# CSCI 420-01 – Neural Networks Machine Learning
# Final Exam

## Part II

James He

1. **What were our definitions of a priori knowledge and a posteriori knowledge?**

   Solution:

   (a) Priori Knowledge: A priori knowledge is knowledge that exists independently of experience

   (b) Posteriori Knowledge: A posteriori knowledge is knowledge or "matters of fact" are types that requrie esperience or emperical evidence

2. **What is the basic premise of learning?**

   Solution: Using a set of observations to uncover an underlying process

3. **What is the essence of machine learning?**

   Solution:

   (a) A pattern exists

   (b) We cannot pin it down mathematically

   (c) We have data on it

4. **What are the components of learning?**

   Solution:

   (a) Target Function

   (b) Data

   (c) Learning Algorithm

   (d) Hypothesis Set

   (e) Final Hypothesis

5. **What are the solution components that comprise a learning model?**

   Solution:

(a) The hypothesis set
(b) The learning algorithm

6. **What is Hoeffding's Inequality and why is it important in machine learning?**
   Solution:
   (a) Hoeffding's Inequality: $\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$, $\epsilon > 0$
   (b) Importance:

7. **What is the (simple) mathematical expression for the generalization error?**
   Solution: $|E_{in}(g) - E_{out}(g)| \equiv generalization\ error$

8. **What is the VC dimension and why is it important in machine learning?**
   Solution: VC dimension $d_{vc}(H)$ - the most points $H$ can shatter

9. **What is the VC generalization bound and why is it "the most important mathematical result in the theory of learning"?**
   Solution:
   (a) VC generalization bound: $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$
   (b) Importance: It is the most important mathematical result in the theory of learning because it is the first result that gives a bound on the generalization error of a learning algorithm in terms of the number of training examples and the complexity of the hypothesis set.

10. **What are the two questions of learning?**
    Solution:
    (a) Can we ake sure the $E_{out}(g)$ is close enough to $E_{in}(g)$?
    (b) Can we make $E_{in}(g)$ small enough?

11. **What is overfitting?**
    Solution: Fitting hte data more than is wwarranted

12. **What are two methods we studied to deal with overfitting?**
    Solution: Regularization and Validation

13. **For each paper that you were responsible for reading, concisely describe the cost function and the optimization procedure they used to learn. (2 * 12 = 24 bullets)**
    Solution:

(a) **Visualizing and Understanding Convolutional Networks**

   i. Cost Function: Cross-entropy loss

   ii. Optimization Procedure: Backpropagation with gradient descent

(b) **Mixture Density Networks**

   i. Cost Function: Negative log-likelihood of a Gaussian mixture model

   ii. Optimization Procedure: Expectation-Maximization algorithm

(c) **A New Learning Algorithm for Stochastic Feedforward Neural Networks**

   i. Cost Function: Mean squared error

   ii. Optimization Procedure: Stochastic gradient descent with momentum

(d) **Statistical Language Models Based on Neural Networks (Chap. 1-4)**

   i. Cost Function: Cross-entropy loss

   ii. Optimization Procedure: Backpropagation through time (BPTT)

(e) **Efficient Estimation of Word Representations in Vector Space**

   i. Cost Function: Negative sampling loss

   ii. Optimization Procedure: Stochastic gradient descent

(f) **Attention Is All You Need**

   i. Cost Function: Cross-entropy loss

   ii. Optimization Procedure: Adam optimizer with learning rate scheduling

(g) **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

   i. Cost Function: masked language model loss and next sentence prediction loss

   ii. Optimization Procedure: Adam optimizer

(h) **Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions**

   i. Cost Function: Mean squared error

   ii. Optimization Procedure: Stochastic gradient descent

(i) **Deep Learning Code Fragments for Code Clone Detection**

   i. Cost Function: Binary cross-entropy

   ii. Optimization Procedure: Gradient descent with regularization

(j) **Extracting and Composing Robust Features with Denoising Autoencoders**

   i. Cost Function: Reconstruction error with sparsity constraint

   ii. Optimization Procedure: Stochastic gradient descent

(k) **Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders**

   i. Cost Function: Maximum likelihood estimation

   ii. Optimization Procedure: Adam optimizer

(l) **beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework**

   i. Cost Function: Reconstruction loss combined with KL divergence

   ii. Optimization Procedure: Stochastic gradient descent with annealing

(m) **Generative Adversarial Nets**

i. Cost Function: Minimax loss function

ii. Optimization Procedure: Alternating gradient descent/ascent