# Guaranteed questions

1. **What is machine learning (ML)?**

   A set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)

2. **What are the main ML types?**

   The main types are:

   (a) Supervised Learning

   (b) Unsupervised Learning

   (c) Semi-Supervised Learning

   (d) Reinforcement Learning

3. **What ML algorithms have you studied after the midterm exam?**

   (a) K-Means Clustering

   (b) Hierarchical Clustering

   (c) Recommender Systems

   (d) Large Scale and Online Learning

   (e) Ensemble Learning

   (f) k-Nearest Neighbors (kNNs)

   (g) Principle Components Analysis (PCA)

   (h) Recurrent Neural Networks

   (i) Reinforcement Learning

   (j) Autoencoders

   (k) Bayesian Networks

4. **Which is more important to you — model accuracy, or model performance, support your answer with an example? This is a partial answer, you need to provide a simple example and your opinion.**

   The model accuracy, or model performance is based on your opinion supported by a simple example (hint: all answers are correct such as either one or both together based on the example you provide).

5. **What are advantages and disadvantages of the Hidden Markov Model?**

**Advantages**

- HMMs are very powerful modeling tools
- Statisticians are comfortable with the theory behind hidden Markov models
- HMMs can be combined into larger HMMs
- Easy to read the model and make sense of it
- The model itself can help increase understanding

**Disadvantages**

- State independence
- Not good for RNA folding problems
- Over fitting
- Local maximums
- Speed

## L15 K-Means Clustering — Tue Mar 3

1. **List, then define the common clustering algorithms.**

   **K-Means clustering:** partitions data into k distinct clusters based on distance to the centroid of a cluster.

   **Hierarchical clustering:** builds a multilevel hierarchy of clusters by creating a cluster tree.

   **Gaussian mixture models:** models clusters as a mixture of multivariate normal density components.

   **Self-organizing maps:** use neural networks that learn the topology and distribution of the data.

   **Hidden Markov models:** use observed data to recover the sequence of states.

2. **What are the two main steps of the k-means algorithm?**

   (a) Assign

   (b) Optimize (Cost Function)

3. **Write the pseudocode of the k-means algorithm.**

```
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \mu_3, \ldots, \mu_K \in \mathbb{R}^n$
repeat
{
    for i = 1 to m
        $c^{(i)}$ := index (from 1 to $K$) of cluster centroid closest to $x^{(i)}$
    for $k = 1$ to $K$
        $\mu_k$ := average (mean) of points assigned to cluster $k$
}
```

4. **How does the k-means algorithm work?**

   The way k-means algorithm works is as follows:

   (a) Specify number of clusters K.

   (b) Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

   (c) Keep iterating until there is no change to the centroids (i.e., assignment of data points to clusters isn't changing).

      i. Compute the sum of the squared distance between data points and all centroids.

      ii. Assign each data point to the closest cluster (centroid).

      iii. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

5. **List advantages and disadvantages of k-means.**

   **Advantages**

   - Easy to implement
   - With a large number of variables, K-Means may be computationally faster than (than what??)
   - k-Means may produce tighter clusters than hierarchical clustering
   - An instance can change cluster (move to another cluster) when the centroids are recomputed.

   **Disadvantages**

   - Difficult to predict the number of clusters (K-Value)
   - Initial seeds have a strong impact on the final results
   - The order of the data has an impact on the final results
   - Sensitive to scale: rescaling your datasets (normalization or standardization) will completely change results. While this itself is not bad, not realizing that you have to spend extra time on to scaling your data might be bad.

# L16 Hierarchical Clustering — Thu Mar 5

1. **What is cluster analysis?**

   - Cluster: A collection of data objects
     - similar (or related) to one another within the same group
     - dissimilar (or unrelated) to the objects in other groups
   - Cluster analysis (or clustering, data segmentation, . . . )
     - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters

- Unsupervised learning: no predefined classes (i.e., learning by observations vs. learning by examples: supervised)

2. **What are the typical applications of cluster analysis?**

   - As a stand-alone tool to get insight into data distribution.
   - As a preprocessing step for other algorithms.

3. **List, then define the two approaches of hierarchical clustering.**

   - Agglomerative: a bottom-up strategy
     - Initially each data object is in its own (atomic) cluster.
     - Then merge these atomic clusters into larger and larger clusters.
   - Divisive: a top-down strategy
     - Initially, all objects are in one single cluster.
     - Then the cluster is subdivided into smaller and smaller clusters.

