Q & A: Linear Regression

Lecture Series „Machine Learning“

Niels Landwehr

Research Group „Data Science“
Institute of Computer Science
University of Hildesheim

Quiz: Linear Regression

- Assume we want to learn a linear regression model $f:\mathbb{R}^2 \to \mathbb{R}$ of the form

$$f(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

  where $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$ is the input and $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^T \in \mathbb{R}^3$ is the parameter vector

- Assume a data set of three training examples $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3)\}$ given by

$$\mathbf{X} = \begin{array}{cc} \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 2 \end{pmatrix} & \begin{array}{c} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{array} \\ \begin{array}{ccc} x_0 & x_1 & x_2 \end{array} & \end{array} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}$$

Slido: 1100407

- Consider the following parameter vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3$. Which of the parameter vectors has the lowest squared loss?

$$\boldsymbol{\theta}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} \qquad \boldsymbol{\theta}_2 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \qquad \boldsymbol{\theta}_3 = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$$

Quiz: Linear Regression

- **Solution**:The predictions for the three different parameter vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3$ are:

$$\hat{\mathbf{y}}_1 = \mathbf{X}\boldsymbol{\theta}_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix} \rightarrow \text{Loss is } \frac{1}{3} \left\| \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} \right\|_2^2 = \frac{1}{3}(0+0+4) = \frac{4}{3}$$

$$\hat{\mathbf{y}}_2 = \mathbf{X}\boldsymbol{\theta}_2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix} \rightarrow \text{Loss is } \frac{1}{3} \left\| \begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} \right\|_2^2 = \frac{1}{3}(4+36+4) = \frac{44}{3}$$

$$\hat{\mathbf{y}}_3 = \mathbf{X}\boldsymbol{\theta}_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} \rightarrow \text{Loss is } \frac{1}{3} \left\| \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} - \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} \right\|_2^2 = 0$$

Quiz: Convexity

- Is the function $f:\mathbb{R}_{>0} \to \mathbb{R}$ given by $f(x) = x^2 + x - \log x$ convex?

  natural logarithm

  - A1: Yes, it is convex everywhere where the function is defined
  - A2: It is convex only for $x > 1$
  - A3: It is not convex

Quiz: Convexity

- Is the function $f : \mathbb{R}_{>0} \to \mathbb{R}$ given by $f(x) = x^2 + x - \log x$ convex?

- **Solution**: Yes, it is convex everywhere where it is defined. We use the criterion that the second derivative of the function is positive everywhere.

  Compute derivatives:

$$f'(x) = 2x + 1 - x^{-1} \qquad \text{Note: } \frac{\partial}{\partial x} \log x = \frac{1}{x}$$

$$f''(x) = 2 + x^{-2}$$

$$= 2 + \frac{1}{x^2}$$

- The second derivative $f''(x)$ is positive everywhere. Therefore the function is convex (special case of Hessian criterion mentioned in the lecture)

Quiz: Gradient Descent

- We want to minimize the function $L:\mathbb{R}\to\mathbb{R}, \quad L(\theta)=\theta^2$ using gradient descent with a learning rate $\eta\in\mathbb{R}$

- Question 1: Which of the following is the correct update rule?

$$
\begin{aligned}
&1. \quad \theta_{i+1}=\theta_i-\eta\theta_i^2\\
&2. \quad \theta_{i+1}=\theta_i+2\eta\theta_i\\
&3. \quad \theta_{i+1}=\theta_i-2\eta\theta_i\\
&4. \quad \theta_{i+1}=\theta_i-\theta_i
\end{aligned}
$$

- Question 2: We initialize the gradient descent procedure with $\theta_0=1$. For which learning rates η will gradient descent find the correct minimum?
  - 1. For all learning rates $\eta\in\mathbb{R}$
  - 2. For all learning rates $\eta\in\mathbb{R}_{>0}$
  - 3. For all learning rates $\eta\in(0,1]$
  - 4. For all learning rates $\eta\in(0,1)$
  - 5. Only for learning rates $\eta<0.01$

Quiz: Gradient Descent

- **Solution update rule:** The correct update rule is

$$\begin{aligned}\theta_{i+1} &= \theta_i - \eta \nabla L(\theta_i) \\ &= \theta_i - 2\eta\theta_i\end{aligned}$$

Quiz: Gradient Descent

- **Solution update rule:** The correct update rule is

$$\begin{aligned}\theta_{i+1} &= \theta_i - \eta \nabla L(\theta_i) \\ &= \theta_i - 2\eta\theta_i\end{aligned}$$

- **Solution convergence:** we first note that the minimum is at $\theta = 0$
- Let's try $\eta = 1$:

$$\left.\begin{aligned}\theta_0 &= 1 \\ \theta_1 &= \theta_0 - 2\theta_0 = -1 \\ \theta_2 &= \theta_1 - 2\theta_1 = 1 \\ \theta_3 &= \theta_2 - 2\theta_2 = -1\end{aligned}\right\} \rightarrow \text{This is not converging}$$

- Basically, $\eta < 1$ is needed such that $|\theta|$ is reduced in every iteration:

$$|\theta_{i+1}| = |\theta_i - 2\eta\theta_i| = |(1-2\eta)\theta_i| = \underbrace{|1-2\eta|}_{<1,\ \text{so}\ \eta<1}|\theta_i|$$

- Of course, η also needs to be positive. Therefore $\eta \in (0,1)$ is the right solution.