Discussion 7 **Alex Francis**
CS 189 Email: afrancis@berkeley.edu
02/21/2017 OH: Wednesday 10:00am - 12:00pm (283E Soda)

1 Feedback

Please give me feedback so that I can get better. Google form: `https://tinyurl.com/alex189-feedback`.

2 Lecture on Stochastic Gradient Descent

Last week, we discussed stochastic gradient descent for the first time in discussion, and I admit that the formulation in this class can seem rather hand-wavy. The idea of gradient descent is to take a step in the direction of the steepest descent of the cost function. In order to do that, we use the update rule,

$$w^{t+1} \leftarrow w^t - \eta \nabla_w J(w)$$

For example, when the combination of the cost function is "average loss" (a) in Shewchuk's notes, we have,

$$w^{t+1} \leftarrow w^t - \eta \nabla_w (\frac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i))$$

$$w^{t+1} \leftarrow w^t - \eta \nabla_w (\frac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i))$$

$$w^{t+1} \leftarrow w^t - \frac{\eta}{n} \sum_{i=1}^{n} \nabla_w (L(h(X_i), y_i))$$

This can be interpreted as taking a step in the direction of steepest descent over the entire cost function.

Stochastic gradient descent, in contrast, takes a step in the direction of steepest descent for only a *single* data point or, sometimes, for small subsets of the data (called a minibatch). SGD is utilized in practice because one can perform individual steps much more quickly. In addition, often large datasets cannot be held in the RAM of an individual computer, and training algorithms are complex, distributed tasks performed across arrays of workers.

But, why is choosing a data vector uniformly at random for gradient descent a good approach? It turns out that the stochastic gradient descent update and the batch gradient descent update are *equal* in expectation! Denote $\tilde{J}(w) = L(h(X_j), y_j)$ for some j chosen uniformly at random from $j \in \{1, \ldots, n\}$.

$$\begin{aligned} \mathbb{E}(\nabla_w J(w) - \nabla_w \tilde{J}(w)) &= \mathbb{E}(\nabla_w J(w)) - \mathbb{E}(\nabla_w \tilde{J}(w)) \\ &= \mathbb{E}(\nabla_w \frac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i)) - \mathbb{E}(\tilde{J}(w)) \end{aligned}$$

The first term is deterministic. There is no random element to it, so the expected value disappears. The second expression can be computed by using the formula for expectation. Recall that $\mathbb{E}(X) = \sum_x xP(X = x)$. Then,

$$\mathbb{E}(\tilde{J}(w)) = \sum_{j=1}^{n} \frac{1}{n} L(h(X_j), y_j)$$

So,

$$\begin{aligned}
\mathbb{E}(\nabla_w J(w) - \nabla_w \tilde{J}(w)) &= \mathbb{E}(\nabla_w \frac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i)) - \mathbb{E}(\tilde{J}(w)) \\
&= \nabla_w \frac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i) - \nabla_w \frac{1}{n} \sum_{j=1}^{n} L(h(X_j), y_j)) \\
&= 0
\end{aligned}$$

Therefore, the SGD update is an *unbiased estimator* for the BGD update in the case where the average loss cost function is used. We have a statistical justification for SGD!

3 Worksheet Problems

Please see the staff solutions, as I wrote them this week.