4. **List all steps of the hierarchical clustering of agglomerative (bottom-up) approach.**

   **Step 1:** Make each data point a single-point cluster → That forms $N$ clusters

   **Step 2:** Take the two closest data points and make them one cluster → That forms $N-1$ clusters

   **Step 3:** Take **the two closest clusters** and make them one cluster → That forms $N-2$ clusters

   **Step 4:** Repeat Step 3 until there is only one cluster

   **Step 5:** Finish

5. **Define the dendrograms, then illustrate how do dendrograms work with a diagram.**

   A dendrogram is a diagram that shows the hierarchical relationship between objects.

   - A binary tree that shows how clusters are merged/split hierarchically
   - Each node on the tree is a cluster; each leaf node is a singleton cluster

     A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster

6. **List, then define all possible methods of merging the clusters that depend on the distance measures.**

   **Single-link** The distance between two clusters is represented by the distance of the **closest pair of data objects** belonging to different clusters.

   **Complete-link** The distance between two clusters is represented by the distance of the **farthest pair of data objects** belonging to different clusters.

   **Average-link** The distance between two clusters is represented by the average distance of **all pairs of data objects** belonging to different clusters.

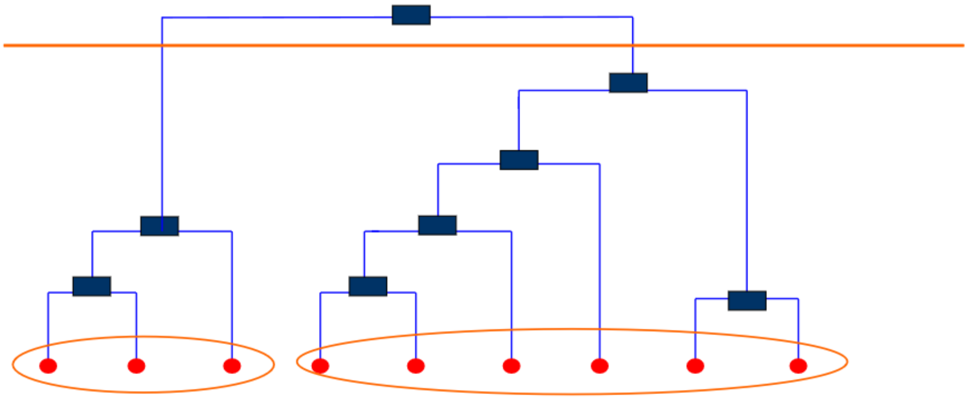

Figure 1: Dendogram

**Centroid distance** The distance between two clusters is represented by **the means of the clusters**.

7. **What are the advantages and disadvantages of hierarchical clustering?**

   **Advantages**

   - Hierarchical clustering outputs a hierarchy, i.e., a structure that is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrograms
   - Easy to implement

   **Disadvantages**

   - It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.
   - Time complexity: not suitable for large datasets
   - Initial seeds have a strong impact on the final results
   - The order of the data has an impact on the final results
   - Very sensitive to outliers

## L17 Recommender Systems — Tue Mar 10

1. **Define the recommendation systems, why use Recommender Systems?**

   Recommendation systems are software agents that elicit the interests and preferences of individual consumers and make recommendations accordingly. They have the potential to support and improve the quality of the decisions consumers make while searching for and selecting products online.

   **Value for the customer**

- Find things that are interesting
- Narrow down the set of choices
- Help to explore the space of options
- Discover new things
- Entertainment

**Value for the provider**

- Additional and probably unique personalized service for the customer
- Increase trust and customer loyalty
- Increase sales, click through rates, conversion etc.
- Opportunities for promotion, persuasion
- Obtain more knowledge about customers

2. **What types of recommendation systems, list them, then draw diagrams show the working mechanism of each?**

