Curve Fitting

The City College of New York
CSc 59929 – Introduction to Machine Learning
Fall 2017 – Erik K. Grimmelmann, Ph.D.

The Challenge

Given an unknown function whose value is known at a number of points, find the polynomial curve that "best" represents the function.

Assumptions and Notation

Let $f(x)$ be an unknown function,

$f(x) \in \mathbb{Q}$,

$x \in A \subseteq \mathbb{Q}$,

$\{x_i\}$ is the set of points at which the values of $f(x)$ are known,

$\{x_i\} = X \subset A$,

$i \in \{1, \ldots, n\} = B \subset \mathbb{N}$,

More Assumptions and Notation

Let $g_k(x)$ *be a polynomial approximation to* $f(x)$,

$$g_k(x) = \sum_{j=0}^{k} c_j x^j \in \mathbb{Q},$$

k is the degree of the polynomial,

$c_j \in \{c_0, c_1, \ldots, c_k\} = \boldsymbol{c} \subset \mathbb{Q},$

such that $c_k \neq 0$

Cost Function

To find the $g_k(x)$ that best fits $f(x)$ for a given set X of points at which the value of $f(x)$ is known, we define a cost function $J(\mathbf{c},X)_R$ and minimize $J(\mathbf{c},X)_R$ with respect to the choice of the polynomial coefficients **c**. The resulting $g_k(x)$ is denoted by $\hat{g}_k(x)$.

The subscript R denotes the regularization included in the cost function. A zero value of R denotes that no regularization term is included.

Cost Function

For the cost function without a regularization component, we choose the ½ of the mean of the squared errors,

$$J_k(\boldsymbol{c}, X) = \frac{1}{2n}\sum_{i=1}^{n}\big(f(x_i) - g(x_i)\big)^2$$

$$= \frac{1}{2n}\sum_{i=1}^{n}\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)^2$$

Because this function is **convex**, we can minimize it using the technique of gradient descent and be confident that it has a unique global minimum.

Gradient of $J_k(\boldsymbol{c}, X)$

$$J_k(\boldsymbol{c}, X) = \frac{1}{2n} \sum_{i=1}^{n} \left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)^2 ,$$

$$\nabla J_k(\boldsymbol{c}, X) = \left(\frac{\partial J_k(\boldsymbol{c}, X)}{\partial c_0}, \frac{\partial J_k(\boldsymbol{c}, X)}{\partial c_1}, \ldots, \frac{\partial J_k(\boldsymbol{c}, X)}{\partial c_k} \right),$$

where $$\frac{\partial J_k(\boldsymbol{c}, X)}{\partial c_m} = \frac{\partial \left(\frac{1}{2n} \sum_{i=1}^{n} \left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)^2 \right)}{\partial c_m}$$

Gradient of $J_k(c, X)$

$$\frac{\partial J_k(\boldsymbol{c}, X)}{\partial c_m} = \frac{\partial \frac{1}{2n}\sum_{i=1}^{n}\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)^2}{\partial c_m}$$

$$= \frac{1}{2n}\sum_{i=1}^{n} 2\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)\frac{\partial}{\partial c_m}\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)\left(-x_i^m\right)$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)\left(x_i^m\right)$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m\right) \text{ where } g_k(x_i) = \sum_{j=0}^{k} c_j x_i^j$$

$J_k(c, X)$ with No Regularization

No regularization

$$J_k(\boldsymbol{c}, X)_0 = \frac{1}{2n}\sum_{i=1}^{n}\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j \right)^2$$

$J_k(c, X)$ with L2 Regularization

L_2 regularization

$$J_k(\mathbf{c}, X)_{L2} = J_k(\mathbf{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k} c_j^2$$

$$= C \cdot J_k(\mathbf{c}, X)_0 + \frac{1}{2}\sum_{j=1}^{k} c_j^2$$

where $C = \dfrac{1}{\lambda}$

L_2 regularization with power weighting

$$J_k(\mathbf{c}, X)_{L2P} = J_k(\mathbf{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k} j^p c_j^2$$

