Least Squares Regression Example.

David Helmbold

University of California, Santa Cruz
dph@soe.ucsc.edu

W'16

Regression Example

Data (with add-a-dimension trick)

x_0	x_1	y
1	2	5
1	3	7
1	4	9

By eye $\theta = (1, 2)$ has zero error. (θ and $\mathbf{w}$ used interchangeably)
Data Matrix X is array of feature vectors:

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)\top} \\ \mathbf{x}^{(2)\top} \\ \mathbf{x}^{(3)\top} \end{bmatrix}$$

Label vector $\vec{y}$ or $\mathbf{y} = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}$

Note $\mathbf{x}^{(i)} \cdot \theta = \mathbf{x}^{(i)\top} \theta$,
So least squares criterion:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{3} (\mathbf{x}^{(i)} \cdot \theta - y^{(i)})^2 = \frac{1}{2} \sum_{i=1}^{3} (\mathbf{x}^{(i)\top} \theta - y^{(i)})^2$$

But

$$X\,\theta - \mathbf{y} = \begin{bmatrix} \mathbf{x}^{(1)} \cdot \theta \\ \vdots \\ \mathbf{x}^{(3)} \cdot \theta \end{bmatrix} - \mathbf{y} = \begin{bmatrix} \mathbf{x}^{(1)} \cdot \theta - y^{(1)} \\ \vdots \\ \mathbf{x}^{(3)} \cdot \theta - y^{(3)} \end{bmatrix}$$

So $J(\theta) = \frac{1}{2}(X\,\theta - \mathbf{y}) \cdot (X\,\theta - \mathbf{y}) = \frac{1}{2}(X\,\theta - \mathbf{y})^\top (X\,\theta - \mathbf{y})$

- Consider SGD for example $\mathbf{x}^{(1)} = (1, 2)$, $y^{(1)} = 5$, and θ initially $(1, 1)$.
- The squared-error on this example is $(1 * \theta_0 + 2 * \theta_1 - 5)^2 = 4$, and its <u>contribution</u> to $J(\theta)$ is 2 (half the squared error).

$$\frac{\partial \text{contribution}}{\partial \theta_0} = \frac{2}{2}(1 * \theta_0 + 2 * \theta_1 - 5) * 1 = -2$$
$$\frac{\partial \text{contribution}}{\partial \theta_1} = \frac{2}{2}(1 * \theta_0 + 2 * \theta_1 - 5) * 2 = -4$$

Previous version had typos above – fixed and changed step size to $1/20$ to keep rest the same

- with step size $\frac{1}{20}$, update θ to $\theta - \frac{1}{20}\nabla_\theta(\text{contribution})$
- New $\theta = (1, 1) - (\frac{-2}{20}, \frac{-4}{20}) = (1.1, 1.2)$

- Gradient step increases both
- Batch gradient - sum up contributions for all examples

- Closed form: minimizing θ is $(X^\top X)^{-1} X^\top \mathbf{y}$
- $X^\top X = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 9 \\ 9 & 29 \end{bmatrix}$
- $(X^\top X)^{-1} = \begin{bmatrix} 29/6 & -9/6 \\ -9/6 & 3/6 \end{bmatrix}$
- $(X^\top X)^{-1} X^\top \mathbf{y} = \begin{bmatrix} 29/6 & -9/6 \\ -9/6 & 3/6 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$