   (a) Content-based recommenders — Characteristic information

   (b) Collaborative filtering recommenders — User-item interactions

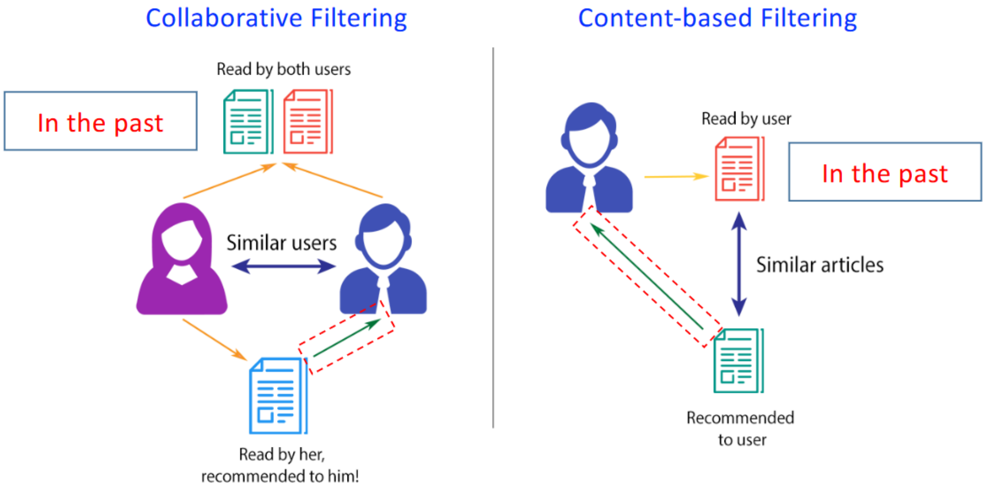

Figure 2: Recommender Types

3. **List advantages and disadvantages of both collaborative filtering and content-based recommenders.**

- **Collaborative filtering recommenders**

  **Advantages**

  - Doesn't require any knowledge about the products themselves

  **Disadvantages**

    - Can't recommend products if you don't have user reviews
    - Difficult to make good recommendations for brand-new users
    - Tends to favor popular products with lots of reviews

- **Content-based recommenders**

  **Advantages**

    - Works even when a product has no user reviews

  **Disadvantages**

    - Needs descriptive data for every product that you want to recommend
    - Difficult to implement for many kinds of large product databases

4. **How to fill rates of users who have not rated any movies?**

   Perform *Mean Normalization*, ?? then assign 0 to each non-ranked movie.??

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 & ? \\ 5 & ? & ? & 0 & ? \\ ? & 4 & 0 & ? & ? \\ 0 & 0 & 5 & 4 & ? \\ 0 & 0 & 5 & 0 & ? \end{bmatrix} \mu = \begin{bmatrix} 2.5 \\ 2.5 \\ 2 \\ 2.25 \\ 1.25 \end{bmatrix} \rightarrow Y = \begin{bmatrix} 2.5 & 2.5 & -2.5 & -2.5 & ? \\ 2.5 & ? & ? & -2.5 & ? \\ ? & 2 & -2 & ? & ? \\ -2.25 & -2.25 & 2.75 & 1.75 & ? \\ -1.25 & -1.25 & 3.75 & -1.25 & ? \end{bmatrix}$$

Figure 3: Mean Normalization

## L18 Large Scale and Online Learning — Thu Mar 12

1. **Supervised Learning, Semi-Supervised, and Unsupervised Learning for what kinds of applications can be used? What is the difference between them in terms of input and output samples?**

2. **What are the differences between Gradient Descent types: Batch, Stochastic, and Mini batch? Which one is the faster to converge?**

   (a) Batch Gradient Descent: Uses all $m$ examples in each iteration

   (b) Stochastic Gradient Descent: Uses 1 example in each iteration

   (c) Mini-batch Gradient Descent: Uses $b$ examples in each iteration

3. **What are the hardware-based solutions can be used to machine learning for big data?**

   **Map-Reduce** Distributing data sets (e.g., training set) across networked computing devices (PCs)

   **Multi-core Machines** Distributing data sets (e.g., training set) across chip-scale computing devices

4. **What are the platforms for online machine learning algorithms?**

   - Hydrosphere.io
   - Prediction.io
   - Azure Machine Learning
   - Amazon Machine Learning
   - Google Prediction
   - BigML
   - DataRobot

## L19 Ensemble Learning — Thu Mar 19

1. **Define *ensemble learning*, illustrate the key motivation of ensemble learning, then draw the general idea diagram of the ensemble learning**

   **Ensemble learning** is a machine-learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine-learning approaches that try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them into one.

   The key motivation is to reduce the error rate. The expectation is that it will become much more unlikely that the ensemble will misclassify an example.

