Logistic Regression

$$P(y=1 \mid \underline{x}, \underline{w}) = \sigma(\underline{w}^T \underline{x})$$

Match $y^{(n)}$ to $\sigma(\underline{w}^T \underline{x}^{(n)})$

Negative log likelihood:

$$NLL = -\sum_{n=1}^{N} \log \sigma(z^{(n)} \underline{w}^T \underline{x}^{(n)})$$

↑ Label $\in \{-1, +1\}$

$z = 2y - 1$

$$\nabla_{\underline{w}} NLL = -\sum_{n=1}^{N} (1 - \sigma_n) z^{(n)} \underline{x}^{(n)}$$

Stochastic Gradient Descent (SGD)

Average gradient

$$\frac{1}{N} \nabla_{\underline{w}} NLL = \frac{1}{N} \sum_{n=1}^{N} \underbrace{g_n}$$

eg. $-(1-\sigma_n) z^n \underline{x}^{(n)}$

Approximation

Batch of B examples at random

$$\approx \frac{1}{B} \sum_{b=1}^{B} g_b$$

Stochastic / Online g.d.
uses a batch with $B=1$

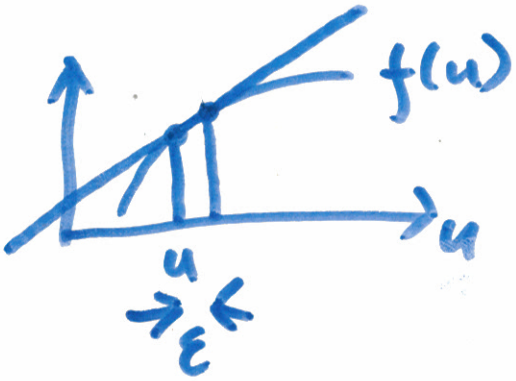
"Minibatch" uses $B \approx 64, 128$

contour of
cost fn

I quoted §10.6
"Numerical Recipes in Fortran"
(2nd Ed.)

Check your Derivatives

$$f'(u) \approx \frac{f(u+\varepsilon) - f(u)}{\varepsilon}$$



Error $O(\varepsilon)$

$\varepsilon \approx 10^{-5}$

$$\approx \frac{f(u+\varepsilon/2) - f(u-\varepsilon/2)}{\varepsilon}$$

Error $O(\varepsilon^2)$

Softmax Regression

Fit one-hot-encoded labels

Target $y = [0\ 0\ 0 \ldots 0\ 1\ 0 \ldots 0]$

cth location indicating class c

Fit f. Our model:

$P(\text{class } c) = f_c$

Each f_k to be positive:

$$f_k \propto e^{\underline{w}^{(k)T}\underline{x}}$$

We want f to be normalized

$$\Sigma_k f_k = 1:$$

$$f_k = \frac{e^{\underline{w}^{(k)T}\underline{x}}}{\Sigma_{k'} e^{\underline{w}^{(k')T}\underline{x}}}$$

Softmax

Model has params

$$W = \{ \underline{w}^{(k)} \}$$

Max. the likelihood of W

In one example at $\underline{x}$ and we see that it has class c:

$$\log f_c = \underline{w}^{(c)T}\underline{x} - \log \sum_{k'} e^{\underline{w}^{(k')T}\underline{x}}$$

$$\nabla_{\underline{w}^{(k)}} \log f_c = \delta_{ck}\underline{x} - \frac{1}{\sum_{k'} \cdots} e^{\underline{w}^{(k)T}\underline{x}}\,\underline{x}$$

δ_{ck}: Kronecker delta

$$= (y_k - f_k)\underline{x}$$

Logistic Regression?

Two classes

$$P(y=1 \mid \underline{x}, W) = \frac{e^{\underline{w}^{(1)T}\underline{x}}}{e^{\underline{w}^{(1)T}\underline{x}} + e^{\underline{w}^{(2)T}\underline{x}}}$$

$$= \frac{1}{1 + e^{(\underline{w}^{(2)T} - \underline{w}^{(1)T})\underline{x}}}$$

$$= \sigma\Big(\underbrace{(\underline{w}^{(1)} - \underline{w}^{(2)})^T}_{\text{"}\underline{w}\text{"}}\underline{x}\Big)$$