$$= C \cdot J_k(\mathbf{c}, X)_0 + \frac{1}{2}\sum_{j=1}^{k} j^p c_j^2$$

$J_k(c, X)$ with L1 Regularization

L_1 regularization

$$J_k(\boldsymbol{c}, X)_{L1} = J_k(\boldsymbol{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k}|c_j|$$

$$= C \cdot J_k(\boldsymbol{c}, X)_0 + \frac{1}{2}\sum_{j=1}^{k}|c_j|$$

where $C = \dfrac{1}{\lambda}$

L_1 regularization with power weighting

$$J_k(\boldsymbol{c}, X)_{L1P} = J_k(\boldsymbol{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k} j^p |c_j|$$

$$= C \cdot J_k(\boldsymbol{c}, X)_0 + \frac{1}{2}\sum_{j=1}^{k} j^p |c_j|$$

$J_k(c, X)$ with L0 Regularization

L_0 regularization

$$J_k(\boldsymbol{c}, X)_{L0} = J_k(\boldsymbol{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k}\left(1-\delta_{c_j,0}\right) \text{ where } \delta_{x,y} = \begin{cases} 1 & \text{if x = y} \\ 0 & \text{if x} \neq \text{y} \end{cases}$$

$$= C \cdot J_k(\boldsymbol{c}, X)_0 + \frac{1}{2}\sum_{j=1}^{k}\left(1-\delta_{c_j,0}\right)$$

where $C = \dfrac{1}{\lambda}$

Gradient of $J_k(c, X)_0$

$$\frac{\partial J_k(\boldsymbol{c}, X)_0}{\partial c_m} = \frac{\partial}{\partial c_m}\left[\frac{1}{2n}\sum_{i=1}^{n}\left(f(x_i) - \sum_{j=0}^{k} c_j x_i^j\right)^2\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m\right)$$

Gradient of $J_k(c, X)_{L2}$

$$\frac{\partial J_k(\mathbf{c}, X)_{L2}}{\partial c_m} = \frac{\partial}{\partial c_m}\left[J_k(\mathbf{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k} c_j^2 \right]$$

$$= \frac{\partial}{\partial c_m}\left[J_k(\mathbf{c}, X)_0 \right] \frac{\partial}{\partial c_m}\left[\frac{\lambda}{2}\sum_{j=1}^{k} c_j^2 \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \frac{\lambda}{2}\sum_{j=1}^{k}\left[\frac{\partial}{\partial c_m} c_j^2 \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \frac{\lambda}{2} 2c_m$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \lambda c_m$$

$$= -C\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + c_m$$

Gradient of $J_k(c, X)_{L2P}$

$$\frac{\partial J_k(\boldsymbol{c}, X)_{L2P}}{\partial c_m} = \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k} j^p c_j^2 \right]$$

$$= \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0 \right] \frac{\partial}{\partial c_m}\left[\frac{\lambda}{2}\sum_{j=1}^{k} j^p c_j^2 \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \frac{\lambda}{2}\sum_{j=1}^{k}\left[\frac{\partial}{\partial c_m} j^p c_j^2 \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \frac{\lambda}{2} 2 m^p c_m$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \lambda m^p c_m$$

$$= -C\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + m^p c_m$$

Gradient of $J_k(c, X)_{L1}$

$$\frac{\partial J_k(\boldsymbol{c}, X)_{L1}}{\partial c_m} = \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k} \left| c_j \right| \right]$$

$$= \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0 \right] \frac{\partial}{\partial c_m}\left[\frac{\lambda}{2}\sum_{j=1}^{k} \left| c_j \right| \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m \right) + \frac{\lambda}{2}\sum_{j=1}^{k}\left[\frac{\partial}{\partial c_m}\left| c_j \right| \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m \right) + \frac{\lambda}{2} 2$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m \right) + \lambda$$

$$= -C\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m \right) + 1$$

Gradient of $J_k(c, X)_{L1P}$

$$\frac{\partial J_k(\boldsymbol{c}, X)_{L1P}}{\partial c_m} = \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k} j^p \left| c_j \right| \right]$$

$$= \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0 \right] \frac{\partial}{\partial c_m}\left[\frac{\lambda}{2}\sum_{j=1}^{k} j^p \left| c_j \right| \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \frac{\lambda}{2}\sum_{j=1}^{k}\left[\frac{\partial}{\partial c_m} j^p \left| c_j \right| \right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \frac{\lambda}{2} 2m^p$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + \lambda m^p$$

$$= -C\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i) \right)\left(x_i^m \right) + m^p$$