2. **List the ensemble methods that minimize variance and bias.**

   **Minimize Variance**

   - Bagging
   - Random Forests

   **Minimize Bias**

   - Functional Gradient Descent
   - Boosting
   - Ensemble Selection

3. **What are the different methods for changing training data? List them, then illustrate the working mechanism of each method, support your working mechanisms with illustration diagrams.**

   **Bagging:** Resample training data (**see figure 4**)

   **Boosting:** Reweight training data (**see figure 5**)

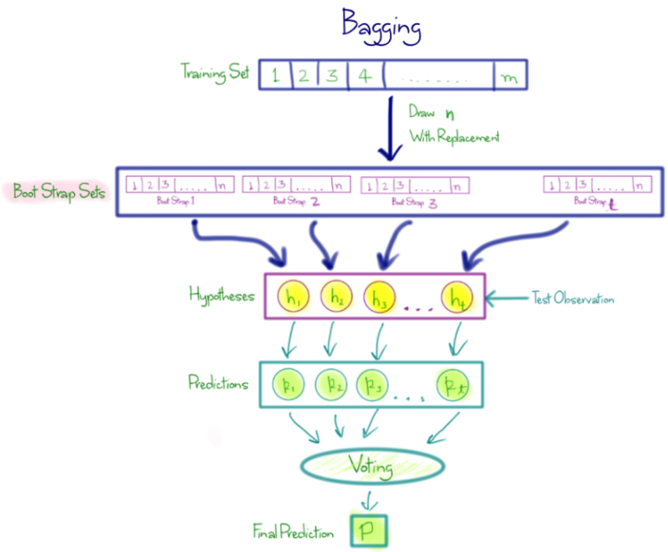

Figure 4: Bagging

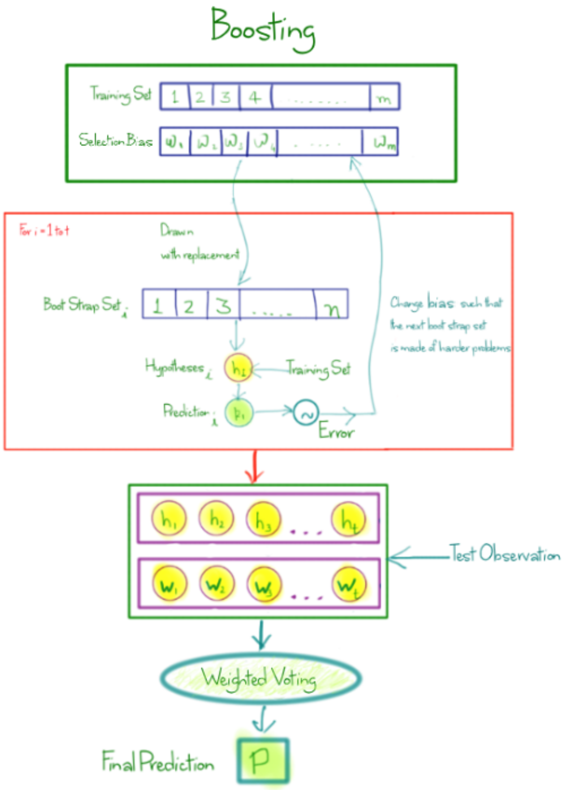

Figure 5: Boosting

4. **Can a set of weak learners create a single strong learner?**

   Yes. You can construct a strong classifier by weighted voting of the weak classifiers.

   The idea is that:

   - Better weak classifier gets a larger weight

- Weak classifiers are added iteratively, which provides increased accuracy through minimization of the cost function.

5. **What are the main features of the Random Forest method?**

   - Random Forest is an ensemble of Decision Trees.
   - They are known to run efficiently on large datasets.
   - Can take care of large number of features.

# L20 k-Nearest Neighbors (kNNs) — Tue Mar 24

1. **What are the Idea, algorithm, and types of the Instance-Based Learning?**

   **Idea**

   - Similar examples have similar labels.
   - Classify new examples like similar training examples.

   **Algorithm**

   - Given some new example $x$ for which we need to predict its class $y$
   - Find most similar training examples
   - Classify $x$ “like” these most similar examples

   **Types**

   **Rote-learner** Memorizes entire training data and performs classification only if attributes of the record match one of the training examples exactly.

   **Nearest Neighbor** Uses $k$ *closest* points (nearest neighbors) for performing classification.

2. **List the k-Nearest Neighbors (k-NNs) Main Steps.**

   (a) For a given instance $T$, get the top $k$ dataset instances that are "nearest" to $T$ (select a reasonable distance measure).

   (b) Inspect the category of these $k$ instances, choose the category $C$ that represent the most instances.

   (c) Conclude that $T$ belongs to category $C$.

