Question 1.
Following your graduation, you are hired by a polling organization as a data analyst. As social media has exploded and transformed the way people interact with each other, it would be a great idea to use messages collected from social media to predict the next US presidency. List three challenges to solving this problem. With reference to existing approaches, describe the design of your system.

Question 2.
Ensemble methods have been very successful in building classifiers. The hot topics include how to create diverse classifiers and how to fuse the decisions from individual classifiers, in particular how to establish the weights that individual classifiers contribute to the ensemble's answer. Describe two existing approaches to solving this problem, discuss their advantages and disadvantages. Make a plan to address one issue or two (related to learning the weights or creating diverse classifiers), briefly describe your new method. Explain the reason why the developed method could outperform the conventional ones.

Question 3.
Marketing or advertising companies would be very interested in being able to predict whether a Twitter message will spread as a meme or not, and even better, construct it so that it will spread. Why is this a hard problem to solve? Describe two approaches using data analytics to predict whether a tweet will go viral or not. How would you validate these approaches?

Question 4.
One of the themes in the machine learning models we've looked at this semester is large numbers of parameters that are changed by tiny amounts. Why do so many apparently different models use such similar techniques? Are there other ways to approach the problem of learning? Is this the best way? Are there also commonalities in the way the amounts to be changed are determined?

Question 5.
Consider if you are in front of a gambling machine. The machine has n arms, pulling each will yield a random amount of reward. The average reward yielded by each arm in long-run is a fixed certain value, but the money you receive in individual rounds is random. E.g. you can ***expect*** pulling arm-2 will produce a return of r2, but the actual returns are random values. The expected return of each arm is unknown – you know there is a fixed value, but not knowing what the value is. The task is to i) design a strategy to earn reward as fast as possible ("fast" is defined in terms of the number you pulling the arms); ii) identify the main challenge in designing such strategies; iii) discuss the up bound of the performance of the optimal policy.

Question 6.
One of the ten challenging problems in data mining research is the imbalance learning problem which has been widely reported in various real-world applications. The class-imbalance learning problem occurs when highly unequal distribution of data among different classes exists in a learning task. The majority class having a relatively large number of data can overwhelm the underlying data distribution, such that the minority class having

a relatively small number of data is underrepresented, leading to difficulty in machine learning algorithms to accurately learn the minority class concepts. List at least two real-world applications facing the imbalance learning problem. Describe two existing algorithms to solving this problem. Discuss their strength and weakness, try to develop an improvement to address one weak point or two. Explain the reason why the developed method could outperform the conventional ones.

Question 7.

Several dimensionality reduction methods such as Principal Component Analysis and Linear Discriminate Analysis are suggested to be used before Modelling to solve 'curse of dimensionality' problem. To further address a nonlinear problem, one of the approaches is to leverage the idea of kernel method to embed the data into a higher-dimensional space. Describe two existing kernel-based dimensionality reduction methods. We already meet the challenge in analysing high dimensional data. Why do we need a higher-dimensional space? How this makes linear separation possible?

Choose one question to answer. Good luck!