

Supplementary notes (relevant to Chap 10 and Chap 11)

- For MLPs:
 - For 2-class classification, book uses logistic sigmoid and cross-entropy error function.
 - For $K > 2$ -classes, book uses softmax logistic function and cross-entropy error function.
- Why?
 - Well, you don't have to really. Any sigmoid that is differentiable can be used with backprop. MSE error function also commonly used.

- Why logistic?? (see Chap.5 and Section 10.5)
- For a single-perceptron, assuming 2-classes that are approximately Gaussian distributed, the optimal weights can be calculated analytically.
- Using this result, together with our familiar decision rule (choose whichever class has the greatest output), means that $P(C1|x)$ is given by the value of this sigmoid.
 - $P(C2|x) = 1 - P(C1|x)$.
- So we're getting an estimate of posterior class probabilities, not just a class prediction.
- For more info – see Bishop, C. M. Neural networks for Pattern Recognition, Oxford press, 1995.

- Why cross-entropy?? (see section 10.7)
- Comes from writing down a likelihood function assuming a Bernoulli distribution for class labels
 - Maximize likelihood = minimize negative log-likelihood, which is cross-entropy.
- So we end up with decent posterior class probability estimates as outputs.

- Why softmax?? (10.7)
- For $K > 2$ classes this is just the generalization of the logistic case, but adding normalization across the K outputs, so they all sum to one for any given input.
- Again, see Bishop's book.