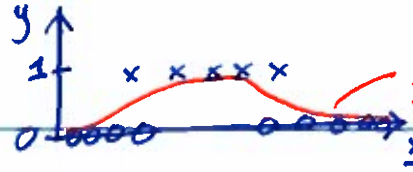
Logistic Regression

for $y \in \{0, 1\}$

Model the outputs: $P(y=1 \mid \underline{x}, \underline{w}) = f(\underline{x}; \underline{w}) = \sigma(\underline{w}^T \underline{x})$



Maximum Likelihood (minimize -ve log likelihood)

$$NLL = -\sum_n \log p(y^{(n)} \mid \underline{x}^{(n)}, \underline{w}) = -\sum_n \log \sigma\left(\underbrace{z^{(n)}}_{2y^{(n)}-1} \underline{w}^T \underline{x}^{(n)}\right)$$

σ_n

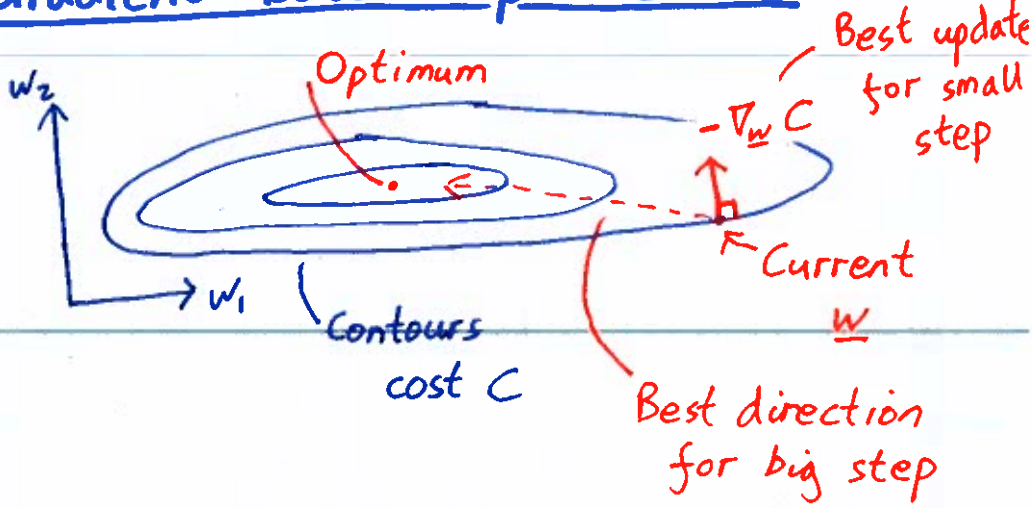
$$\nabla_{\underline{w}} NLL = -\sum_n \nabla_{\underline{w}} \log \sigma_n \qquad \frac{d \log a}{da} = \frac{1}{a} \text{ + chain rule}$$

$$= -\sum_n \frac{1}{\sigma_n} \nabla_{\underline{w}} \sigma_n \qquad \frac{d\sigma(a)}{d(a)} = \sigma(a)(1-\sigma(a))$$

$$= -\sum_n \frac{1}{\sigma_n} \sigma_n (1-\sigma_n) \nabla_{\underline{w}} z^{(n)} \underline{w}^T \underline{x}^{(n)}$$

$$= \underbrace{-\sum_n (1-\sigma_n) z^{(n)} \underline{x}^{(n)}}_{\sigma(-z^{(n)} \underline{w}^T \underline{x}^{(n)})} = -\sum_n (y^{(n)} - \underbrace{f^{(n)}}_{\sigma(\underline{w}^T \underline{x}^{(n)})}) \underline{x}^{(n)}$$

Gradient-based optimization



Finding better directions

- Non-linear conjugate gradients
- L-BFGS
- Newton's method ...

