Issues of Backpropagation

Presentation order of training samples

Sequential or random presentation:

在一个 training epoch 中，training examples 可以按固定的顺序（presented in the same sequential order），或随机（presented in a different random order）进行训练。

随机进行训练通常会产生更好的结果。

Initialization

Random initial state:

The network weights are initialized to some choice of random numbers with a range typically between -0.5 and 0.5 (the inputs are usually normalized to numbers between 0 and 1).

即使学习条件相同，随机的初始权重不同也会导致训练的结果不同。

Hidden layer

多增加 hidden 层不会增加识别（discrimination）的表现能力（representational power）。

- 两个 hidden 层的网络更强大，但一个 hidden 层的网络对于实践中遇到的许多任务可能足够准确
- 一个 hidden 层的网络训练起来更快

A heuristic to start with:
One hidden layer, with n hidden neurons,

n=(inputs + output_neurons) / 2

Stopping Criteria

The stopping criteria is checked at the end of each epoch:

- The error (mean absolute or mean square) at the end of an epoch is below a threshold（All training examples are propagated and the error is calculated. The threshold is determined heuristically -e.g. 0.01）
- Maximum number of epochs is reached
- Early stopping using a validation set

Learning rate

在梯度下降时，我们可以有很多选择。这些选择的主要变化是：the learning rate and local minima。选择学习率对于找到真正的 global minimum of the error distance 至关重要。

BP 的学习速度太小将使得训练进展非常缓慢。学习速度过大会训练的更快，但可能只会在相对差的解决方案之间振荡。

在保证可以收敛的情况下，学习率越大越好。

Momentum

之前的梯度都只考虑上一次的数据，如果数据变化很大，那么梯度的变化也会很大，这样不稳定。而 momentum 则类似于“惯性”，它要考虑过去的梯度的影响，从而削弱当前数据的影响。

$$\Delta w(t) = -\eta \frac{\partial E_e}{\partial w(t)} + \alpha \Delta w(t-1)$$

where t is the index of the current weight change.

注：$\Delta w(t)$ 是当前时间 t 的梯度，$\Delta w(t-1)$ 是之前时间（t-1）的梯度。

Momentum term simply makes the following change to the weight update rule, where α is the momentum term:

- If α=0, this is the same as the regular backpropagation, where the weight update is determined purely by the gradient descent
- If α=1, the gradient descent is completely ignored, and the update is based on the 'momentum', previous weight update rule
- Typical value for α is generally between 0.6 and 0.9

$$\Delta w(t) = -\eta \frac{\partial E_e}{\partial w(t)} + \alpha \Delta w(t-1)$$

（1-α）

注：通常地，我们设 $\eta = (1-\alpha)$

Momentum 有以下影响：

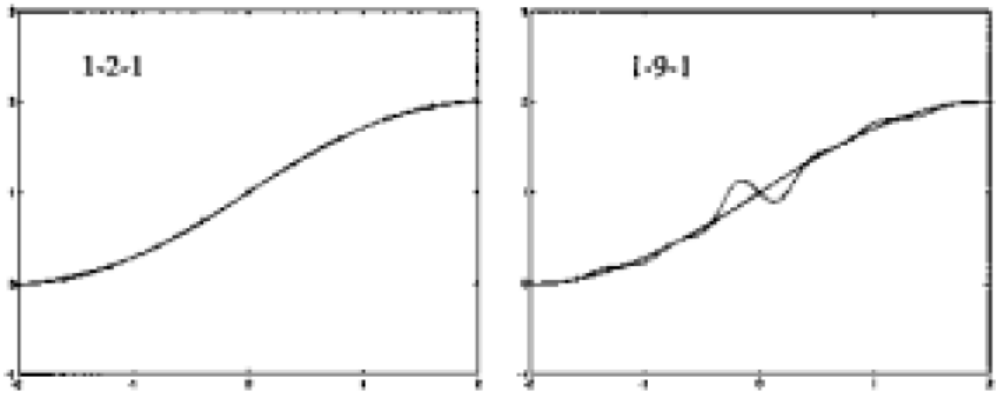
- it smooths the weight changes and suppresses cross-stitching, that is cancels side-to-side oscillations across the error valley（即不让梯度来回震荡）
- 当所有 weight 变化都在同一方向时，momentum 放大学习率，导致更快的收敛
- 能够从 error surface 上的 small local minima 中逃脱（惯性会让梯度冲出去）

Generalization & Overfitting

网络应该能够将其学到的东西泛化（generalize）到所有的样本。而有些时候，模型在训练的时候 error 很小，但在陌生的数据上误差很大：这说明模型只记住了 training example，而没有学会泛化到其他情况（这种又叫 overfitting）。

造成 overfitting 的常见原因：free parameters 的数量大于 training examples 的数量

$f(x) = 1 + \sin\left(\frac{6\pi}{4}x\right)$ was sampled to create 11 training examples



注：1-2-1 代表 1 个 input layer，2 个 hidden layers，1 个 output layer。

Overfitting may be prevented by **early stopping**, **network pruning**（剪枝）, and applying **regularization techniques**.

Techniques to overcome overfitting

Weight decay

Weight decay: Decrease each weight by some small factor during each iteration（权重衰减：每次迭代后都将权重减小。这是为了让权重的值保持 small）.

大权重会以两种不同的方式影响泛化：

- 权重过大可能导致 hidden units 的输出功能过于粗糙，可能具有不连续性
- 如果output units 的激活函数没有限制，那么权重过大可能导致 output units 的输出远远超出合理范围

权重过大的主要风险是：非线性节点输出可能位于 transfer function 的平面部分，而这里导数为零。这会使训练不可逆转地停止。

- Add penalty term to the error function
- ✓ Penalizes large weights to reduce variance
- ✓ Standard weight decay equation

$$MSE_{reg} = MSE + \gamma \cdot MSW$$

penalizes large weights

weight decay penalty term 导致权重收敛到比它们本来要小的绝对值。

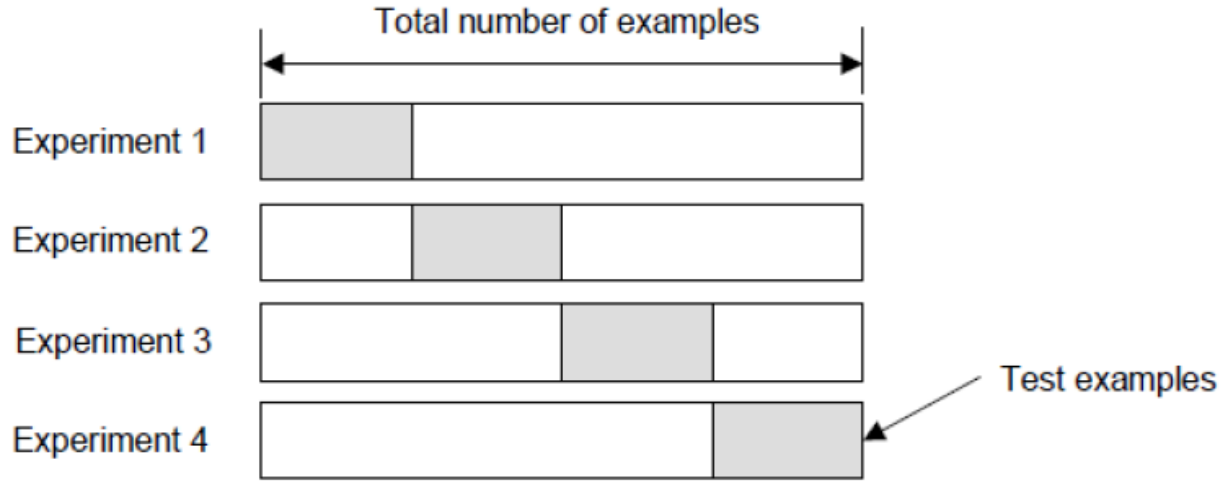
Cross-validation

交叉验证：除了 training data 外，还有一组 validation data。validation data 可以用来验证当前模型的性能，其不参与训练。

K-fold cross validation

对于 cross-validation，一个问题是小数据集可能没有足够的数量来构建 validation dataset。而 overfitting 往往会小数据集造成更大影响。

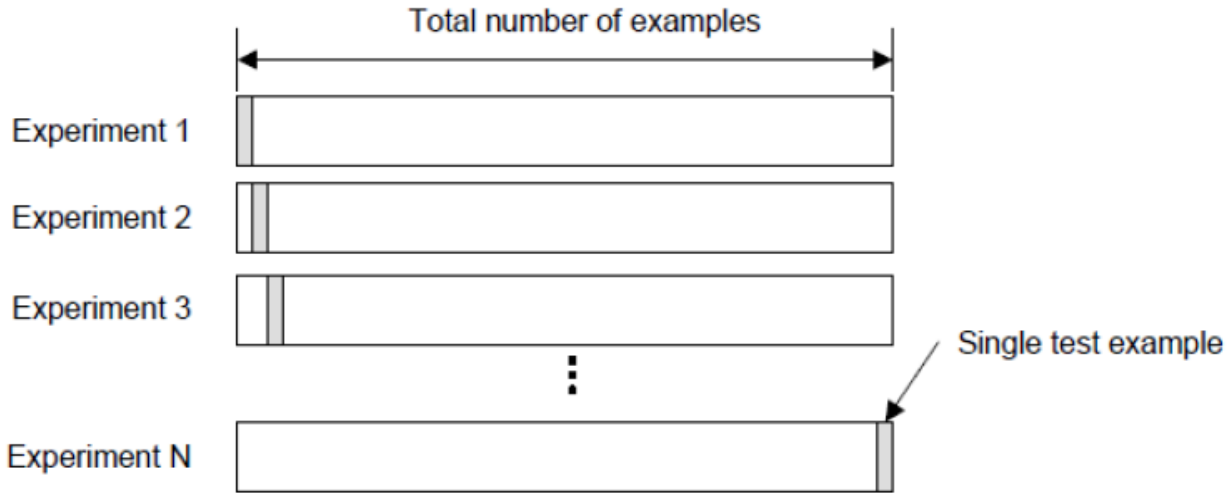
K-fold cross validation：它将原始数据分成 K 组 (K-Fold)，将每个子集数据分别做一次验证集，其余的 K-1 组子集数据作为训练集，这样会得到 K 个模型。这 K 个模型分别在验证集中评估结果，最后的误差 MSE (Mean Squared Error) 加和平均就得到交叉验证误差



Leave-one-out cross-validation

Leave-one-out cross-validation (LOOCV)：使用原始样本中的单个训练集作为验证数据，其余数据用作训练数据。重复这样，将样本中的每个都用作验证数据一次。

这和 K-fold cross validation 很想，不过 K-fold cross validation 要把数据先分成 K 份，再从 K 份中拿；而 LOOCV 直接从原数据中拿。



Limitations & Capabilities of MLP

用 BP 算法来训练的 MLPs 可以进行 function approximation 和 pattern classification。

从理论上讲，它可以：

- 执行任何线性和非线性映射
- 能在任意程度上近似任何 reasonable function
- 克服 perceptron 的局限性

实际上：

- 可能并不总是找到解决方案-会被困在 local minima 中
- 对起始条件敏感（权重初始化）
- 对 hidden layers and neurons 的数量敏感
- 对 learning rate 的值敏感