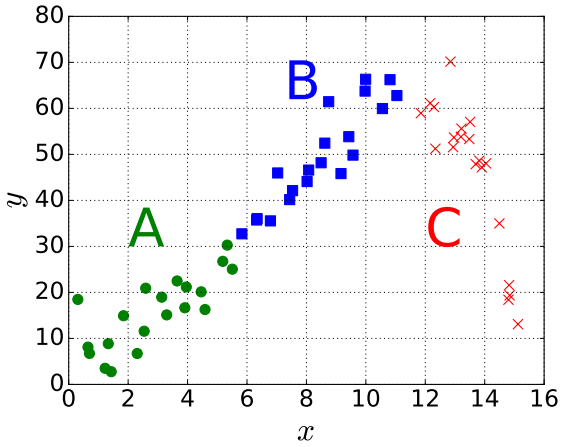
Name: _____

ML is from Mars, Validation is...

11. (10 points) It's 2030, and MIT's Subsurface Ice eXplorer (SIX) instrument has just sent back exciting data giving the concentration of water ice y at depth x beneath the surface of the north pole on Mars!

    Your task, as one of the mission specialists (back on Earth), is to figure out what hypothesis best models the data, which look like this:



    Due to how the SIX sampling drill works, the datapoints shown in this plot come from three disjoint subsets:

    A: depth $x = 0$ to x around 6 (circles)

    B: depth x around 6 to x around 12 (squares)

    C: depth x approximately above 12 (the symbol $\times$)

    And as an ML expert, you know that while you may train your model on one subset of data, you should test it on a different subset of data.

    (a) Suppose your hypothesis is that ice concentration is linearly related to depth, i.e. $y = \theta x + \theta_0$. You employ mean square error (MSE) for the objective function, and use dataset A for training, and dataset B for testing (since they are conveniently disjoint!). Let us say that that MSE below 30 is LOW, and MSE above 100 is HIGH. Judging from the above plot, will the MSE for training be LOW or HIGH? How about for testing? Explain why.

    Training error: ◯ LOW ◯ HIGH
    Testing error: ◯ LOW ◯ HIGH
    Explanation:

(b) Continuing with the hypothesis that ice concentration is linearly related to depth, you now employ datasets A and B (combined) for training, and dataset C for testing. Judging from the above plot, will the MSE for training be LOW or HIGH? How about for testing? Are your choices for training and testing datasets good ones? Explain.

Training error: ◯ LOW ◯ HIGH
Testing error: ◯ LOW ◯ HIGH
Explanation:

(c) Realizing that Mars is unlikely to be a snowball of ice (although it's possible Earth once was!), you switch to a family of hypotheses with nonlinear feature transforms, $y = \theta^T \phi_k(x) + \theta_0$, where $\phi_k(x)$ is a vector of polynomials up to order k. Can you think of any good way to evaluate what order k is the best to choose? Explain.

Explanation: