Unconstrained optimization III

Topics we'll cover

1. Stochastic gradient descent for logistic regression
2. Stochastic gradient descent more generally

Recall: gradient descent for logistic regression

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots$, until convergence:

$$w_{t+1} \;=\; w_t \;+\; \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \Pr_{w_t}(-y^{(i)}|x^{(i)})$$

Each update involves the entire data set, which is inconvenient.

Stochastic gradient descent: update based on just one point:

- Get next data point (x, y) by cycling through data set
- $w_{t+1} \;=\; w_t \;+\; \eta_t \, y \, x \Pr_{w_t}(-y|x)$

Decomposable loss functions

Loss function for logistic regression:

$$L(w) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})}) \;=\; \sum_{i=1}^{n} (\text{loss of } w \text{ on } (x^{(i)}, y^{(i)}))$$

Most ML loss functions are like this: for training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$,

$$L(w) \;=\; \sum_{i=1}^{n} \ell(w; x^{(i)}, y^{(i)})$$

where $\ell(w; x, y)$ captures the loss on a single point.

Gradient descent and stochastic gradient descent

For minimizing

$$L(w) \; = \; \sum_{i=1}^{n} \ell(w; x^{(i)}, y^{(i)})$$

Gradient descent:

- $w_o = 0$
- while not converged:
  - $w_{t+1} = w_t - \eta_t \sum_{i=1}^{n} \nabla \ell(w_t; x^{(i)}, y^{(i)})$

Stochastic gradient descent:

- $w_o = 0$
- Keep cycling through data points (x, y):
  - $w_{t+1} = w_t - \eta_t \nabla \ell(w_t; x, y)$

Variant: mini-batch stochastic gradient descent

Stochastic gradient descent:

- $w_o = 0$
- Keep cycling through data points (x, y):
  - $w_{t+1} = w_t - \eta_t \nabla \ell(w_t; x, y)$

Mini-batch stochastic gradient descent:

- $w_o = 0$
- Repeat:
  - Get the next batch of points B
  - $w_{t+1} = w_t - \eta_t \sum_{(x,y) \in B} \nabla \ell(w_t; x, y)$