Deep Learning
Summary of lecture 2

Dr. Richard E. Turner (ret26@cam.ac.uk)

Engineering Tripos Part IB
Paper 8: Information Engineering

Summary of lecture 2

fitting method 1: maximum likelihood fit

$$G(\boldsymbol{w}) = -\sum_n \left[y^{(n)} \log \mathrm{x}(\boldsymbol{z}^{(n)};\boldsymbol{w}) + (1-y^{(n)})\log\left(1-\mathrm{x}(\boldsymbol{z}^{(n)};\boldsymbol{w})\right)\right]$$

relative entropy / data fit

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} G(\boldsymbol{w})$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w}) = -\sum_n (y^{(n)} - x^{(n)})\boldsymbol{z}^{(n)}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w})$$

Summary of lecture 2

fitting method 1: maximum likelihood fit

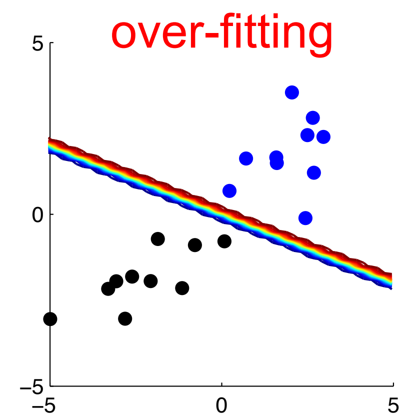
$$G(\boldsymbol{w}) = -\sum_n \left[y^{(n)} \log \mathrm{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) + (1 - y^{(n)}) \log \left(1 - \mathrm{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w})\right)\right]$$

relative entropy / data fit

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} G(\boldsymbol{w})$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w}) = -\sum_n (y^{(n)} - x^{(n)}) \boldsymbol{z}^{(n)}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w})$$



Summary of lecture 2

fitting method 1: maximum likelihood fit

$$G(\boldsymbol{w}) = -\sum_n \left[y^{(n)} \log \mathrm{x}(\boldsymbol{z}^{(n)};\boldsymbol{w}) + (1-y^{(n)}) \log\left(1-\mathrm{x}(\boldsymbol{z}^{(n)};\boldsymbol{w})\right)\right]$$

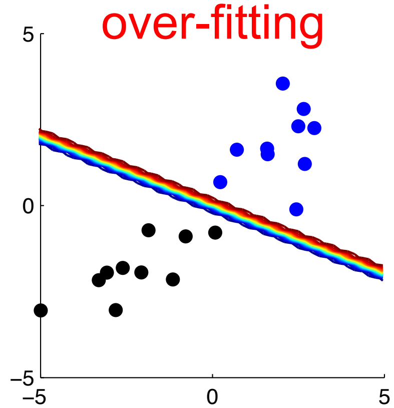
relative entropy / data fit

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} G(\boldsymbol{w})$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w}) = -\sum_n (y^{(n)} - x^{(n)})\boldsymbol{z}^{(n)}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w})$$

over-fitting



fitting method 2: regularised maximum likelihood

$$E(\boldsymbol{w}) = \tfrac{1}{2}\sum_i w_i^2$$

"regulariser" prevents extreme weights

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} M(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \left[G(\boldsymbol{w}) + \alpha E(\boldsymbol{w})\right]$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w}) = -\sum_n (y^{(n)} - x^{(n)})\boldsymbol{z}^{(n)} + \alpha \boldsymbol{w}$$

weight decay

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w})$$

Summary of lecture 2

fitting method 1: maximum likelihood fit

$$G(\boldsymbol{w}) = -\sum_n \left[y^{(n)} \log \mathrm{x}(\boldsymbol{z}^{(n)};\boldsymbol{w}) + (1-y^{(n)}) \log\left(1-\mathrm{x}(\boldsymbol{z}^{(n)};\boldsymbol{w})\right)\right]$$