3. **What are the three require things to implement the k-NNs?**

   (a) **Feature Space** (Training Data)

   (b) **Distance metric** (to compute distance between instances)

(c) The **value of** $\mathbf{k}$ (the number of nearest neighbors to retrieve from which to get majority class)

4. **How to classify an unknown instance (sample) using the k-NNs?**

   (a) Compute distance to other training instances

   (b) Identify $k$ nearest neighbors (k-NNs)

   (c) Use class labels of nearest neighbors to determine the class label of the unknown instance

5. **What are the two common distance metrics used for k-NNs?**

   **Euclidean Distance (Continuous distribution):** the square root of the sum of the squared differences between a new point (x) and an existing point (y)

   **Manhattan Distance:** the distance between real vectors using the sum of their absolute difference

6. **List Advantages and Disadvantages of k-NNs.**

   **Advantages**

   - Simple technique that is easily implemented
   - Building model is inexpensive
   - Extremely flexible classification scheme — does not involve preprocessing
   - Well suited for
     - Multi-modal classes (classes of multiple forms)
     - Records with multiple class labels
   - Can sometimes be the best method

   **Disadvantages**

   - Classifying unknown records are relatively expensive
     - Requires distance computation of k-nearest neighbors
     - Computationally intensive, especially when the size of the training set grows
   - Accuracy can be severely degraded by the presence of noisy or irrelevant features
   - Nearest-neighbor classification expects class conditional probability to be locally constant

## L21 Principle Components Analysis (PCA) — Thu Mar 26

1. **Define principle components analysis (PCA), then list the 3 main fields could be used to and 3 application examples.**

**Principle Components Analysis (PCA)**

- A technique used to reduce the dimensionality of the data set to 2D or 3D.
- PCA allows us to compute a linear transformation that maps data from a high dimensional space to a lower dimensional sub-space.
- The goal of PCA is to reduce the dimensionality of the data while retaining as much information as possible in the original dataset.

Can be used to:

(a) Reduce number of dimensions in data

(b) Find patterns in high-dimensional data

(c) Visualize data of high dimensionality

Example applications:

(a) Face recognition

(b) Image compression

(c) Gene expression analysis

2. **What do we mean by the variance and covariance? List the differences between the variance and covariance.**

**Variance and Covariance:** Measure of the *spread* of a set of points around their center of mass (mean)

**Variance:** Measure of the deviation from the *mean* for points in *one dimension*

**Covariance:** Measure of how much each of the dimensions vary from the mean with *respect to each other*

- Covariance is measured between two dimensions
- Covariance sees if there is a relation between two dimensions
- Covariance between one dimension is the variance

3. **Illustrate the main tasks of the PCA Process — step 1.**

- Subtract the *mean* from each of the data dimensions.
- All the $x$ values have $x$ subtracted and $y$ values have $y$ subtracted from them.
- This produces a dataset whose mean is zero.
- Subtracting the mean makes variance and covariance calculation easier by simplifying their equations.
- The variance and co-variance values are not affected by the mean value.

4. **How we could derive new datasets through the PCA Process — step 5?**

(a) Final Data = *Row Feature Vector* × *Row Zero Mean Data*

(b) *Row Feature Vector* is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top.

(c) *Row Zero Mean Data* is the mean-adjusted data transposed; i.e., the data items are in each column, with each row holding a separate dimension.

## L22 Recurrent Neural Networks — Tue Apr 7

1. **Define RNNs.**

   Neural nets that allow previous outputs to be used as inputs while having hidden states.

2. **Are RNNs Supervised or Unsupervised Learning?**

   Supervised: used for time series analysis.

3. **What is the major difference between RNNs and FNNs? illustrate that.**

   In feed-forward neural networks, the connections between nodes do not form a cycle. However, in recurrent neural networks, the connections between nodes form cyclic paths.

4. **List types AND architectures of RNNs, then draw the architecture of traditional RNNs.**

