theta\}) - y\right)^2 + \sum_{\theta\in\Theta} \lambda_\theta \theta^2 \right)$$

$$= \frac{\sum_{i=1}^{n} (y - g_\theta(\mathbf{x}_i|\Theta \setminus \{\theta\}))\, h_\theta(\mathbf{x}_i|\Theta \setminus \{\theta\})}{\sum_{i=1}^{n} h_\theta(\mathbf{x}_i)^2 + \lambda_\theta}$$

$$= \frac{\theta \sum_{i=1}^{n} h_\theta^2(\mathbf{x}_i) + \sum_{i=1}^{n} h_\theta(\mathbf{x}_i)\, e_i}{\sum_{i=1}^{n} h_\theta(\mathbf{x}_i)^2 + \lambda_\theta}, \qquad e_i := y_i - \hat{y}(\mathbf{x}_i|\Theta).$$

ALS hyperparametsers

Regularization λ

Initialization σ

ALGORITHM 2: Alternating least squares (ALS)

Input: Training data S, regularization parameters λ, initialization σ
Output: Model parameters $\Theta = (w_0, \mathbf{w}, \mathbf{V})$
$w_0 \leftarrow 0$; $\mathbf{w} \leftarrow (0, \ldots, 0)$; $\mathbf{V} \sim \mathcal{N}(0, \sigma)$;
repeat
    $\hat{\mathbf{y}} \leftarrow$ predict all cases S;
    $\mathbf{e} \leftarrow \mathbf{y} - \hat{\mathbf{y}}$;
    $w_0 \leftarrow w_0^*$;
    for $l \in \{1, \ldots, p\}$ **do**
        $w_l \leftarrow w_l^*$;
        update e;
    end
    for $f \in \{1, \ldots, k\}$ **do**
        init $q_{\cdot,f}$;
        for $l \in \{1, \ldots, p\}$ **do**
            $v_{l,f} \leftarrow v_{l,f}^*$;
            update e, q;
        end
    end
until *stopping criterion is met*;

$$\theta \leftarrow \theta^*.$$