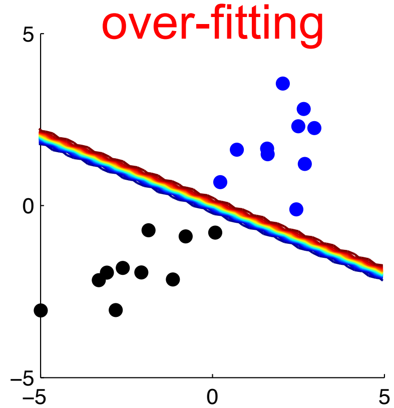
relative entropy / data fit

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\arg\min}\; G(\boldsymbol{w})$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w}) = -\sum_n (y^{(n)} - x^{(n)})\boldsymbol{z}^{(n)}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} G(\boldsymbol{w})$$

over-fitting



fitting method 2: regularised maximum likelihood

$$E(\boldsymbol{w}) = \tfrac{1}{2}\sum_i w_i^2$$

"regulariser" prevents extreme weights

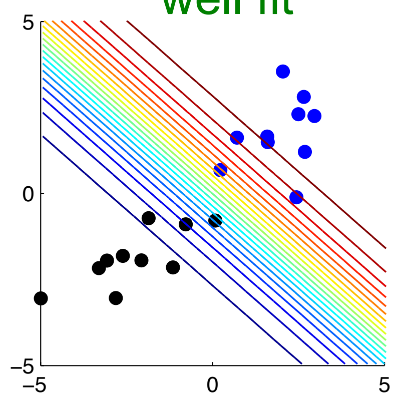
$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\arg\min}\, M(\boldsymbol{w}) = \underset{\boldsymbol{w}}{\arg\min}\left[G(\boldsymbol{w}) + \alpha E(\boldsymbol{w})\right]$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w}) = -\sum_n (y^{(n)} - x^{(n)})\boldsymbol{z}^{(n)} + \alpha\boldsymbol{w}$$

weight decay

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w})$$

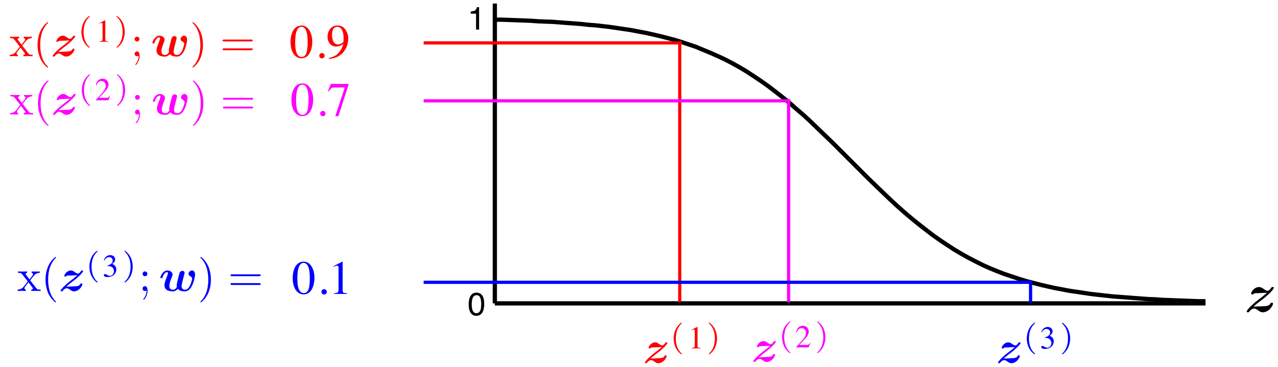
well fit



Question

Observe 3 labelled data points with scalar inputs and I have single neuron:

$$\mathrm{x}(\boldsymbol{z}\,;\boldsymbol{w}) = p(y\, = 1|\boldsymbol{z}\,,\boldsymbol{w})$$



What is the probability of the observed labels given the inputs and weights?

$$p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = \prod_{n=1}^{N} p(y^{(n)}|\boldsymbol{z}^{(n)},\boldsymbol{w})$$

A. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9^2 \times 0.7$

B. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9 \times 0.3 \times 0.1$

C. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9^2 \times 0.3$

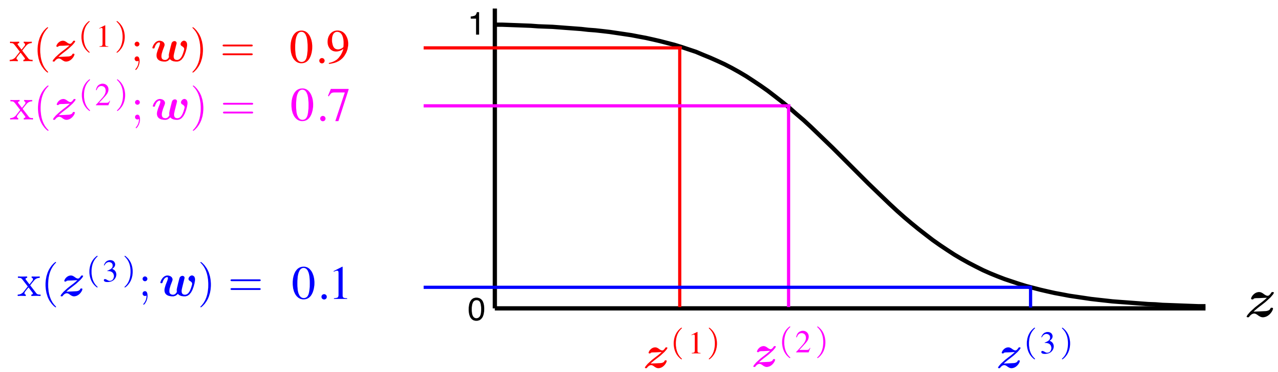
D. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9 \times 0.7 \times 0.1$

E. I don't know!

Question

Observe 3 labelled data points with scalar inputs and I have single neuron:

$$\mathrm{x}(\boldsymbol{z}\,;\boldsymbol{w}) = p(y\,=1|\boldsymbol{z}\,,\boldsymbol{w})$$



What is the probability of the observed labels given the inputs and weights?

$$p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = \prod_{n=1}^{N} p(y^{(n)}|\boldsymbol{z}^{(n)},\boldsymbol{w})$$

A. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9^2 \times 0.7$

B. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9 \times 0.3 \times 0.1$

C. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9^2 \times 0.3$

D. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9 \times 0.7 \times 0.1$

E. I don't know!

Learning by improving the likelihood of the parameters

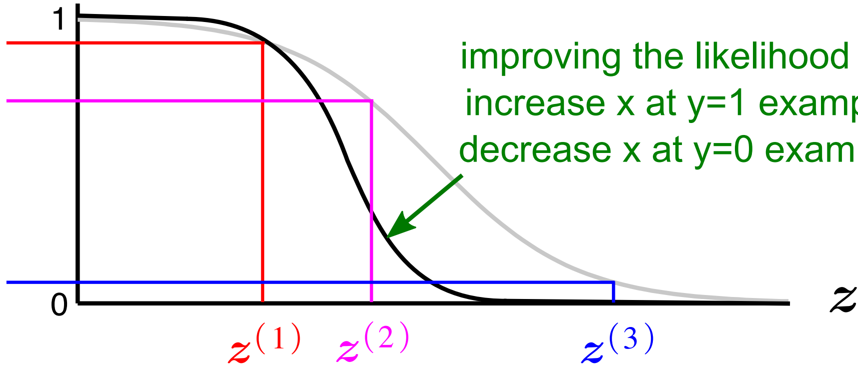
Observe 3 labelled data points with scalar inputs and I have single neuron:

$$\mathrm{x}(\boldsymbol{z}\,;\boldsymbol{w}) = p(y\;=1|\boldsymbol{z}\;,\boldsymbol{w})$$

$\mathrm{x}(\boldsymbol{z}^{(1)};\boldsymbol{w}) = 0.9$

$\mathrm{x}(\boldsymbol{z}^{(2)};\boldsymbol{w}) = 0.7$

$\mathrm{x}(\boldsymbol{z}^{(3)};\boldsymbol{w}) = 0.1$



$y^{(1)} = 1$

$y^{(2)} = 0$

$y^{(3)} = 0$

What is the probability of the observed labels given the inputs and weights?

$$p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = \prod_{n=1}^{N} p(y^{(n)}|\boldsymbol{z}^{(n)},\boldsymbol{w})$$

also known as the likelihood of the parameters

A. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9^2 \times 0.7$

B. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9 \times 0.3 \times 0.1$

C. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9^2 \times 0.3$

D. $p(\{y^{(n)}\}_{n=1}^{N}|\{\boldsymbol{z}^{(n)}\}_{n=1}^{N},\boldsymbol{w}) = 0.9 \times 0.7 \times 0.1$

E. I don't know!