Deep Learning Summary of lecture 2

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

Engineering Tripos Part IB Paper 8: Information Engineering

fitting method 1: maximum likelihood fit

$$G(\boldsymbol{w}) = -\sum_{n} \left[y^{(n)} \log \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) + (1 - y^{(n)}) \log \left(1 - \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) \right) \right] \text{ relative entropy } f$$

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

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

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

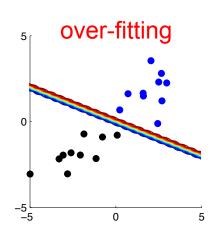
fitting method 1: maximum likelihood fit

$$G(\boldsymbol{w}) = -\sum_{n} \left[y^{(n)} \log \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) + (1 - y^{(n)}) \log \left(1 - \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) \right) \right] \quad \frac{\text{relative entropy } / \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w})}{\text{data fit}}$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{arg \,min}} G(\mathbf{w})$$

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

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



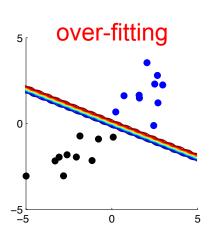
fitting method 1: maximum likelihood fit

$$G(\boldsymbol{w}) = -\sum_{n} \left[y^{(n)} \log \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) + (1 - y^{(n)}) \log \left(1 - \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) \right) \right] \quad \frac{\text{relative entropy } / \mathbf{x}}{\text{data fit}}$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{arg \,min}} G(\mathbf{w})$$

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

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



fitting method 2: regularised maximum likelihood

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{i} w_{i}^{2}$$
 "regulariser" prevents extreme weights
$$\boldsymbol{w}^{*} = \operatorname*{arg\,min}_{\boldsymbol{w}} M(\boldsymbol{w}) = \operatorname*{arg\,min}_{\boldsymbol{w}} [G(\boldsymbol{w}) + \alpha E(\boldsymbol{w})]$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w}) = -\sum_{n} (y^{(n)} - x^{(n)}) \boldsymbol{z}^{(n)} + \alpha \boldsymbol{w}$$
 weight decay
$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w})$$

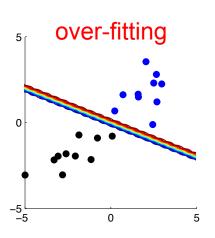
fitting method 1: maximum likelihood fit

$$G(\boldsymbol{w}) = -\sum_{n} \left[y^{(n)} \log \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) + (1 - y^{(n)}) \log \left(1 - \mathbf{x}(\boldsymbol{z}^{(n)}; \boldsymbol{w}) \right) \right]$$
 relative entropy / data fit

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{arg \,min}} G(\mathbf{w})$$

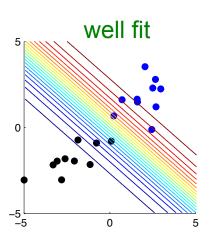
$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} G(\mathbf{w}) = -\sum_{n} (y^{(n)} - x^{(n)}) \mathbf{z}^{(n)}$$

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



fitting method 2: regularised maximum likelihood

$$\begin{split} E(\boldsymbol{w}) &= \frac{1}{2} \sum_{i} w_{i}^{2} & \text{"regulariser" prevents extreme weights} \\ \boldsymbol{w}^{*} &= \mathop{\arg\min}_{\boldsymbol{w}} M(\boldsymbol{w}) = \mathop{\arg\min}_{\boldsymbol{w}} \left[G(\boldsymbol{w}) + \alpha E(\boldsymbol{w}) \right] \\ \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w}) &= -\sum_{n} (y^{(n)} - x^{(n)}) \boldsymbol{z}^{(n)} + \alpha \boldsymbol{w} & \text{weight decay} \\ \boldsymbol{w} &\leftarrow \boldsymbol{w} - \eta \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} M(\boldsymbol{w}) \end{split}$$



Question

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

$$\mathbf{x}(\boldsymbol{z}^{(1)}; \boldsymbol{w}) = 0.9$$
 $\mathbf{x}(\boldsymbol{z}^{(1)}; \boldsymbol{w}) = 0.7$
 $\mathbf{x}(\boldsymbol{z}^{(2)}; \boldsymbol{w}) = 0.7$
 $\mathbf{x}(\boldsymbol{z}^{(3)}; \boldsymbol{w}) = 0.1$
 $\mathbf{x}(\boldsymbol{z}^{(3)}; \boldsymbol{w}) = 0.1$
 $\mathbf{x}(\boldsymbol{z}^{(3)}; \boldsymbol{w}) = 0.1$
 $\mathbf{x}(\boldsymbol{z}^{(3)}; \boldsymbol{z}^{(3)})$

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:

$$\mathbf{x}(\mathbf{z}^{(1)}; \mathbf{w}) = 0.9$$
 $\mathbf{x}(\mathbf{z}^{(1)}; \mathbf{w}) = 0.7$
 $\mathbf{x}(\mathbf{z}^{(2)}; \mathbf{w}) = 0.7$
 $\mathbf{x}(\mathbf{z}^{(3)}; \mathbf{w}) = 0.1$
 $\mathbf{x}(\mathbf{z}^{(3)}; \mathbf{w}) = 0.1$
 $\mathbf{x}(\mathbf{z}^{(3)}; \mathbf{z}^{(3)})$
 $\mathbf{z}^{(3)}$

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

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:

$$\mathbf{x}(\boldsymbol{z}^{(1)};\boldsymbol{w}) = 0.9$$
 improving the likelihood will:
$$\mathbf{x}(\boldsymbol{z}^{(2)};\boldsymbol{w}) = 0.7$$
 increase x at y=1 examples decrease x at y=0 examples
$$\mathbf{x}(\boldsymbol{z}^{(3)};\boldsymbol{w}) = 0.1$$

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

$$\mathbf{z}^{(1)} = 0$$

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

$$p(\{y^{(n)}\}_{n=1}^N|\{m{z}^{(n)}\}_{n=1}^N,m{w})=\prod_{n=1}^N p(y^{(n)}|m{z}^{(n)},m{w})$$
 also known as the likelihood of the parameters $p(\{y^{(n)}\}_{n=1}^N|\{m{z}^{(n)}\}_{n=1}^N,m{w})=0.02\times0.7$

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.
$$p(\{y^{(n)}\}_{n=1}^{N}|\{oldsymbol{z}^{(n)}\}_{n=1}^{N},oldsymbol{w})=0.9^2\times0.3$$

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!