1. (a) Discount factor is very close to 1, so it makes the agent strive for a long term high rewards. Also, learning rate is low, so the result shows that agent doesn't consider recent information.

   (b) I observed that there is no much difference because

   (c) When I increased value of epsilon a lot, I observed the difference because agent selects different action. At every decision get a random number. It is more exploratory policy.

2. (a) We don't know thief but the performance is different from problem 1. The packages can be stolen.

   (b) Now, we know thief, so the performance takes long and very slow. However, agent can choose optimal action.

   (c) Best learning rate is 0.5 and discount factor is 0.9. Expected discount reward is much higher than using other value of learning rate and discount factor.

   3. The result is not much different because the result is very straightforward with best learning rate and discount factor.