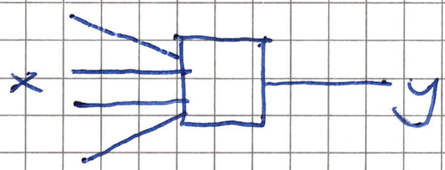
Bayesian neural networks

W3a

Consider a neuron with

- inputs x_i $\quad i=1, \ldots, I$
- weights w_i $\quad i=1, \ldots, I$
  and bias w_0
- output $y(a)$ or $y(x_i; w)$
  where $a = \sum_i w_i x_i$ is the activation



Example: Classifier

Train (adjust w) to minimize the loss function

$$M(w) = G(w) + \alpha E_w(w)$$

with the error function

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

where the training data $t^{(n)}$ is either 0 or 1 (two classes)

$\boxed{W3a}$

The regularizer

$$E_w(w) = \frac{1}{2} \sum_i w_i^2$$

is designed to avoid overfitting.

We interpret the output

$$y(x;w) \equiv p(t=1 \mid x, w)$$

i.e the probability that input x belong to class 1.

Together with $p(t=0 \mid w, x) = 1-y$

we get

$$p(t \mid w, x) = y^t (1-y)^t = \exp\left[t \log(y) + (1-t)\log(1-y)\right]$$

So the error function can be interpreted as minus the log likelihood

$$p(D \mid w) = e^{-G(w)}$$

Similarly, the regularizer can be interpreted as a log prior

$$p(w \mid \alpha) = \frac{1}{Z_w(\alpha)} \exp(-\alpha E_w)$$

where this is a Gaussian for the quadratic E_w.

W3a

In conclusion, the loss function then corresponds to the inference of w given the data

$$p(w|D,\alpha) = \frac{p(D|w)\,p(w|\alpha)}{p(D|\alpha)}$$

$$= \frac{e^{-G(w)}\,e^{-\alpha E_w(w)}/Z_w(\alpha)}{p(D|\alpha)}$$

$$= \frac{1}{Z_M}\,e^{-M(w)}$$

w^* that minimizes $M(w)$ is the point estimate for the parameters of the neural network.

Instead, we want to employ the Bayesian approach, which involves marginalization

$$p(t^{(N+1)}|x^{(N+1)},D,\alpha) = \int d^K w\; p(t^{(N+1)}|x^{(N+1)},w,\alpha) \times p(w|D,\alpha)$$

for predicting $t^{(N+1)}$ for input $x^{(N+1)}$

W3a

This approach is an example of probabilistic programming which is an important trend in the ML community.

Different kind of uncertainties should be addressed:

- Epistemic uncertainties:
  from uncertainties in the model (e.g the weights used in our NN).
  Can be reduced with more data.

- Aleatoric uncertainty:
  from inherent noise in training data.
  Included in the likelihood function (but will not be reduced with more data of the same quality)