Assignment 2 (Sol.)
Introduction to Machine Learning
Prof. B. Ravindran

1. Let $A^{m \times n}$ be a matrix of real numbers. The matrix AA^T has an eigenvector x with eigenvalue b. Then the eigenvector y of A^TA which has eigenvalue b is equal to

   (a) x^TA
   (b) A^Tx
   (c) x
   (d) Cannot be described in terms of x

   Sol. (b)

   $$(AA^T)x = bx$$

   Multiplying by A^T on both sides and rearranging,

   $$(A^T)(AA^Tx) = A^T(bx)$$
   $$(A^TA)(A^Tx) = b(A^Tx)$$

   Hence, A^Tx is an eigenvector of A^TA, with eigenvalue b.

2. Let $A^{n \times n}$ be a row stochaistic matrix - in other words, all elements are non-negative and the sum of elements in every row is 1. Let b be an eigenvalue of A. Which of the following is true?

   (a) $|b| > 1$
   (b) $|b| \leq 1$
   (c) $|b| \geq 1$
   (d) $|b| < 1$

   Sol. (b)
   Note that $Ax = bx$ where x is an eigenvector and b is an eigenvalue. Let x_max be the largest element of x. Let A_{ij} denote the ith row, jth column element of A, and x_j denote the jth element of x. Now,

   $$\sum_{j=1}^{j=n} A_{ij}x_j = bx_i$$

   Let us consider the case where x_i corresponds to the maximum element of x. The RHS is equal to bx_{max}. Now, since $\sum_{j=1}^{j=n} A_{ij} = 1$ and $A_{ij} > 0$, the LHS is less than or equal to x_max. Hence,

   $$bx_{max} \leq x_{max}$$

Hence,

$$|b| \leq 1$$

3. Let u be a $n \times 1$ vector, such that $u^T u = 1$. Let I be the $n \times n$ identity matrix. The $n \times n$ matrix A is given by $(I - kuu^T)$, where k is a real constant. u itself is an eigenvector of A, with eigenvalue -1. What is the value of k?

   (a) -2

   (b) -1

   (c) 2

   (d) 0

   Sol. (c)

$$(I - kuu^T)u = -u$$
$$u - ku(u^T u) = -u$$
$$2u - ku = 0$$

   Hence, $k = 2$

4. Which of the following are true for any $m \times n$ matrix A of real numbers.

   (a) The rowspace of A is the same as the columnspace of A^T

   (b) The rowspace of A is the same as the rowspace of A^T

   (c) The eigenvectors of AA^T are the same as the eigenvectors of $A^T A$

   (d) The eigenvalues of AA^T are the same as the eigenvalues of $A^T A$

   Sol. (a), (d)
   Since the rows of A are the same as the columns of A^T, the rowspace of A is the same as the columnspace of A^T. The eigenvalues of AA^T are the same as the eigenvalues of $A^T A$, because if $AA^T x = \lambda x$ we get $A^T A(A^T x) = \lambda(A^T x)$. (b) is clearly not necessary. (c) need not hold, since although the eigenvalues are same, the eigenvectors have a factor of A^T multiplying them.

5. Consider the following 4 training examples

   - $x = -1, y = 0.0319$
   - $x = 0, y = 0.8692$
   - $x = 1, y = 1.9566$
   - $x = 2, y = 3.0343$

   We want to learn a function $f(x) = ax + b$ which is parametrized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

   (a) $(1, 1)$

(b) $(1, 2)$

(c) $(2, 1)$

(d) $(2, 2)$

Sol. (a)
The line $y = x + 1$ is the one with minimum squared error out of all the four proposed.

6. You are given the following five training instances

   - $x_1 = 2, x_2 = 1, y = 4$
   - $x_1 = 6, x_2 = 3, y = 2$
   - $x_1 = 2, x_2 = 5, y = 2$
   - $x_1 = 6, x_2 = 7, y = 3$
   - $x_1 = 10, x_2 = 7, y = 3$

   We want to model this function using the K-nearest neighbor regressor model. When we want to predict the value of y corresponding to $(x_1, x_2) = (3, 6)$

   (a) For $K = 2, y = 3$

   (b) For $K = 2, y = 2.5$

   (c) For $K = 3, y = 2.33$

   (d) For $K = 3, y = 2.666$

   Sol. (b), (c)
   Points 3 and 4 are the two closest points to $(x_1, x_2) = (3, 6)$. Hence, the 2-nearest neighbor prediction would be the average of their y values, which is 2.5. The third closest point is point 2, on adding which we get an average y value of 2.33.

7. Bias and Variance can be visualized using a classic example of a dart game. We can think of the true value of the parameters as the bull's-eye on a target, and the arrow's value as the estimated value from each sample. Consider the following situations, and select the correct option(s)

   (a) Player 1 has low variance compared to player 4

   (b) Player 1 has higher variance compared to player 4

   (c) Bias exhibited by player 2 is more than that done by player 3.

