ICTS 4156: Introduction to ML

Razvan Bunescu

Lecture notes, February 24, 2021

1 Notes on lecture slides material

1.1 Gradient descent

Let $J(w) = \frac{1}{2}(w-4)^2 + 1$ and the initial guess $w_0 = 0$, for which $J(w_0) = 9$. Let the learning rate be $\eta = 0.5$. The gradient of J is $\nabla J(w) = w - 4$.

1. The gradient at w_0 is $\nabla J(w_0) = w_0 - 4 = -4$. The gradient update step is
   $w_1 = w_0 - \eta \nabla J(w_0) = 0 - 0.5 * -4 = 2$.
   So, $w_1 = 2$ for which $J(w_1) = 3$.

2. The gradient at w_1 is $\nabla J(w_1) = w_1 - 4 = -2$. The gradient update step is
   $w_2 = w_1 - \eta \nabla J(w_1) = 2 - 0.5 * -2 = 3$.
   So, $w_2 = 3$ for which $J(w_2) = 1.5$.

3. The gradient at w_2 is $\nabla J(w_2) = w_2 - 4 = -1$. The gradient update step is
   $w_3 = w_2 - \eta \nabla J(w_2) = 3 - 0.5 * -1 = 3.5$
   So, $w_4 = 3.5$ for which $J(w_3) = 1.125$.

4. and so on ... for ever?

Bonus points: For different values of η, plot on the same graph $J(w)$ and the points (in blue) corresponding to the gradient steps. Try $eta = 0.1, 0.5, 1, 2, ...$

"Until $J(w)$ does not improve": how do we quantify this? One method is to look at the *relative* change.

$$\Delta J = \left| \frac{J(w_t) - J(w_{t-1})}{J(w_{t-1})} \right| \tag{1}$$

If ΔJ is too small (0,0001) for a number of epochs (5 or 10), then you may consider stopping.

1.2 Feature scaling

Suppose feature x_j has values $1, 2, 3$ in the training data (3 training examples). The sample mean is $m_j = \frac{1+2+3}{3} = 2$. The sample standard deviation is $\sigma_j = \sqrt{\frac{(1-m_j)^2+(2-m_j)^2+(3-m_j)^2}{3}} = \sqrt{\frac{2}{3}}$. Then the feature x_j which had values $1, 2, 3$ will be scaled to the following values:

1. For training example 1, the new feature will be $\hat{x}_j = \frac{1-m_j}{\sigma_j} = \frac{1-2}{\sigma_j}$.
2. For training example 2, the new feature will be $\hat{x}_j = \frac{2-m_j}{\sigma_j} = \frac{2-2}{\sigma_j}$.
3. For training example 3, the new feature will be $\hat{x}_j = \frac{3-m_j}{\sigma_j} = \frac{3-2}{\sigma_j}$.

1.3 Optimization

We want to find w that minimizes $J(w) = \dfrac{1}{2N} \sum f(w)$.

$\nabla J(w) = 0$

$\dfrac{1}{2N} \sum \nabla f(w) = 0$

$\dfrac{1}{2} \sum \nabla f(w) = 0$

$\nabla J'(w) = 0$, where $J'(w) = N * J(w)$

w that minimizes $J(w) = (w-4)^2$ will also minimize $J'(w) = 10 * (w-4)^2$.