Details non-examinable

(might cover later)

Often "batch" methods:

C and $\nabla_{\underline{w}} C$ use whole data set

Stochastic / Online or "Minibatch" use data subsets

Stochastic Gradient Descent

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

Average Gradient over examples:

$$g = \frac{1}{N} \nabla_w C = \frac{1}{N} \sum_{n=1}^{N} \nabla_w C_n$$

C_n ← Cost for nth example

Monte Carlo Approximation:

Sample mini-batch of B examples:

$$\approx \frac{1}{B} \sum_{b=1}^{B} \nabla_w C_{n_b} = \hat{g}$$

$n_b \sim$ sampled from $\{1..N\}$

SGD

Initialize $w \leftarrow w^{(0)}$ (0 ok for logistic regression)

for $t = 1 \ldots T$:

$$w \leftarrow w - \eta_t \hat{g}^{(t)}$$

η_t ← step-size

$\hat{g}^{(t)}$ ← gradient est. at time t using current w

Softmax Regression

Multi-class classification

Target $y^{(n)} = [0\ 0\ 0 \ldots 0\ 1\ 0 \ldots 0]^T$

↑ c^{th} location

Example n has label "$y = c$"

Fit $\underline{f}$, predictions from our model:

$$P(\text{class} = c \mid \underline{x}, W) = f_c(\underline{x}; W)$$

Positive score for each class:

$$s_k = e^{\underline{w}^{(k)T}\underline{x}} \qquad k = 1 \ldots K$$

(K = # classes)

Want $\underline{f}$ to be normalized: $\sum_k f_k = 1$

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

$$\underset{K\times 1}{\underline{f}} = \text{softmax}\left(\underset{K\times D}{W}\ \underset{D\times 1}{\underline{x}}\right)$$

D = # features

Model has parameters W ($K \times D$)

$$W = \begin{pmatrix} \text{---} \; \underline{w}^{(1)T} \; \text{---} \\ \vdots \\ \text{---} \; \underline{w}^{(k)T} \; \text{---} \end{pmatrix}$$

Maximize likelihood of W

For SGD we use one example at $\underline{x}$ with class label c

$-\log$ prob. of this example given W

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

$$-\nabla_{\underline{w}^{(k)}} \log f_c = \underbrace{-\delta_{kc}}_{\text{KroneckerDelta}}\underline{x} + \underbrace{\frac{1}{\sum_{k'} \cdots} e^{\underline{w}^{(k)T}\underline{x}}}_{f_k}\underline{x}$$

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

Logistic Regression

$k=2$, two classes

$$P(y=1 \mid \underline{x}, W) = \frac{e^{\underline{w}^{(1)T}\underline{x}}}{e^{\underline{w}^{(1)T}\underline{x}} + e^{\underline{w}^{(0)T}\underline{x}}}$$

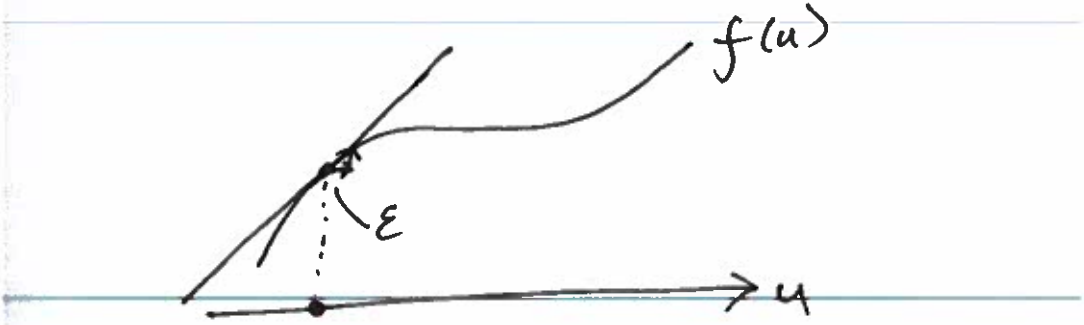
$$= \frac{1}{1 + e^{\underline{w}^{(0)T}\underline{x} - \underline{w}^{(1)T}\underline{x}}}$$

$$\left[\sigma(a) = \frac{1}{1+e^{-a}}\right]$$

$$= \sigma\left(\underbrace{(\underline{w}^{(1)} - \underline{w}^{(0)})^T}_{\text{"}\underline{w}\text{"}}\underline{x}\right)$$

Parameters are redundant.

Check your derivatives



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

correct $\lim \varepsilon \to 0$

With "doubling" floating point: $\varepsilon = 10^{-5}$, error $O(\varepsilon)$

Notes: Central difference

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

Error $O(\varepsilon^2)$

(Non-examinable)

Probability estimation question

Iain is juggling.

Assume he has a 1% chance of dropping each catch.

(not really independent, but let's pretend)

Estimate without a computer/calculator the probability he manages 100 catches in a row on first attempt.