**Chapter 1 – Introduction**

1. What 3 components are required to define a learning system?

The 3 components of a learning system are experience (E), well-specified task (T) and using a performance metric (P).

1. How do you make a well-specified task? What needs to be specified?

To make a task well-specified

1. Explain how learning can be seen as search?

Learning can be seen as search because the algorithms search a large space of possible hypotheses to find the one that best fits the observed data.   
The learner searches through the space to locate the hypothesis that is most consistent with the available training examples.

**Chapter 2 – Concept Learning**

1. What is the inductive hypothesis and why is it important?

The inductive hypothesis is a hypothesis that approximates the target function well over a sufficiently large set of training examples such that it will also approximate the target function well over unseen examples. The inductive hypothesis is important because it means we can use known training examples to predict unseen examples.

1. What is the difference between supervised, unsupervised and semi-supervised learning?

Supervised learning is the task where we are trying to predict/classify a dependent variable based on training examples.

Unsupervised learning is the task where the dataset doesn’t have an explicit dependent variable and we are instead trying to find structure or patterns in the data (groups or clusters)

Semi-supervised learning when we fit a model on a dataset that is only partially labelled.

1. Differences between concept learning, classification and regression?

Classification and regression are both examples of supervised learning.

Classification is where the dependent variable is discrete. Additionally, classification is where we want to select a category from a set of categories to group a new observation into based off training examples.

Concept learning is a subclass of classification where we want to classify a binary class label.

Regression is where the dependent variable is numeric/continuous. In regression we want to predict this numeric dependent variable and examine the relationships between variables.

1. What is the hypothesis space of the Candidate elimination algorithm?

The hypothesis space of the candidate elimination algorithm are all the possible pairs of S and G pairs.

1. What is an inductive bias?

The inductive of a learning algorithm is the set of assumptions that the learner uses to predict the labels of unseen examples. Inductive bias is what allows learners to generalize to unseen examples.

1. Inductive biases of the algorithms Find S and Candidate algorithm?

The inductive bias of the candidate elimination algorithm is that the target concept can be represented in the hypothesis space.   
The inductive bias of the find-s algorithm is that the target concept can be represented in the hypothesis space and that all instances are negative instances.

1. What is an unbiased learner?

An unbiased learner is a learning algorithm with no inductive bias. This means it cannot generalize and thus predict unseen examples. It a wrote-learner and therefore not a good learner.

**Chapter 3 – Decision Trees**

1. What does information theory have to do with machine learning?

Compression is the same as machine learning. For example, in lossy compression we generalize details of image to save space. In Decision Tree’s we generalize information from dataset and predict using it. Thus, they are similar. The base of both is information theory.

1. What is the decision tree hypothesis space?

The hypothesis of decision trees is space of all possible decision trees.

1. How does the ID3 algorithm search the hypothesis space?

ID3 searches the hypothesis space by iterating through trees from simple to most complex using the information gain heuristic.

1. What is the inductive bias of a decision tree algorithm?

The inductive bias of a decision tree algorithm are the assumptions that shorter trees are preferred over longer trees and that trees that place high information gain attributes close to the root are preferred over those that do not.

1. What is the difference between a preference bias and a restriction bias?

A preference bias is a bias placed on how the algorithm searches the space not the space itself. For example, ID3 incompletely searches a complete hypothesis space. In comparison, a restriction bias

In comparison, a restriction bias places a bias on the hypothesis space. For example, candidate elimination sets a restriction on the hypothesis representation and thus completely searches an incomplete hypothesis space.

1. Which bias is better preference or restriction bias?

The preference bias is more desirable than restriction bias. This is because if we search the space enough, we will find true hypothesis but if we have a restriction bias (and the true hypothesis is not within the constrained space function is not within the space).

1. What is Ocam’s Razor hypothesis behind Minimum Description Length?

The Occam’s razor hypothesis behind Minimum Description Length states that the shortest hypothesis that describes the data well is the best hypothesis.

1. What is not usually included in the Occam’s razor analysis?

For Occam's razor analysis to be true you had to have the optimal representation language for the data and for the tree.  Most people do not take the time to find the optimal language.

1. Is an algorithm with a larger inductive bias good or worse?

No straight answer. If the bias is leading you in the right direction, a larger bias may be better. However, if the bias is leading you in the wrong direction then you don’t want a stronger bias.

1. What is overfitting?

Overfitting is a phenomenon where the hypothesis that best fits the training data is not the hypothesis that best fit the test data. When overfitting occurs, you will see that the training accuracy is much higher than the test accuracy.

1. What increases overfitting?

Overfitting increases when the number of training instances decreases, number of features (attributes) increases, noise in the data increases or the signal is more complex.

1. What are some approaches to overfitting?

Stopping the tree early, post-pruning using a validation set (post error prune or post rule prone), statistical testing (such as randomisation testing)

1. What is the difference between Sample Error and True Error?

Sample error is the fraction of instances in the sample taken that the learner misclassifies. In comparison true error is the probability the learner will misclassify a single randomly drawn instance from the population.

1. What is Bias and Variance?

Bias is the difference between the learner’s hypothesis and the true target concept. It is the difference between the expected prediction and the actual value. In contrast, variance is the how much the different hypotheses differ from one another.

1. What are the 4 sources of error?

The four sources of error are variations in the train/test data, internal randomness in the algorithm and random classification error.

1. How can we minimize these sources of error?

To minimize the internal randomness and the variation in the training data use cross-validation. To reduce the random classification and variation in the test data use randomization testing.

1. Real cause of overfitting?

The real cause of overfitting is the multiple comparisons problem.

1. How does randomisation testing help overfitting?

Using randomisation testing indicates whether the result of the learner is purely by chance or not. Therefore, it shows whether the learner is actually finding patterns and correlations in the data, informing whether you are overfitting or not.

