Stochastic gradient descent

Many learning problems look like

$$\arg \min_w L(w, x_i, y_i) + \lambda \|w\|_p$$

↑

loss on examples

↑

eg. SVM.

$$\text{loss is } \sum_i \text{ hinge loss} \left(w, \text{ example } i \right)$$

eg. logistic regression

$$\text{loss is } \sum_i \text{ exp loss}.$$
In these cases, two important obs.

1) We don't so much care about

\[ E = \sum \ell (w, \text{example}_i) \]

for empirical loss

\[ \sum \ell (w, \text{example}_i) \cdot = \int \]

for true loss

i.e. examples

i.e. everything

This means that the min of the optimization problem \( w^* \) may not be best \( w \) (which is \( w^{opt} \))

\[ w^{opt} = \arg \min \mathcal{L} \]

\[ w^* = \arg \min \mathcal{E} \]

error incurred by exact optimization

\[ \Rightarrow \text{This means inexact opt might be OK.} \]
2) Exact optimization could be hugely expensive

(millions of examples; millions of feats)

Idea: Pick example (feature) at random, compute gradient for that alone, take small step.

STOCHASTIC GRADIENT DESCENT

- How far?

Steplengths \( s_i \) should have property

\[
\begin{align*}
\cdot & \quad s_N \to 0, \quad N \to \infty \\
\cdot & \quad \sum_{i=1}^{\infty} s_i \to \infty, \quad R \to \infty
\end{align*}
\]

(i.e., path must be infinitely long, but steps must become infinitely small)
Logistic regression with $\ell_2$ norm

$$\min_w \sum_i \left[ L_e(w, y, x) + \frac{\lambda}{N} \|w\|^2 \right] = E + R$$

for each step, choose $I$ examples, say at $k$th

$$g^{(n)}(x) = \left[ \frac{x_k}{1 + e^{w^{(n)}x_k}} - \left( \frac{y_k + 1}{2} \right) x_k \right]$$

$$+ 2 \frac{\lambda}{N} w^{(n)}$$

$$w^{(n+1)} = w^{(n)} + \delta_i \cdot g^{(n)}$$

Notice, because we choose UAR,

$$E(g^{(n)}) = \nabla_w [E + R]$$
typically, step lengths look like

\[ s_i = \frac{1}{i} \]

(this satisfies constraints).

We get (with work)

\[ \frac{1}{(w_t - w^*)^2} \]

grows linearly in \( t \)

i.e. early steps really help, late steps don't do much.

Notice: algorithm is on-line
What about SVM?

\[ \sum_i \left[ l_i (w, x_i, y_i) + \frac{\lambda}{N} w^T w \right] \]

which is not differentiable.

\[ l_i = \max(0, 1 - y_i(w^T x_i + b)) \]

\[ l_i, y_i = 1 \]

however, this is convex.

we invoke the sub-gradient
Consider the graph of a function 

\[(x_1, x_2, x_3, \ldots, x_n, f(x_1, \ldots, x_n))\]

This is a surface.

Assume \( f \) is differentiable.

Surface has a normal

\[N = k \left( \frac{-\partial f}{\partial x_1}, \frac{-\partial f}{\partial x_2}, \ldots, +1 \right)\]

\[= k \left( -\nabla f, +1 \right)\]

Example: curve in the plane is graph of \( f \) of 1 var

Notice we can read gradient off normal
Now, if the function is not diff, we can come up with normal cone. All of these vectors are normals at this pt.

Notice, if needs to be convex so we know what interior is for this construction.

Graph of fn of two vars.
- read off gradient corresponding to any element of normal cone
- this is subgradient \( \nabla \)
- moving backward along subgradient will guarantee descent for small enough step
- subgradient of diff. fn = gradient

\[ \text{SVM} \]

\[ \nabla_{\omega} L(\omega) = \begin{cases} 0 & \text{if } L_h(\omega) = 0 \\ -y_i x_i & \text{otherwise} \end{cases} \]
Stochastic Subgradient Descent is

\[
\begin{array}{l}
\text{choose } k\text{'th example at random}
\end{array}
\]

\[
\begin{array}{l}
\text{if right, } w^{(n+1)} = w^{(n)} (1 - \frac{\lambda \delta_i}{2N})
\end{array}
\]

\[
\begin{array}{l}
\text{if wrong, } w^{(n+1)} = w^{(n)} - \delta_i \left[ -y_i x_i \pm \frac{\lambda w^{(n)}}{2N} \right]
\end{array}
\]

This is amazingly effective.
$l_1$ regularization

$$||w||_1 = \sum_i |w_i|$$
$$||w||_2 = \sum_i w_i^2$$

Notice: in $l_2$ norm, small $w_i$ have little effect; hence, not much advantage in trying to zero; in $l_1$ norm, effect of small $w_i$ is substantial; they tend to go to zero, resulting in sparsity (desirable)
\[ \min_{\mathbf{w}} \sum_{i} l_{ve}(w, y_i, x_i) + \lambda \| \mathbf{w} \|_1, \]

- not differentiable
- subgradient unlikely to enforce sparsity

alternative (equivalent)

\[ \min_{\mathbf{w}} \sum_{i} l_{ve}(w, y_i, x_i) + \sum_{e} h_{ve}. \]

\[-h_k \leq w_k \leq h_k\]

(Notice this is a box problem; solve with interior point method; excellent version due to Koh et al.)