COMP3411/9814: Artificial Intelligence

Week 8 Extension: Variations on Backpropagation

Russell & Norvig: 18.7

Gradient Descent (Backpropagation)

We define an **error function** E to be (half) the sum over all input patterns of the square of the difference between actual output and desired output

$$E = \frac{1}{2}\sum(z-t)^2$$

If we think of E as height, it defines an error **landscape** on the weight space. The aim is to find a set of weights for which E is very low. This is done by moving in the steepest downhill direction.

$$w \leftarrow w - \eta\,\frac{\partial E}{\partial w}$$

Parameter η is called the learning rate.

Variations on Backprop

- Cross Entropy
  - problem: least squares error function unsuitable for classification, where target = 0 or 1
  - mathematical theory: maximum likelihood
  - solution: replace with cross entropy error function
- Weight Decay
  - problem: weights “blow up”, and inhibit further learning
  - mathematical theory: Bayes’ rule
  - solution: add weight decay term to error function
- Momentum
  - problem: weights oscillate in a “rain gutter”
  - solution: weighted average of gradient over time

Cross Entropy

For classification tasks, target t is either 0 or 1, so better to use

$$E = -t\log(z) - (1-t)\log(1-z)$$

This can be justified mathematically, and works well in practice – especially when negative examples vastly outweigh positive ones.
It also makes the backprop computations simpler

$$\frac{\partial E}{\partial z} = \frac{z-t}{z(1-z)}$$

$$\text{if} \quad z = \frac{1}{1+e^{-s}},$$

$$\frac{\partial E}{\partial s} = \frac{\partial E}{\partial z}\frac{\partial z}{\partial s} = z-t$$

Maximum Likelihood

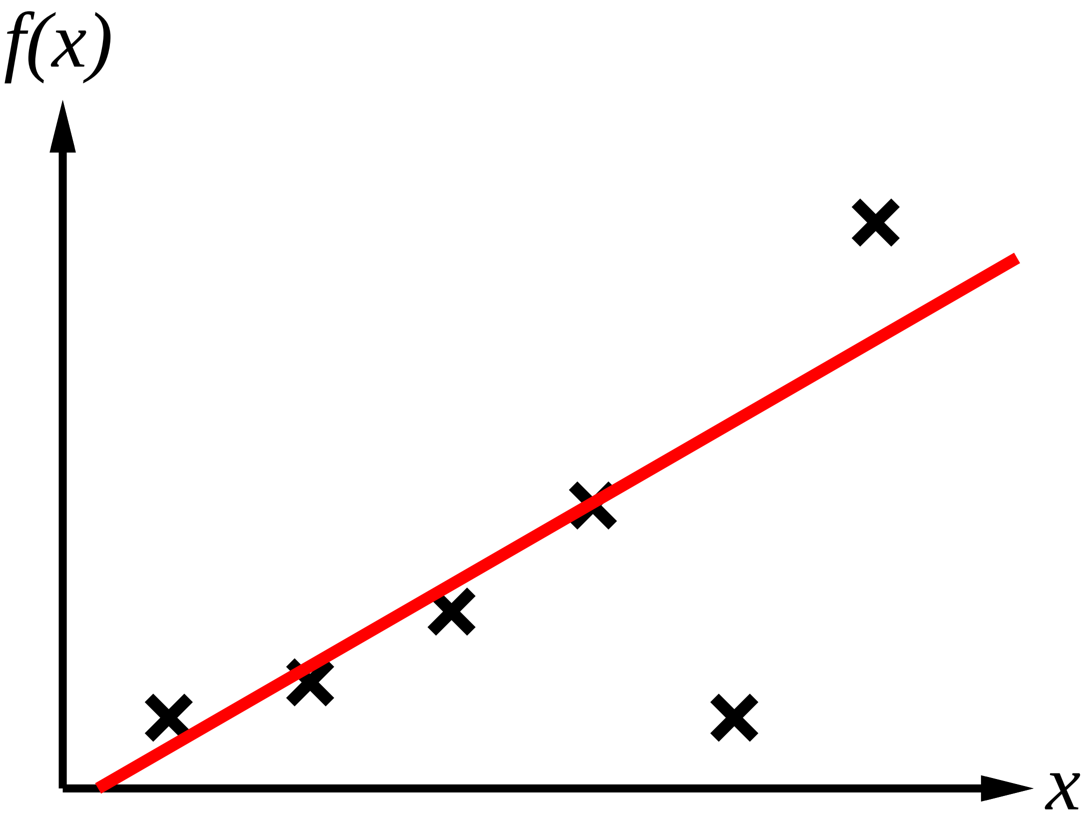
H is a class of hypotheses

$P(D|h) =$ probability of data D being generated under hypothesis $h \in H$.

