Homework1

name: Leiya Ma, UID: 2018042, email: lma2@caltech.edu

1. Basics:

- **Question A:** Hypothesis set is also called model class which the model we come up with to represent the whole data set; The hypothesis of a linear model is expressed as:

$$f(x|w,b) = w^T x - b$$

- **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:**

  (i) Training data is the set of examples you use to learn. Each input in this set has a specified output.

  (ii) Validation data is what you use to do the iteration optimization until you find the "optimal" model.

  (iii) 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:

  (i) The input space X, is the word vector (dictionary) of each email;

  (ii) The output space Y, is the classification result whether this email is a spam or not.

  (iii) The target function X->Y, is the classifier to determine whether the email is a spam or not.

2. Bias Variance Tradeoff

The simulation is achieved in Matlab and the related .mat file is attached in homework file named problem2.mat.

- **Question A:** the expected values of y1 and y2 over all pairs in *D* are : $E_D(y_1) = -0.0084, E_D(y_2) = 0.0037$

  The theoretical value of these value can be represented as:

$$\int_{-1}^{1} sin(\pi x)\frac{1}{2}dx = 0$$

- **Question B:** for each pair $D^{(i)}$, the parameters for the hypotheses $g^{(i)}(x)$, $h^{(i)}(x)$ are calculated and reported in problem2.mat. eg, mean_eg are parameters for $g^{(i)}(x)$; eh, mean_eh are parameters for $h^{(i)}(x)$:

$$g^{(i)}(x) = b$$

$$h^{(i)}(x) = ax + c$$

  The expected value of $g^{(i)}(x)$ can be represented as:

$$E_{out}[g^{(i)}] = \int_{-1}^{1} \left(b^{(i)} - sin(\pi x)\right)^2 \frac{1}{2}dx = \left(b^{(i)}\right)^2 + \frac{1}{2}$$

  The expected value of $h^{(i)}(x)$ can be represented as:

$$E_{out}[h^{(i)}] = \int_{-1}^{1} \left(a^{(i)}x + c^{(i)} - sin(\pi x)\right)^2 \frac{1}{2}dx = \frac{\left(a^{(i)}\right)^2}{3} + c^2$$

  The expected value of $E_{out}[g^{(i)}]$ and $E_{out}[h^{(i)}]$ over *D* for the two hypotheses are:

$$E_{D^{(i)}\sim D}[E_{out}(g^{(i)})] = 0.7535$$

$$E_{D^{(i)}\sim D}[E_{out}(h^{(i)})] = 1.8971$$

- **Question C:** As the average function $\bar{g}(x)$ can be represented as:

$$\bar{g}(x) = \frac{1}{|D|}\sum_{D} g^{(i)}(x)$$

  the parameters are calculated and reported in problem2.mat too, where mean_b, mean_a and mean_c are parameters related to this question:

$$\bar{g}(x) = mean\_b, \bar{h}(x) = mean\_a \cdot x + mean\_c$$

$$mean\_b = \frac{1}{|D|}\sum_{D} b^{(i)} = -0.0024$$

$$mean\_a = \frac{1}{|D|}\sum_{D} a^{(i)} = 0.7531$$

$$mean\_c = \frac{1}{|D|}\sum_{D} c^{(i)} = 0.0100$$

- **Question D:**

  (i) For hypotheses class g, the expected bias calculated is:

$$\mathrm{E}_x[Bias(x)] = E_x\left[\left(\bar{g}(x) - f(x)\right)^2\right] = bias\_g = 0.5000$$

$$\mathrm{E}_x[Var(x)] = E_x\left[E_D\left[\left(g^{(i)}(x) - \bar{g}(x)\right)^2\right]\right] = var\_g = 0.2535$$

where,

$$\mathrm{E}_x[Bias(x)] + \mathrm{E}_x[Var(x)] = 0.7535$$

equals what we have calculated in Question B:

$$E_{D^{(i)}\sim D}\left[E_{out}\left(g^{(i)}\right)\right] = 0.7535$$

  (ii) For hypotheses class h, the expected bias calculated is:

$$\mathrm{E}_x[Bias(x)] = E_x\left[\left(\bar{h}(x) - f(x)\right)^2\right] = bias\_h = 0.1891$$

$$\mathrm{E}_x[Var(x)] = E_x\left[E_D\left[\left(h^{(i)}(x) - \bar{h}(x)\right)^2\right]\right] = var\_h = 1.7080$$

where,

$$\mathrm{E}_x[Bias(x)] + \mathrm{E}_x[Var(x)] = 1.8971$$

equals what we have calculated in Question B:

$$E_{D^{(i)}\sim D}\left[E_{out}\left(h^{(i)}\right)\right] = 1.8971$$

- Here we prove that $E_{out} = bias + var$ exists, using $v(x)$ to represent both $\mathrm{g}(x)$ and $\mathrm{h}(x)$

$$E_D\left[E_{out}(v(x))\right] = E_D\left[E_x\left[\left(v(x) - f(x)\right)^2\right]\right]$$

  where

$$E_x\left[\left(v(x) - f(x)\right)^2\right] = E_x\left[\left(v(x) - \bar{v}(x) + \bar{v}(x) - f(x)\right)^2\right]$$

$$= E_x\left[\left(v(x) - \bar{v}(x)\right)^2\right] + E_x\left[\left(\bar{v}(x) - f(x)\right)^2\right] + 2E_x\left[\left(v(x) - \bar{v}(x)\right)\right]E_x\left[\left(\bar{v}(x) - f(x)\right)\right]$$

  where the red part equals 0 when calculating the expected number over *D* and thus can be eliminated.

