1. Which of these statements about Dropout is FALSE:
   a. Dropout simulates an ensemble of network architectures
   b. Dropout helps to prevent overfitting
   c. Dropout encourages redundancy
   d. Dropout encourages the weight values to be small

2. Which type of Autoencoder explicitly forces the hidden features not to change much when the inputs are slightly altered?
   a. Variational Autoencoder
   b. Sparse Autoencoder
   c. Denoising Autoencoder
   d. Contractive Autoencoder

3. The best way to deal with the problem of temporal correlations in Deep Q-Learning is:
   a. Cross Entropy Minimization
   b. Increased Momentum
   c. Experience Replay
   d. Back Propagation Through Time

4. Considering a Singular Value Decomposition X = U S V^T, what are the special properties of matrices U, S and V?
   a. U is orthogonal, V is upper triangular and S is symmetric
   b. U, V are upper triangular, and S is diagonal
   c. U, V are symmetric and S is orthogonal
   d. U, V are unitary and S is diagonal

5. Two common methods for unsupervised pre-training of neural networks are:
   a. Deep Boltzmann Machine and Bayesian Inference
   b. Weight Initialization and Autoencoder
   c. Bayesian Inference and Weight Initialization
   d. Autoencoder and Deep Boltzmann Machine

6. When training on linearly separable data using the Perceptron Learning Rule, what will happen if both the learning rate and the initial weights are scaled up by a large factor?
   a. The data will be learned successfully, but in a larger number of epochs
   b. The data will be learned successfully, in a smaller number of epochs
   c. The data will be learned successfully, in about the same number of epochs
   d. Learning may become unstable and fail to converge

7. Reinforcement Learning is when an agent is:
   a. Present multiple times (over time) with the same examples of inputs and their target outputs
   b. Only presented with the inputs and not target outputs, so it aims to find structure in these inputs

c. Not presented with target outputs, but instead given a reward signal that it aims to maximize
d. Presented once with examples of inputs and their target outputs

8. When using Batch Normalization, in the Testing phrase, the Mean and Variance of the activations at each node are typically:
   a. Pre-computed from the training set
   b. Estimated using running averages
   c. Either of the above
   d. None of the above

9. When comparing a Hopfield Network with a Boltzmann Machine, which statement is FALSE?
   a. The range of activations is {-1,1} for one model and {0,1} for the other
   b. One model is used for retrieval, the other for generation
   c. The formula for the energy function is different for the two models
   d. The updates are deterministic for one model, and stochastic for the other

10. The Context Layer in a Simple Recurrent Network:
    a. is computed from the current input and the previous hidden layer
    b. is comprised of the inputs In a sliding window around the current timestep
    c. is a copy of the hidden layer from the previous timestep
    d. is computed from the current input and the previous output

11. Which statement about word2vec is FALSE?
    a. Representations for the same word at the input and output layers are different
    b. It aims to maximise the log probability of a word, based on the surrounding words
    c. The tanh activation function is used at the hidden nodes
    d. Performance improves if frequent words are samples less often

12. Which of these is NOT a method for dealing with the problem of vanishing or exploding gradients?
    a. Batch Normalization
    b. Rectified Linear Unit
    c. Weight Initialization
    d. Conjugate Gradients

13. For the Generative Adversarial Networks discussed in this course, the game between the Generator and Discriminator:
    a. Is never zero-sum
    b. Can be either zero-sum or not, but the non-zero-sum version produces better images
    c. Is always zero-sum
    d. Can be either zero-sum or not, but the zero-sum version produces better images

14. The principle 'The most likely hypothesis is the simplest one consistent with the data.' is called:
    a. Perceptron Learning
    b. Bayes' Rule
    c. Maximum Likelihood
    d. Ockham's Razor