$\log P(D|h)$ is called the likelihood.

ML Principle: Choose $h \in H$ which maximizes the likelihood,
i.e. maximizes $P(D|h)$ [or, maximizes $\log P(D|h)$]

Least Squares Line Fitting



Derivation of Least Squares

Suppose data generated by a linear function h, plus Gaussian noise with standard deviation σ.

$$
\begin{aligned}
P(D|h) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(d_i - h(x_i))^2} \\
\log P(D|h) &= \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2 - \log(\sigma) - \frac{1}{2}\log(2\pi) \\
h_{ML} &= \text{argmax}_{h\in H} \log P(D|h) \\
&= \text{argmin}_{h\in H} \sum_{i=1}^{m} (d_i - h(x_i))^2
\end{aligned}
$$

(Note: we do not need to know σ)

Derivation of Cross Entropy

For classification tasks, d is either 0 or 1.

Assume D generated by hypothesis h as follows:

$$
\begin{aligned}
P(1|h(x_i)) &= h(x_i) \\
P(0|h(x_i)) &= (1-h(x_i)) \\
\text{i.e.} \quad P(d_i|h(x_i)) &= h(x_i)^{d_i}(1-h(x_i))^{1-d_i}
\end{aligned}
$$

then

$$
\begin{aligned}
\log P(D|h) &= \sum_{i=1}^{m} d_i \log h(x_i) + (1-d_i)\log(1-h(x_i)) \\
h_{ML} &= \text{argmax}_{h\in H} \sum_{i=1}^{m} d_i \log h(x_i) + (1-d_i)\log(1-h(x_i))
\end{aligned}
$$

(Can be generalized to multiple classes.)

Bayes Rule

H is a class of hypotheses

$P(D|h) =$ probability of data D being generated under hypothesis $h \in H$.

$P(h|D) =$ probability that h is correct, given that data D were observed.

Bayes' Theorem:

$$\begin{aligned} P(h|D)P(D) &= P(D|h)P(h) \\ P(h|D) &= \frac{P(D|h)P(h)}{P(D)} \end{aligned}$$

$P(h)$ is called the prior.

Example: Medical Diagnosis

Suppose we have a 98% accurate test for a type of cancer which occurs in 1% of patients. If a patient tests positive, what is the probability that they have the cancer?

Weight Decay

Assume that small weights are more likely to occur than large weights, i.e.

$$P(w) = \frac{1}{Z} e^{-\frac{\lambda}{2}\sum_j w_j^2}$$

where Z is a normalizing constant. Then the cost function becomes:

$$E = \frac{1}{2}\sum_i (z_i - t_i)^2 + \frac{\lambda}{2}\sum_j w_j^2$$

This can prevent the weights from "saturating" to very high values.

Problem: need to determine λ from experience, or empirically.

Momentum

If landscape is shaped like a "rain gutter", weights will tend to oscillate without much improvement.

Solution: add a momentum factor

$$\delta w \leftarrow \alpha\,\delta w + (1-\alpha)\frac{\partial E}{\partial w}$$

$$w \leftarrow w - \eta\,\delta w$$

Hopefully, this will dampen sideways oscillations but amplify downhill motion by $\frac{1}{1-\alpha}$.

Conjugate Gradients

Compute matrix of second derivatives $\frac{\partial^2 E}{\partial w_i \partial w_j}$ (called the Hessian).

Approximate the landscape with a quadratic function (paraboloid).

Jump to the minimum of this quadratic function.

Natural Gradients (Amari, 1995)

Use methods from information geometry to find a "natural" re-scaling of the partial derivatives.