  Therefore,

$$E_D\left[E_x\left[\left(v(x) - f(x)\right)^2\right]\right] = E_D\left[E_x\left[\left(v(x) - \bar{v}(x)\right)^2\right]\right] + E_D\left[E_x\left[\left(\bar{v}(x) - f(x)\right)^2\right]\right]$$

$$= bias + var$$

3. The Perceptron

- **Question A**: To include the bias term, we can include an additional element in $\vec{x}$ and $\vec{w}$ separately:

$$\vec{x} = [1, x_1, x_2, \cdots, x_n]$$
$$\vec{w} = [b, w_1, w_2, \cdots, w_n]$$

- Question B:

t	b	w1	w2	x1	x2	y
0	0	0	1	1	-2	+1
1	1	1	-1	0	3	+1
2	2	1	2	1	-2	+1
3	3	2	0			

Since now every point is in the correctly classified, we end the iteration at b=3, w1=2, w2=0;

- Question C: In a 2D data set, the number of the data points in the smallest data set that is not linearly separable is 3. If the 3 points are +, -, +, and they are in the same line, then they cannot be linearly separable. Similarly, in n-dimensional set, the number is also 3.
- Question D: The PLA algorithm will not coverage. Because the perceptron is a linear classifier, therefore it will never get to the state with all the input vectors classified correctly if the training set D is not linearly separable.

4. Gradient Descent

- Question A: The gradient of the squared loss function for linear regression is:

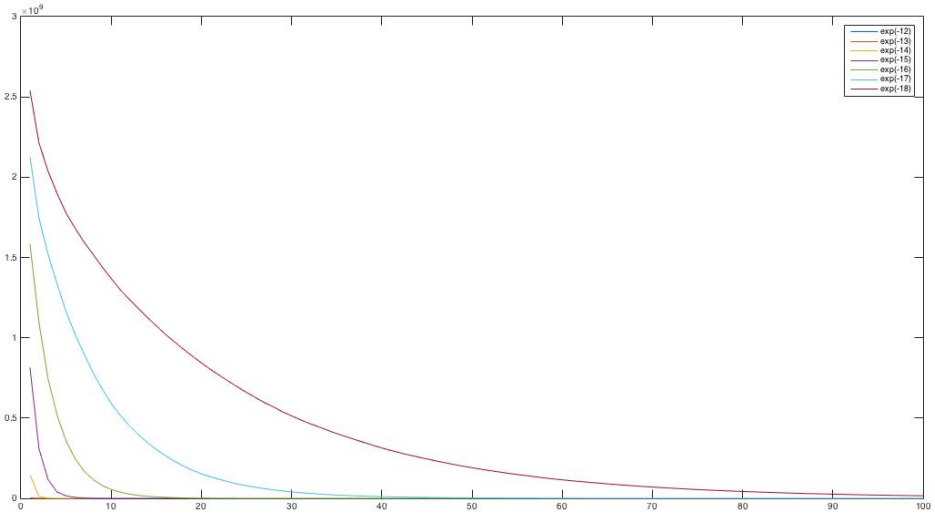
$$\partial L = \partial \sum_{i=1}^{n} (y_i - \boldsymbol{w}^T x_i)^2 = \sum_{i=1}^{n} 2(y_i - \boldsymbol{w}^T x_i)\,(-x_i)$$

- Question B: the function is written in Matlab. The file name is func.m, and the some of the parameters in this question and the next one are recorded in result4.mat

  The result is:

$$\mathrm{w} = [-5.9805, 3.9696, -11.9214, 8.9287]$$

- Question C: the following diagram is showing the error of loss function in different step size.



- Question D: When using the closed form solution for this problem, the result is $\mathrm{w} = (x^T x)^{-1}(x^T y) = [-5.9993, 3.9998, -12.0012, 9.0005]$, which matches up what got from Question B.

- Question E: The first reason is in most of the linear regression problem, there is no closed form solution; The second reason is even in one of the few cases where closed form exists, it may be impractical to use this formula as this solution is extremely expensive to use sometimes and needs large computation space.

- Question F: When SGD is used on a non-convex learning problem, it may still able to find the local optimal point. But it is not guaranteed to be the overall optimal point among all the data.