ML Fundamentals: Linear Regression

February 2, 2021

1 Hypothesis

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$
$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

2 Cost Function

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2$$

3 Parameters update using Gradient Descent

$$\begin{aligned}
\theta_j :&= \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y \right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y \right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y \right) \\
&= \left(h_\theta(x) - y \right) \cdot \frac{partial}{\partial \theta_j} \left(\sum_{i=0}^{n} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y \right) x_j \\
\theta_j :&= \theta_j + \alpha \left(y^{(i)} - h_\theta \left(x^{(i)} \right) \right) x_j^{(i)}
\end{aligned}$$

4 Batch Gradient Descent

Repeat until convergence {

$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta \left(x^{(i)} \right) \right) x_j^{(i)}$ (for every j)

}

5 Stochastic Gradient Descent

Loop {

for i=1 to m {

$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta \left(x^{(i)} \right) \right) x_j^{(i)}$ (for every j)

}

}

6 Probabilistic Interpretation

Why least mean squared error would be a reasonable choice for the linear regression.

$$
\begin{aligned}
y^{(i)} &= \theta^T x^{(i)} + \epsilon^{(i)} \\
\epsilon^{(i)} &\sim \mathcal{N}\left(0, \sigma^2\right) \\
p\left(\epsilon^{(i)}\right) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\epsilon^{(i)}\right)^2}{2\sigma^2}\right) \\
p\left(y^{(i)} \mid x^{(i)}; \theta\right) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right)
\end{aligned}
$$

Likelihood of parameters θ, is probability of y given x with the parameters θ. Note, the right terms to use are **likelihood of parameters**, and **probability of data**.

$$
\begin{aligned}
L(\theta) &= L(\theta; X, \vec{y}) = p(\vec{y} \mid X; \theta) \\
&\quad \text{Errors are from iid, independently and identically distributed} \\
L(\theta) &= \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}; \theta\right) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right) \\
&\quad \text{Lets maximize the log likelihood} \\
\ell(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2
\end{aligned}
$$

Hence, we can see that maximizing the log likelihood is equivalent to minimizing mean squared error

7 Locally Weighted Linear Regression

$$
\begin{gathered}
\text{Fit } \theta \text{ to minimize} \\
\sum_i w^{(i)} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \\
\text{where, } w^{(i)} = \exp\left(-\frac{\left(x^{(i)} - x\right)^2}{2\tau^2}\right) \\
\text{where, } \tau \text{ is called bandwidth parameter}
\end{gathered}
$$

Note, linear regression is **parametric** algorithm, where it has fixed set of parameters. Where as, locally weighted linear regression is called **non-parametric** as the numbers of parameters grows with the size of training set.