   (a) One-to one

   (b) One-to-many

   (c) Many-to-one

   (d) Many-to-many

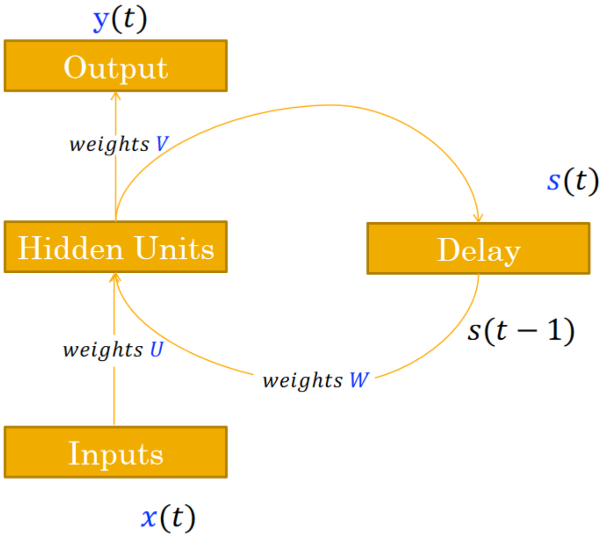

Figure 6: Architecture of a traditional RNN

5. **List, then illustrate the three main training approaches of RNNs.**

**Back-propagation through-time:** Unfolding RNNs in time and using the extended version of back-propagation.

**Extended Kalman Filtering (EKF):** A set of mathematical equations that provides efficient computational means to estimate the state of a process, in a way that minimizes the mean of squared error (cost function) on a linear system.

**Real-Time Recurrent Learning (RTRL):** Computing the error gradient and update weights for each time step.

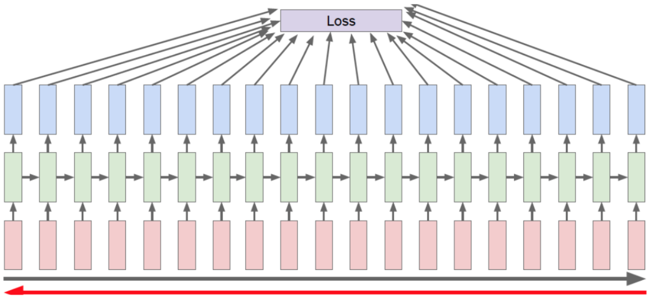

Figure 7: Back-propagation through time

6. **What are the pros and cons of the typical RNNs architecture?**

**Advantages**

(a) Possibility of processing input of any length

(b) Model size $s$ not increasing with size of input

(c) Computation considers historical information

(d) Weights are shared across time

**Disadvantages**

(a) Computation being slow

(b) Difficulty of accessing information from a long time ago

(c) Cannot consider any future input for the current state

## L23 Reinforcement Learning — Thu Apr 9

1. **List the four main machine learning types.**

2. **Define the reinforcement learning with a diagram, then compare between the reinforcement learning and supervised learning.**

3. **Draw the generic learning model to learn from data. Then define the main operations of it through indicating each operation (i.e. Sensor Data, Feature Extraction, etc.) and related steps.**

4. **What are the key features and elements of the reinforcement learning?**

5. **List the 3 types of reinforcement learning.**

6. **What makes reinforcement learning different from other machine learning paradigms?**

## L24 Autoencoders — Tue Apr 14

1. **What are autoencoders? List the general types of autoencoders based on size of hidden layer?**

2. **What are the main differences between PCA and autoencoders?**

3. **List the key elements AND components of autoencoders? Then illustrate the components.**

4. **List, then explain the 3 main properties AND 4 hyperparameters of autoencoders.**

5. **List the 8 types AND 5 applications of autoencoders.**

## L25 Bayesian Networks — Thu Apr 16

1. **What are Bayesian networks (BNs)? List BN components and importance.**

2. **List types of probabilistic relationships, then provide 7 real-world Bayesian network applications.**

3. **Define hidden Markov model (HMM), then list and illustrate components of HMM.**

   A Hidden Markov Model (HMM) is a sequence of random variables, $Z_1, Z_2, \ldots, Z_t, \ldots$ such that the distribution of $Z_t$ depends only on the(hidden) state $x_t$ of an associated Markov chain.

   A Hidden Markov Model (HMM) is composed of the following:

   - $\mathcal{X}$ : a finite set of states.
   - $\mathcal{Z}$ : a finite set of observations.
   - $T : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$, i.e., transition probabilities.
   - $M : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}_+$, i.e., observation probabilities.
   - $\pi : \mathcal{X} \rightarrow \mathbb{R}_+$, i.e., prior probability distribution on the initial state.

4. **List, with illustration, the 4 main inference algorithms of Hidden Markov Model.**

5. **What are advantages and disadvantages of Hidden Markov Model?**

### Advantages

- HMMs are very powerful modeling tools
- Statisticians are comfortable with the theory behind hidden Markov models
- HMMs can be combined into larger HMMs
- Easy to read the model and make sense of it
- The model itself can help increase understanding

### Disadvantages

- State independence
- Not good for RNA folding problems
- Over fitting
- Local maximums
- Speed