1. Difference between Type 1 and Type 2 error?

Type 1 error (False Positive) is the error of rejecting a null hypothesis when its actually true whereas Type 2 error (False Negative) is the error of failing to reject a null hypothesis when it is in fact not true.

1. Which error (type 1 or type 2) do we worry about more?

We are more worried about Type 1 error than Type 2 error. Type 1 error is the traditional definition of error and more important as we have drawn the conclusion that the null hypothesis is false when, in fact, it is true.

1. What 2 things must be true for an ensemble to improve results?

Firstly, the individual classifiers must be better than random and secondly individual classifiers must be different enough from one another (diverse) to cover a wide variety of examples.

1. What are the 3 reasons ensembles work?

The 3 reasons can be summarized as statistical, computational and representational.

Statistical: The training data may not be enough to narrow the hypotheses down to a single good hypothesis. In this case if we take a bunch of learners combine their votes the hope is we will get closer to the true hypothesis.

Computational: Finding an optimal single hypothesis that fits the data may take a long time. Instead of just searching for a long time, running a bunch of individual classifiers for a shorter period of time letting each classifier find suboptimal parameters but the hope is when you combine them you should get a more optimal overall set of parameters.

Representational: Hypothesis space may not contain the true hypothesis because of our restrictive representation. However, if we let ensemble run, we may be able to combine the individual classifier results in the ensemble to find an approximate of the target function that is outside of the hypothesis space.

1. What are the 4 methods to create ensembles?

Manipulating training data

Manipulating the input features

Manipulating the output features

Injecting randomness

1. When does bagging ensembles perform better than cross-validated?

Bagging always does better than cross validation (CV). Bagging always does better than CV because some of the datapoints are duplicated causing a warping of the distribution of datapoints. This means the trees are more different from each other than if you use CV only.

1. Why does random forest perform so well?

Random forest performs so well because they perform bagging (bagging on the data) and feature/attribute bagging. The tree’s produced in the ensemble are very diverse thus covering a larger variety of cases.

1. What does the hypothesis space of an ANN look like?

The hypothesis space of an ANN is the set of possible weight matrices corresponding to the network’s dimensions.

1. Difference between GD and SGD?

In GD the error is summed over all examples before updating the weights. In contrast in SGD weights are updated each training example. This is the main difference between SGD and GD.

1. Why don’t ANN’s get stuck in local minima?

In practice networks are so large (have a large number of weights) thus correspond to error surfaces with a very high dimensional space. When GD falls into a local minimum with respect to one weight it won’t fall in the local minima on all the other weights. Hence the more weights (higher dimensions) you can get out of local minima as it only needs one point to jump out of the local minima.

1. What is the inductive bias of backpropagation?

The inductive bias of backpropagation is a smooth interpolation between data points. For example, if there are two positive training instances (with no negatives between them) backpropagation will label the points in between also as positive (smooth interpolation).

1. What is the main advantage of hidden units?

The main advantage of hidden units is they can discover intermediate representations of the data. This is a method to compress the data by making implicit concepts explicit.

1. What makes an ANN overfit? Does it overfit?

ANN’s can overfit. What makes them overfit is having more hidden layers than required, running backpropagation for too many iterations.

1. How important is the architecture and parameter settings?

The parameter and architecture of the network are very important and determine the performance of the network.

1. Determine a good architecture?

Trial and error. Start with most complex structure with the hope that weights in irrelevant parts are low so those parts get ignored.

1. What is the general way GA’s search the space?

GA’s search the hypothesis space by a randomised beam search method to seek maximally fit hypothesis.

1. What is the GA hypothesis space?

The hypothesis space of GA are the sets of possible hypotheses.

1. What is the difference between mutation and crossover?

Crossover takes two parents’ hypotheses from the current population and creates two offspring hypotheses by recombining portions of both parents. In comparison, in mutation proportion of the population are chosen at random and random mutations are performed (for example randomly flip a bit).

1. What is early convergence and how do you stop it?

Is

1. Difference between Lamarkian and Baldwin effect?

**Baldwin**: Train NN using backprop and then evaluate each NN based on their final fitness. But when we cross them over, we use the initial weights they started with. In essence Baldwin is saying this NN has a good fitness/chance of going into the next generation because when trained it got good weights. However, when going to the next generation use initial weights as that knowledge **(knowledge is not passed on)**

**Lamarkian:** Pass on knowledge in their lifetime. Train neural nets then cross them over to generate offspring neural nets. The things those neural nets learnt was passed onto their children. **(knowledge is passed on)**

1. What are the main characteristics of swarm intelligence?

Collect behaviours (global behaviour) that result from local interactions of the individuals with each other or with the environment.

1. What are the properties of swarm intelligence?

1. Many agents follow very simple rules

2. No central control structure

3. Simple local interactions lead to emergence of complex global behaviour

1. Where does swarm intelligence get its power?

Interactive collaboration.

1. ACO Hypothesis space?

The hypothesis space of ACO is the vector of pheromone values.

1. Advantage of ACO over simulated annealing and GA’s?

1. Near optimal solutions

2. When graph changes dynamically, ACO will adapt and continue to work

1. PSO Hypothesis space?

The hypothesis space is the vector of each particles position

1. Why does PSO always converge?

PSO will always converge due to one-way communication (as only best information is passed on), just may not converge to the best solution.

1. One way or Multi-way information sharing?

Case by case scenario.

Multiway sharing will not converge as quickly and therefore is more likely to end in the global minima. But it does take longer to converge. Further, low quality characteristics are still able to propagate through multiple iterations in GA’s vs in PSO only the best information is shared.

One-way sharing will converge quicker and can get stuck in local minima (early convergence).