Question 1. Explain the principle of the gradient descent algorithm. Accompany your explanation with a diagram and pseudo-code.

Question 2. Derive the gradient descent training rule for linear regression assuming that the hypothesis is $\theta_0 + \theta_1 x_1 + \theta_2 x_2$ and the training data consists of $m$ points.

Question 3. We observe data $(x^{(i)}, y^{(i)}), i = 1, 2, \ldots, n$ from $n$ people, where $x^{(i)}$ is the person’s height and $y^{(i)}$ is the person’s weight.

1. Explain how to construct a linear regression model for this data.
2. Suppose we suspect that the weight of a person is not linearly related to their height but rather is related to the square root of their height. Explain how to modify the linear regression model to accommodate this.

Question 4. What is overfitting? Give an example to illustrate. What is underfitting? Give an example.

Question 5. In ridge regression the hypothesis and cost function are

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n, J(\theta) = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{1}{\lambda} \sum_{j=1}^{n} \theta_j^2$$

1. Explain how use of larger or smaller values of $\lambda$ affects the parameters $\theta$. How can an appropriate choice of $\lambda$ help with overfitting?
2. Describe $k$-fold cross-validation
3. Explain how it can be used to select $\lambda$. 