Gradient of $J_k(c, X)_{L0}$

$$\begin{aligned}
\frac{\partial J_k(\boldsymbol{c}, X)_{L0}}{\partial c_m} &= \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0 + \frac{\lambda}{2}\sum_{j=1}^{k}\left(1-\delta_{c_j,0}\right)\right] \\
&= \frac{\partial}{\partial c_m}\left[J_k(\boldsymbol{c}, X)_0\right] + \frac{\partial}{\partial c_m}\left[\frac{\lambda}{2}\sum_{j=1}^{k}\left(1-\delta_{c_j,0}\right)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m\right) + \frac{\lambda}{2}\left[\sum_{j=1}^{k}\frac{\partial}{\partial c_m}\left(1-\delta_{c_j,0}\right)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g_k(x_i)\right)\left(x_i^m\right) - \frac{\lambda}{2}\frac{\partial}{\partial c_m}\delta_{c_m,0} \\
&= \textbf{undefined}
\end{aligned}$$

$\delta_{c_m,0}$ is not differentiable since it is not a continuous function of c_m.

Thus, gradient descent alone cannot be used to optimize $J_k(\boldsymbol{c}, X)_{L0}$.

Gradient Descent Fitting Algorithm

1. Compute the gradient of the cost function, $\nabla J_k\left(\mathbf{c}, X\right)$, by summing over all (or a fixed number of randomly selected) training samples, $\left(x_i, f(x_i)\right)$,
2. Update the coefficients, **c**,

$$c_j := c_j + \Delta c_j, \text{ where}$$

$$\Delta c_j = -\eta \frac{\partial J\left(\mathbf{c}, X\right)}{\partial c_j}$$

and where η is the learning rate such that $0 < \eta < 1$.

Gradient Descent Fitting Algorithm

3. Repeat steps 1. and 2. until the coefficients converge,

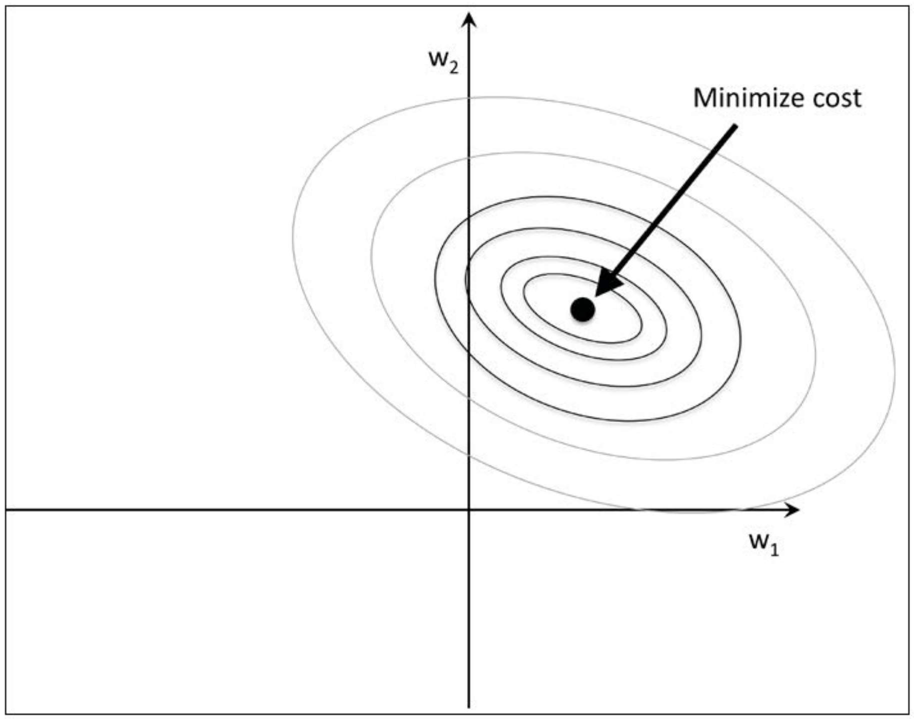
   that is, until

$$\|\Delta \boldsymbol{c}\| < \varepsilon \ (or \ \|\Delta \boldsymbol{c}\|_1 < \varepsilon), \text{where}$$

