Advanced Predictive Modeling

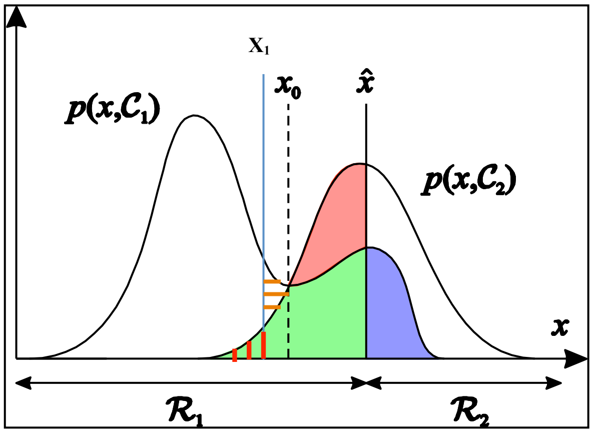
Quiz 3, Fall 2017 Total Marks = 15; *Time: 15 min.*

Declaration: By submitting this quiz for grading, I affirm that I have neither given nor received help from another examinee and acknowledge that **this is a closed-book, closed-notes test.**

Name: ________________________________

Signature: ________________________________

Q3 (3+3 pts). A figure from the notes related to a 2-class problem is reproduced below. Suppose your classifier labels all objects with “x” < x1 as belonging to Class 1, and labels all others as Class 2. Indicate by shading the appropriate regions (i) the probability that a point belonging to Class 2 is misclassified. (ii) the Error in excess of the Bayes error rate that you classifier incurs.



▥ the probability that a point belonging to Class 2 is misclassified

▤ the error in excess of the Bayes error rate that you classifier incurs.

Q2. (2 +2 pts). What is the difference between the assumption made in the design of the QDA classifier as compared to that made for the LDA (Linear Discriminant Analysis) classifier? Suggest a situation (with reasoning) where LDA is likely to perform better than QDA in terms of classification accuracy.

LDA assumes the data are normally distributed with identical class covariance matrices (ie, the same covariance matrix over all classes) while QDA assumes different covariance matrixes across classes. LDA is likely to perform better when you don’t have enough data and the true separator is linear. In this case QDA may not be able to capture the general trend and could overfit to a nonlinear decision boundary.

Q3 (2+1+ 2 pts). In the design of support vector machines (SVMs),

(i) what are “support vectors”?
(ii) what is the role of the slack penalty
(iii) what is the “kernel trick” and how does it allow one to expand on the classification abilities of SVMs?

(i) Support vectors are the points closest to the hyperplane that is orthogonal to the weight vector and maximizes the margin between the classes.

(ii) The slack penalty governs the amount of deviation from a simple (hyperplane) boundary. The lower the penalty, the less the cost of a solution with many points on the “wrong” side of the boundary and more the emphasis on solution that focuses on the margin term. This leads to “less curvy” solutions. Larger penalty for slack will correspond to decreased bias but increased variance.

(iii) The kernel trick is an implicit transformation and inner product of the original data in a high dimensional space and allows us to find a linear boundary in a suitable higher-dimensional space without explicitly mapping the features into this space.
Concretely, let ϕ: x-> ϕ(x). The inner product in the transformed space is equivalent to evaluating the kernel function K(x_i,x_j) = ϕ(x_i)^T ϕ(x_j) Therefore, we can say that the kernel function implicitly transforms the data into a higher dimensional space and evaluates the inner product, without requiring explicit evaluation of ϕ(x)