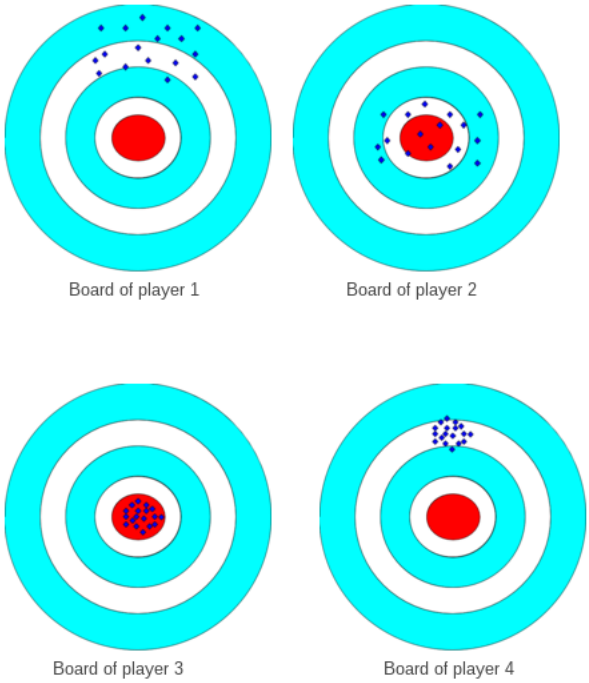

Figure 1: Figure for Q7

Sol.(b)
We can think of the true value of the population parameter as the bulls-eye on a target, and of the sample statistic as an arrow fired at the bulls-eye. Bias and variability describe what happens when a player fires many arrows at the target. Bias means that the aim is off, and the arrows land consistently off the bulls-eye in the same direction. The sample values do not center about the population value. Large variability means that repeated shots are widely scattered on the target. Repeated samples do not give similar results but differ widely among themselves.

(a) Player 1 - Shows high bias and high variance
(b) Player 2 - Shows low bias and high variance
(c) Player 3 - Shows low bias and low variance
(d) Player 4 - Shows high bias and low variance

8. Choose the correct option(s) from the following.

(a) When working with a small dataset, one should prefer low bias/high variance classifiers over high bias/low variance classifiers.
(b) When working with a small dataset, one should prefer high bias/low variance classifiers over low bias/high variance classifiers.
(c) When working with a large dataset, one should prefer high bias/low variance classifiers over low bias/high variance classifiers.

(d) When working with a large dataset, one should prefer low bias/high variance classifiers over high bias/low variance classifiers.

Sol. (b), (d)
On smaller datasets, variance is a concern since even small changes in the training set may change the optimal parameters significantly. Hence, a high bias/ low variance classifier would be preferred. On the other hand, with a large dataset, since we have sufficient points to represent the data distribution accurately, variance is not of much concern. Hence, one would go for the classifier with low bias even though it has higher variance.

9. Consider a modified k-NN method in which once the k nearest neighbours to the query point are identified, you do a linear regression fit on them and output the fitted value for the query point. Which of the following is/are true regarding this method.

   (a) This method makes an assumption that the data is locally linear.

   (b) In order to perform well, this method would need dense distributed training data.

   (c) This method has higher bias compared to K-NN

   (d) This method has higher variance compared to K-NN

   Sol. (a), (b), (d)
   Since we do a linear fit in the k-neighborhood, we are making an assumption of local linearity. Hence, (a) holds. The method would need dense distributed training data to perform well, since in the case of the training data being sparse, the k-neighborhood would end up being quite spread out (not really local anymore). Hence, the assumption of local linearity would not give good results. Hence, (b) holds. The method has higher variance, since we now have two parameters (slope and intercept) instead of one in the case of conventional k-NN. (In the conventional case, we just try to fit a constant, and the average happens to be the constant which minimizes the squared error).

10. The Singular Value Decomposition (SVD) of a matrix R is given by USV^T. Consider an orthogonal matrix Q and $A = QR$. The SVD of A is given by $U_1 S_1 {V_1}^T$. Which of the following is/are true?
    Note-There can be more than one correct option.

    (a) $U = U_1$

    (b) $S = S_1$

    (c) $V = V_1$

    Sol. (b), (c)
    The matrix V_1 represents the eigenvectors of $A^T A$. Now since

    $$A^T A = (R^T Q^T)(QR)$$

    Since Q is orthogonal, $Q^T Q = I$. Therefore,

    $$A^T A = (R^T I R)$$
    $$A^T A = (R^T R)$$

Since these matrices are equal, their eigenvectors will be the same. V represents the eigenvectors of R^TR and V_1 the eigenvectors of A^TA. Hence, $V = V_1$. Also, since S represents the eigenvalues of A^TA (as well as AA^T, since they the set of eigenvalues is the same as both), $S = S_1$. However, U need not be equal to U_1, since $AA^T = (QR)(R^TQ^T) = Q(RR^T)Q^T$.

11. Assume that the feature vectors defining the training data are not all linearly independent. What happens if we apply the standard linear regression formulation considering all feature vectors?

    (a) The coefficients $\hat{\beta}$ are not uniquely defined.

    (b) $\hat{y} = X\hat{\beta}$ is no longer the projection of y into the column space of X.

    (c) X is full rank.

    (d) X^TX is singular.

    Sol. (a), (d)
    Since X is such that its columns are not linearly independent, it will not have full rank. As a result, X^TX also wont have full rank. Hence, it will be singular. The coefficients will not be uniquely defined since there would be multiple feasible solutions as a result of X^TX being singular and hence non-invertible.