CS 7646 Summer 2016 - Quiz 2

June 14, 2016

Student Name _____

GT ID# _____

1 Fill in the blanks

1. Answer choices: learner, model, supervised, unsupervised, parametric, instance, classification, regression, discriminative, generative. (use each once)

   (a) A __parametric__ learner trains and stores a fixed number of coefficients to use for prediction.

   (b) An algorithm that improves performance at some task by exposure to data is called a __learner__.

   (c) A __discriminative__ learner approximates the conditional distribution of $P(y|X)$.

   (d) If your desired predictions are drawn from a limited, discrete set of outputs, you should use a __classification__ learner.

   (e) A __unsupervised__ learner is commonly used to cluster unlabelled observations into related groups.

   (f) If you wish to make a continuous numerical approximation, you would use a __regression__ learner.

   (g) A __supervised__ learner requires labelled training data in order to make predictions.

   (h) If you require the ability to sample new data (including observations) from your system, you should choose a __generative__ learner.

   (i) A system of equations or rules into which you can input observations to obtain predictions is called a __model__.

   (j) A __instance__ learner answers queries by directly consulting the entire training data set.

2. The method of dividing your data set into N distinct subsets to train N different models and compare their results is called __N-fold cross validation__.

3. Which learning algorithm would be a better choice if your primary concern is each of the following: (Answer choices: KNN Learner (K), or Polynomial Learner (P), may be used multiple times.)

   (a) You have no idea what kind of relationship exists between your observations and the value you wish to predict. __K (KNN has no bias — makes no assumptions)__

   (b) The learning algorithm must consume as little memory (or storage) as possible after training is complete. __P (after training, you only need the coefficients)__

   (c) Querying the model must be extremely fast and require little computational power. __P (query is just m1*x1+m2*x2…)__

   (d) You need a prediction only a few times per day, but a new data point will be added to the training set each second. __K (no retraining required to add a data point)__

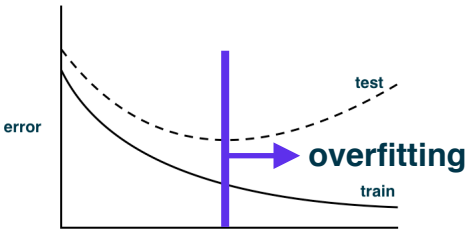
2 True/False

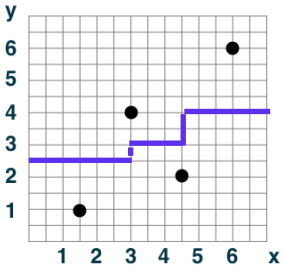
4. The AdaBoost algorithm improves its performance by iteratively *underweighting* poorly-predicted training points, eventually discarding them as insignificant outliers. **False - this is BrownBoost**

5. Bootstrap Aggregation reduces overfitting by building multiple learners sampled with replacement from the same training data. **True**

3 Analyzing a Learning Technique

6. Given this plot (below, left side) of train error and test error (on y) vs some hyperparameter H (on x), draw a vertical line at the value of H that provides the *optimal* fit to the data, and clearly indicate the region in which the model is *overfitting* the training data.





Remember: we are overfitting when training error gets better but test error gets worse.

7. Give one reason why supervised regression learning may not be the ideal way to develop a stock market trading strategy. **Supervised regression learning at best gives a single numeric prediction. It does not give us an actual trading policy (whether to buy/sell, how many shares, how long to hold the shares, etc).**

4 KNN

Classification is by majority vote and the order of the nearest neighbors does not matter.

8. Given a test point query into a *classification* 5NN Learner, if the nearest neighbors *in order of increasing distance* have y values $4, 4, 3, 3, 3$, the final value of y_{pred} will be **3**

9. The right-hand plot above shows the (X, y) training tuples for a *regression* 2NN Learner. The single X feature and y are both continuously-valued functions. Draw the prediction function (KNN query output) across the entire plot from $x = 0$ to $x = 7$.

x1 and x2 can only split 4 points one way, 1 point the other.

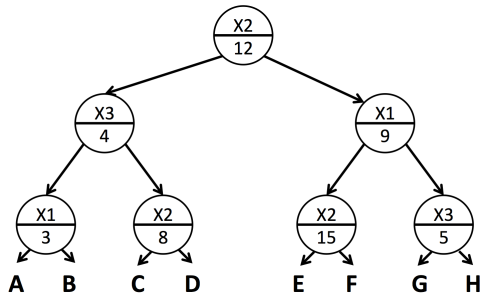
5 Decision Trees (CART)

10. The below left data points have reached some node of a CART decision tree during training:

x3 tries to split in the middle (index 2) then searches backwards. So it also splits four points to 1 point.

Only x4 can split 3 points to 2 points.

x_1	x_2	x_3	x_4	y
4	0	6	6	14.3
5	1	3	9	9.6
4	0	6	6	2.0
2	0	7	8	6.1
4	0	8	7	14.3



Per the CART algorithm, what *factor* and *value* will be selected for this split? **x4, 7**

12. The CART decision tree (above right) is queried with test point $(X_1, X_2, X_3) = (5, 9, 4)$. Which lettered leaf node will be reached by this query? **D (9 < 12 [left], 4 >= 4 [right], 9 >= 8 [right])**

13. During training, leaf node D is created with y training values 4,2,6,4. If leaf D is reached during a query, y_{pred} will be **4** if classification, or **4** if regression.

Classification chooses the most common value from 4, 2, 6, 4.

Regression takes the mean value of 4, 2, 6, 4. (4+2+6+4) / 4 = 16 / 4 = 4