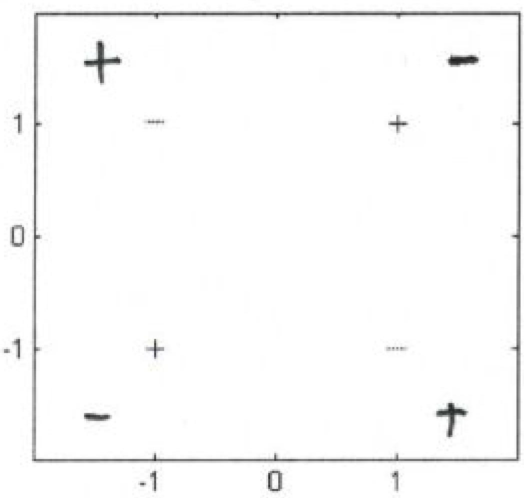
18th September 2019 AML: Midsem Exam Solution Total Marks: 56

A1 [4+3+3 Marks].

(a) w = $(0,0,0,1)^T$

(b)

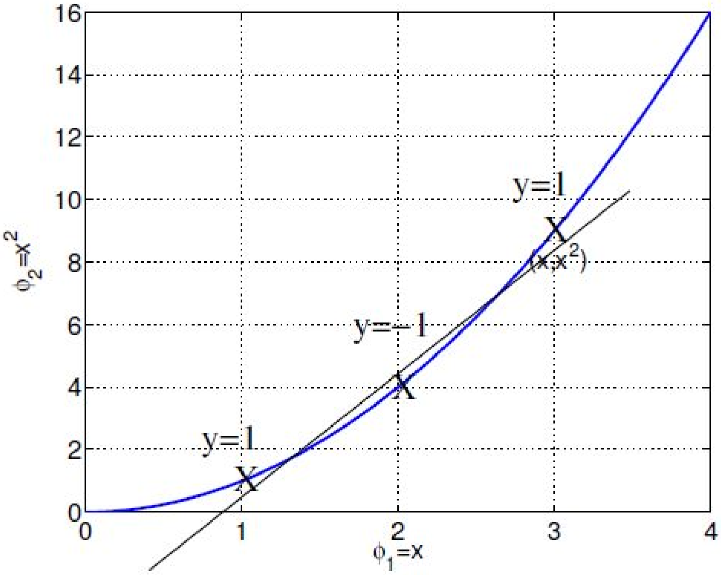


(c) $1+ x_1x_1' + x_2x_2' + x_1x_2x_1'x_2'$

A2 [3+3+2+2+2 Marks].

(a) (x = 1, y = 1), (x = 2, y = −1), (x = 3, y = 1)

(b)



(c) Greater
(d) Greater
(e) **True:** We need to normalize the margin for it to be meaningful. For example, a simple scaling of the feature vectors would lead to a larger margin. Such a scaling does not change the decision boundary, however, and so the larger margin cannot directly inform us about generalization.

A3 [9+5 Marks]

A. 1. Fixed rules are functions without any parameters (e.g., summation or multiplication of the kernels) and do not need any training.
2. Heuristic approaches use a parameterized combination function and find the parameters of this function generally by looking at some measure obtained from each kernel function separately. These measures can be calculated from the kernel matrices or taken as the performance values of the single kernel-based learners trained separately using each kernel.
3. Optimization approaches also use a parametrized combination function and learn the parameters by solving an optimization problem. This optimization can be integrated to a kernel-based learner or formulated as a different mathematical model for obtaining only the combination parameters.

B. We can use different heuristics to estimate the weighting function values using conditional class probabilities, $\Pr(y_i = y_j | x_i)$ and $\Pr(y_j = y_i | x_j)$, calculated with a nearest-neighbor approach. However, each kernel function corresponds to a different neighborhood and $n_m(.,.)$ is calculated on the neighborhood induced by $k_m(.,.)$. For an unlabeled data instance x, they take its class label once as +1 and once as −1, calculate the discriminant values $f(x|y =+ 1)$ and $f(x|y =- 1)$, and assign it to the class that has more confidence in its decision (i.e., by selecting the class label with greater $yf(x|y)$ value).

A4 [4 Marks]. We assume that instead of two underlying distributions, there are actually three underlying distributions, which we will denote $q^{(o)}$, $q^{(g)}$ and $q^{(i)}$. We then consider $p^{(o)}$ to be a mixture of $q^{(o)}$ and $q^{(g)}$, and consider $p^{(i)}$ to be a mixture of $q^{(g)}$ and $q^{(i)}$. One can intuitively view the $q^{(o)}$ distribution as a distribution of data that is truly out-of-domain, $q^{(i)}$ as a distribution of data that is truly in-domain and $q^{(g)}$ as a distribution of data that is general to both domains. Thus, knowing $q^{(g)}$ and $q^{(i)}$ is sufficient to build a model of the in-domain data. The out-of-domain data can help us by providing more information about $q^{(g)}$ than is available by just considering the in-domain data. For example, in part-of-speech tagging, the assignment of the tag “determiner” (DT) to the word “the” is likely to be a general decision, independent of domain. However, in the Wall Street Journal, “monitor” is almost always a verb (VB), but in technical documentation, it will most likely be a noun. The $q^{(g)}$ distribution should account for the case of “the/DT”, the $q^{(o)}$ should account for “monitor/VB” and $q^{(i)}$ should account for “monitor/NN.”

A5 [3 Marks]

A6 [3+3 Marks]. Similar to the techniques discussed in class for dictionary learning, the same techniques can be applied to transform learning. Steps to perform classification will differ.

A7 [7 Marks].

A. True
B. False
C. False
D. False
E. True
F. False
G. False