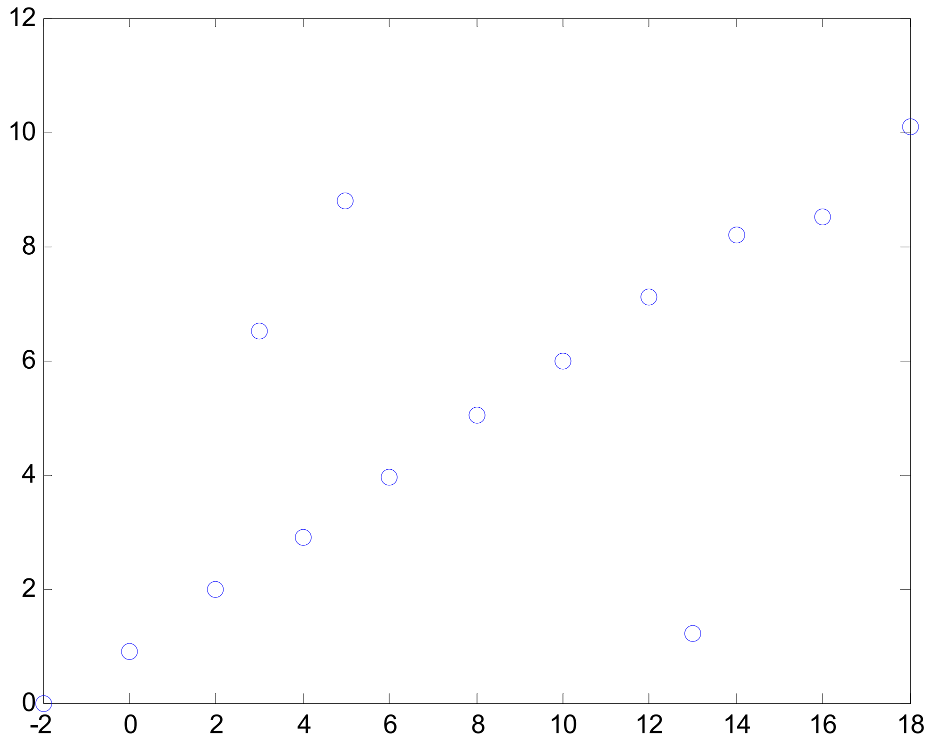
Assignment 2 (Due: May 20, 2018)

1. (**Programming**) RANSAC is widely used in fitting models from sample points with outliers. Please implement a program to fit a straight 2D line using RANSAC from the following sample points:
(-2, 0), (0, 0.9), (2, 2.0), (3, 6.5), (4, 2.9), (5, 8.8), (6, 3.95), (8, 5.03), (10, 5.97), (12, 7.1), (13, 1.2), (14, 8.2), (16, 8.5) (18, 10.1). Please show your result graphically.

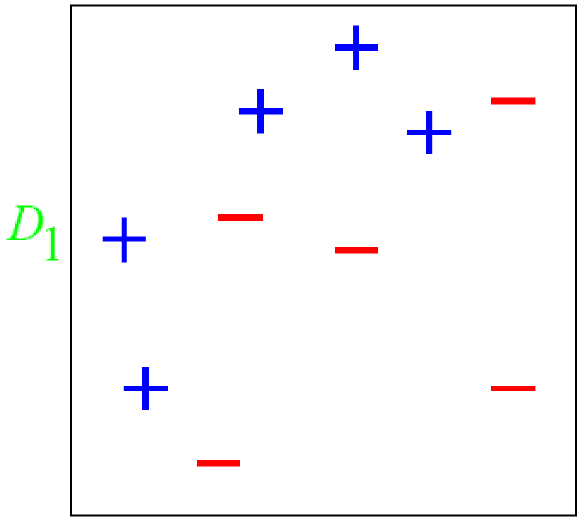


2. (**Programming**) AdaBoost is a powerful classification tool, with which a strong classifier can be learned by composing a set of weak classifiers. In our lecture, we use a vivid example to demonstrate the basic idea of AdaBoost. Now, your task is to implement this demo.

   Training:
   There are 10 samples on a 2-D image plane and information of the ith sample is given as (x_i, y_i, l_i), where (x_i, y_i) is its coordinate and l_i is its label. 10 samples are (80, 144, +1), (93, 232, +1), (136, 275, -1), (147, 131, -1), (159, 69, +1), (214, 31, +1), (214, 152, -1), (257, 83, +1), (307, 62, -1), (307, 231, -1). Weak classifiers are vertical or horizontal lines as described in our lecture. The final trained strong classifier actually is a function having the form,

   Label = strongClassifier(x, y)

Finally, test your resultant strong classifier to verify whether it can correctly classify all the training samples.



3. (**Math**) There are n p-dimensional data points and we can stack them into a data matrix, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \in R^{p\times 1}, \mathbf{X} \in R^{p\times n}$

   The covariance matrix of $\mathbf{X}$ is $C = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$, where $\mu = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ (actually, it is the mean of the data points)

   Based on discussions in our lecture, we know that if α_1 is the eigen-vector associated with the largest eigen-value of *C*, the data projections along α_1 will have the largest variance.

   Now let's consider such an orientation α_2. It is orthogonal to α_1; and among all the orientations orthogonal to α_1, the variance of data projections to α_2 is the largest one.
   Please prove that: α_2 actually is the eigen-vector of *C* associated to *C*'s second largest eigen-value. (we can assume that α_2 is a unit-vector)

4. (**Math**) In our lecture, we mentioned that for logistic regression, the cost function is,

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{m} y_i \log\left(h_\theta(\boldsymbol{x}_i)\right) + (1 - y_i)\log\left(1 - h_\theta(\boldsymbol{x}_i)\right)$$

   Please verify that the gradient of this cost function is

$$\nabla_\theta J(\boldsymbol{\theta}) = \sum_{i=1}^{m} \boldsymbol{x}_i \left(h_\theta(\boldsymbol{x}_i) - y_i\right)$$