ε is the convergence threshold, $\varepsilon > 0$

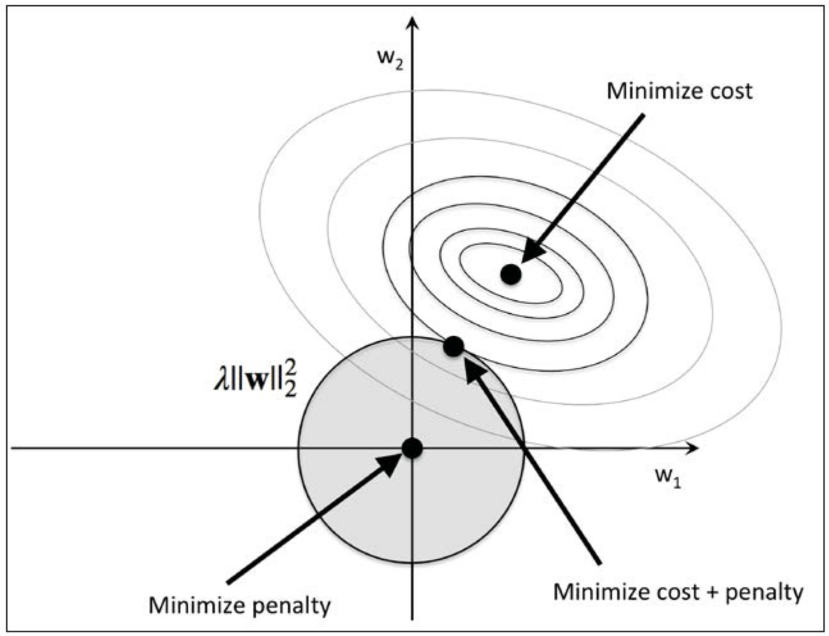
or for a set number of iterations.

No Regularization



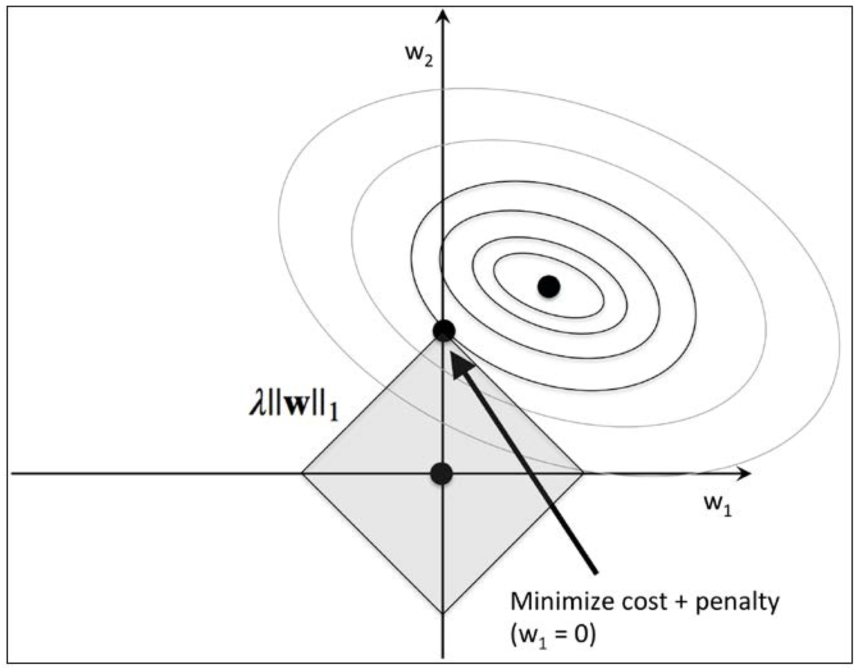
From Textbook

L2 Regularization



From Textbook

L1 Regularization



From Textbook