AAE 722 Homework I (2020)

Due July 17th

1. Bayes rule (5 pts)

We have 2 opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black ball and a white ball. You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball, it is black. You now pick the second ball from that same bag. What is the probability that this ball is also black? (Remember to apply Bayes rule to answer the question: $P(B|A) = \frac{P(A|B)\cdot P(B)}{P(A)}$

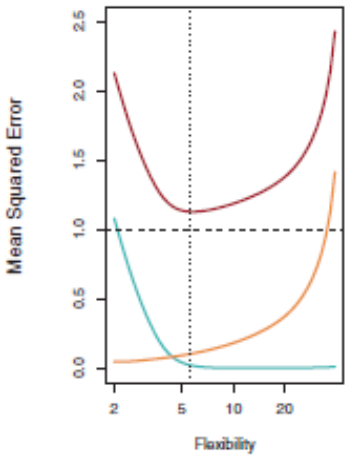
2. Type of machine learning systems (5 pts)

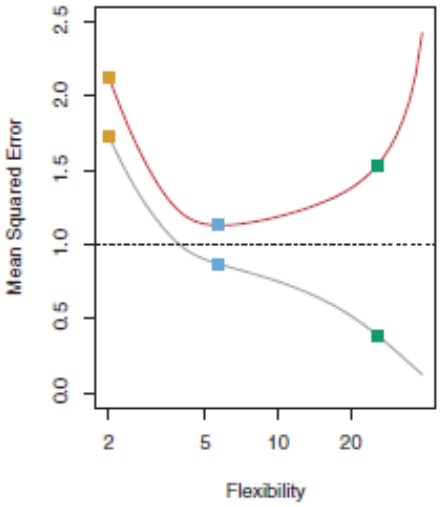
For each of the following tasks, identify which type of learning is involved (supervised, reinforcement, or unsupervised). Notice that some of them can fit more than one type.

(1) Recommending a book to a user in an online bookstore
(2) Playing "Go" game
(3) Categorizing movies into different types
(4) Learning to play music
(5) Deciding the maximum allowed debt for each bank customer(credit limit)

3. Bias-variance decomposition (10 pts)

(1) What is the expected test MSE (mean squared error) for a test sample point of (x_0, y_0), or so-called expected prediction error (EPE)? Use $\hat{f}(\cdot)$ for the prediction function. What does it measure?
(2) The expected test MSE (or EPE) in (1) can be decomposed into three quantities: the variance of the prediction, the squared bias of the prediction and the variance of the error term (ϵ). Write down the formula and explain the meanings of individual elements.
(3) The figure below on the left illustrates the curves of expected MSE, squared bias, variance, and irreducible error curves as we go from less flexible to more flexible machine learning methods. Label each curve and explain why each curve has the displayed shape.
(4) The figure below on the right plots the training error, testing error and the irreducible error curves. Label each one and explain why each curve has the shape displayed in the figure.





4. Prediction and estimation (10 pts)

Suppose we consider to predict $y_i = \mu + \epsilon_i$ using $\hat{u} = \alpha\bar{y}$, where $\bar{y}$ is the sample mean, ϵ has mean 0 and variance σ^2.

(1) Derive the optimal α^* for minimizing the quadratic loss of prediction, $E(\ell(\alpha\bar{y}, y)) = E[(y - \alpha\bar{y})^2]$ using the bias-variance decomposition formula.
(2) Is $\alpha^*\bar{y}$ derived in (1) a biased estimator of μ? Why?

5. R application: use the dataset of house values (`homes2004.csv`; the variables are defined in `homes2004code.txt`) provided to do the following and submit the `html` file of your code and outputs. (30 pts)

Note that for the coding questions, please report the results and write up your analysis according to the instructions.

(1) Figure out the numbers of rows with `NA` value and remove them from the data. Report the dimensions of the data before and after the cleaning.

(2) Report the frequency of educational level of the sample householders.

(3) Count the states in the data and report the number.

(4) Report the `mean`, `max`, `min` of the `BATHS` and `VALUE` variables.

(5) Generate a dummy variable indicating the number of bedrooms is greater than 2. Report the frequency table of the generated variable.

(6) Report the average current market value and the average purchase price by the numbers of full bedrooms in unit.

(7) Generate the scatter plots of the numbers of full bedrooms and the average current market value.

(8) Run a linear regression with at least 3 explanatory variables to explore the effects of housing characteristics on the current market value of the unit. Report the output and analyze your results, for example, by describing the partial effect like "1 more bedroom increases the home value by $xx with other factors fixed."

Hint: You can use the `factor()` function or the `ifelse` statement to generate dummy variables.

(9) Report and analyze the diagnostic plots of the model you specified in (8).