Exercise 2

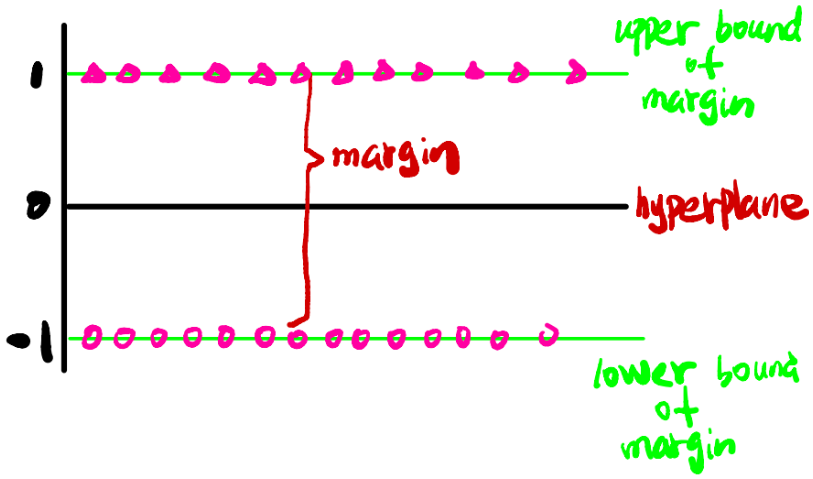
1)

Logistic regression does not work. The error message is PerfectSeparationError: Perfect separation detected, results not available. This happens when all or nearly all of the values in one of the predictor categories are associated with only one of the binary outcome values. Under this condition, the problem occurs because the program runs into numerical issues when computing exp(p/1-p). Mathematically, for minimizing the negative log-likelihood (the loss function), the parameter vector does not converge because there is no minimum for the loss function under this condition. To deal with this issue, we need manually set a maximum and minimum for the coefficients which cannot converge. To deal with numerical issue, we can change exp(p/1-p) by exp((p/1-p) − t/2), where t = max(1/1-p) – min(1/1-p).

The solution of soft-margin and hard-margin SVM on dataset A are the same. This illustrates that Dataset A is linearly separable and there is no points drop within the margin that makes hard-margin SVM move away from the optimal hyperplane.

2)

Number of values that smaller or equal to one is 2000, which means that all points drop on boundary of the margin. Just looks like caricature below.



Note that not all points on the boundary of margin belongs to support vector. Only resulted alpha(solved from Lagrange dual problem) larger than 0 indicates points are support vector. Theoretically, two support vector is enough to find the hyperplane.

For dataset A, there are 2 support vector and can write a linear combination of solution parameter vector as:

$$W = 0.5X_{1998} - 0.5X_{1999}$$

For dataset B, the Hard-margin SVM does not stop running because it cannot converge for this dataset. The reason for this is that the dataset B is not linearly separable. To calculate the solution parameter vector, I used 192 points as support vector. Firstly, I find dual_coef_ and support_vectors_ and then multiply them together to get solution parameter vector. The empirical prediction accuracy of the soft-margin SVM on the test set is 97.15% and that of the logistic regression is 96.90%.