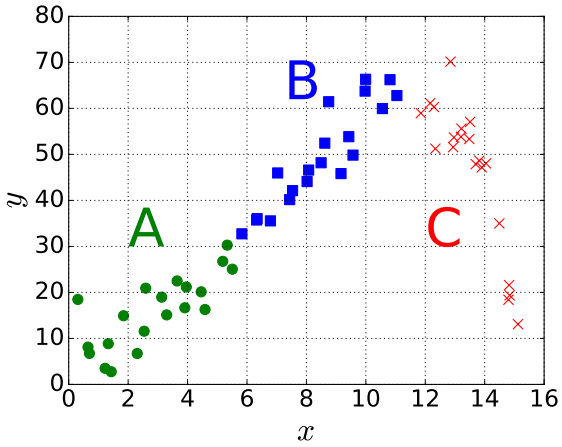
Name: _____

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.

    Solution:
    Training Error: LOW.
    Testing Error: LOW.

    Both errors are LOW because training on dataset A should produce a straight line which fits both A and B very well.

    (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.

Solution:
Training Error: LOW.
Testing Error: HIGH.

Training error will be LOW because training on dataset A and B should produce a straight line which fits both A and B very well. However, extrapolating forward the straight line produced will not be a good fit for dataset C leading to a HIGH testing MSE.

Are these choices for training and testing error good ones? If we are trying to model all of the data (i.e. the data in subsets A, B, *and* C), the union of subsets A and B is not representative; it misses out on the behavior in subset C. Similarly, subset C is not representative; it misses out on the behavior in subsets A&B. A better choice of training data would be one that has points from every subset; similarly, a better choice of testing data would have points from every subset.

(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.

Solution:
Training Set: randomly select data points from across all three datasets (A, B, C). A good percentage could be 80% data for training.

Testing Set: use the remaining 20% points not chosen for training to be part of the test set.

The reason one would want to choose randomly from across all datasets is because the data for training and for testing should come from the same sample distribution, even if they are disjoint datapoints.

Alternatively, use cross-validation. With cross-validation, you could use all the data for training then determine the best k by minimizing the error output by cross-validation. This would mean no need for a single separate test set.