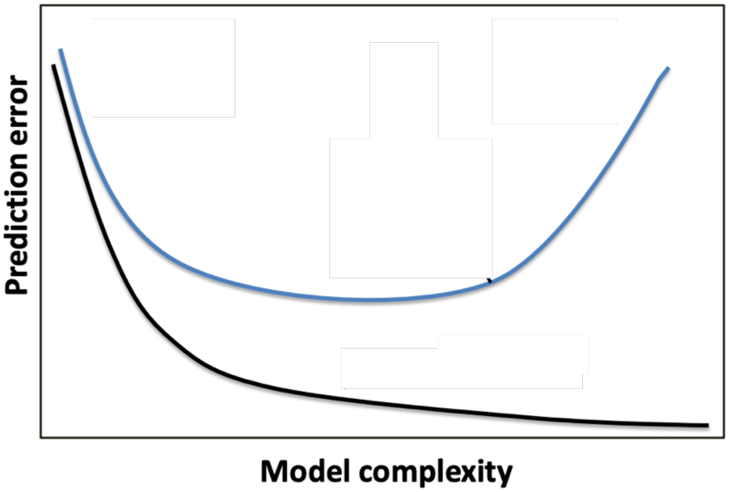
CS105 Midterm 2 Review

1. The following figure depicts training and validation curves of a learner with increasing model complexity



(a) Which of the curves is more likely to be the training error and which is more likely to be the validation error? Indicate on the graph by filling the dotted lines.

(b) In which regions of the graph are bias and variance low and high? Indicate clearly on the graph with four labels: “low variance”, “high variance”, “low bias”, “high bias”.

(c) In which regions does the model overfit or underfit? Indicate clearly on the graph by labeling “overfit” and “underfit”.

(2) Experimental design questions: For each of the listed descriptions below, circle whether the experimental set up is ok or problematic. If you think it is problematic, briefly state all the problems with their approach:

(a) A project team reports a low training error and claims their method is good.
(b) A project team claimed great success after achieving 98 percent classification accuracy on a binary classification task where one class is very rare (e.g., detecting fraud transactions). Their data consisted of 50 positive examples and 5 000 negative examples.
(c) A project team split their data into training and test. Using their training data and cross-validation, they chose the best parameter setting. They built a model using these parameters and their training data, and then report their error on test data.
(d) A project team performed a feature selection procedure on the full data and reduced their large feature set to a smaller set. Then they split the data into test and training portions. They built their model on training data using several different model settings, and report the best test error they achieved. Split data -> Choose features -> train model

(3) What is dimensionality reduction? What are the various applications of dimensionality reduction?

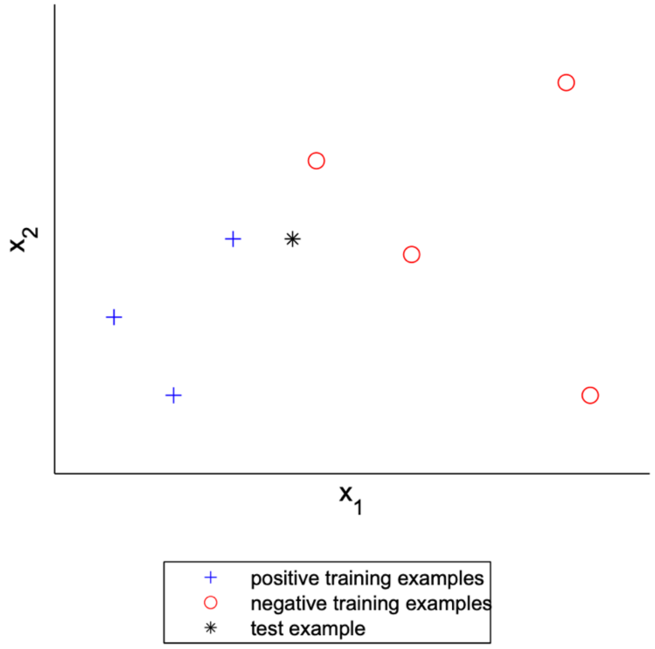
(4) Bias vs variance: Below, we list several classifiers and actions that might affect their bias and variance. Indicate how the bias and variance change (decrease, no change, or increase) in response to the action. :

(a) Reduce the number of leaves in a decision tree:
Bias _____________ Variance ___________

(b) Increase k in a k-nearest neighbor classifier:
Bias _____________ Variance ___________

(c) Increase the number of training examples in logistic regression:
Bias _____________ Variance ___________

(5) The figure depicts training data and a single test point for the task of classification given two continuous attributes X1 and X2. For each value of k, circle the label predicted by the k-nearest neighbor classifier for the depicted test point.



(a) Predicted label for k = 1:

(b) Predicted label for k = 3:

(c) Predicted label for k = 5:

(6) What are some challenges or considerations we have to make when applying KMeans ?

(7) Consider a supervised classification problem where we have 50 training examples and 10 features, where the features are stored in a 50 by 10 matrix X and the labels are stored in a 50 by 1 vector y (you can assume that the examples are in a random order).

Assume we are training a model that depends on a parameter 'k' with the following interface:

- model = train(X,y,k); % Train model on {X, y} with parameter k
- Y_Predict = predict(model, X_test); % Predict using the model on X_test.

Assume that k can be either 1, 2, or 3. Give informal pseudo-code describing how to choose k using 2-fold cross-validation.

(8) What does it mean to scale features? Why is it important to scale features for ML algorithms??

(9) What is the purpose of boosting and bagging?

(10) Describe bagging technique?

(11) Write pseudo code for a general boosting technique??