1 Fall 2017: Problem 7

7. a) i. One-hot encoding
Explanation: We have categorical data that has no numerical interpretation, so it makes the most sense to use a one-hot encoding.

ii. Divide by 50
Explanation: We would like our features to have approximately the same magnitude, so we divide by 50.

iii. Divide by 1 billion
Explanation: We would like our features to have approximately the same magnitude, so we divide by 1,000,000.

iv. Omit
Explanation:The company name is unlikely to be indicative of how the stock will perform in the future, so we omit it.

b) i. α) 1
β) sigmoid
γ) NLL
Also okay: 1 unit, linear, hinge or 2 units, softmax, NLL
Explanation: We can try to predict the probability that the company will have an IPO (and use a threshold probability of 0.5 to decide which classification to make), which can be done with a single sigmoid unit and NLL loss.
Other valid solutions are to use a single linearly activated unit with hinge loss (which is effectively SVM to perform the classification into the two classes) or two units with softmax activation and NLL loss (and choosing the larger of the two probabilities for our classification).

ii. α) 1
β) linear
γ) squared-error
Explanation: We would like to predict a single numerical value that spans the real numbers, so we will use a linear activation and squared error.

iii. α) 100
β) sigmoid
γ) NLL (individually)
Explanation: Here, we are asked to perform 100 independent 2-class classification problems. So, we can have 100 separate sigmoid activated units each with their own NLL loss (which we sum to get the total loss). Each unit is responsible for performing the prediction for one specific client.
We can also adapt the other solutions from (a), just using 100 copies of whatever approach we choose.