# Homework1

1. Basics:

* Question A: Hypothesis set is also called model class which will be used to validate and test the data; The hypothesis of a linear model is expressed as:
* Question B: Overfitting occurs when a training set is too complex that has too many parameters related to the number of observation. As a result, the model class (w, b) derived from it cannot be applied to the whole data set which is really large compared with the former set.
* Question C: The most common way to prevent overfitting is using techniques such as cross-validation, model selection so that the model can fully learn the training set rather than just “memorize” it.
* Question D:

1. Training data is the set of examples you use to learn. Each input in this set has a specified output.
2. Validation data is what you use to do the iteration optimization until you find the “optimal” model.
3. Test data is used to calculate the test error to demonstrate the quality of the final model (after the selection process (ii)).

* Question E: The fundamental training data sample assumption is that the training data is neither overfit nor underfit, it properly represents the large scale data set we would like to predict.
* Question F: In spam email example:

1. The input space X, is the word vector (dictionary) of each email;
2. The output space Y, is the classification result whether this email is a spam or not.
3. The target function X->Y, is the classifier to determine whether the email is a spam or not.
4. Bias Variance Tradeoff
5. The Perceptron

* Question A: To include the bias term, we can include an additional element in and separately: