

## ASSIGNMENT 6, Solutions

### 1 (300pts) Regularization

For these problems we write the penalty term as  $\Omega_R(\mathbf{w})$  and the final error to optimize will be  $E(\mathbf{w}) = R_{emp}(\mathbf{w}) + \lambda\Omega_R(\mathbf{w})$ , where  $\lambda$  is the regularization parameter.

(a) Binary weights:

$$\Omega_R(\mathbf{w}) = \sum_i (w_i^2 - 1)^2$$

where the sum is over all the weights  $w_i$ .

(b) Similar weights:

$$\Omega_R(\mathbf{w}) = \sum_i \sum_j (w_i - w_j)^2$$

where the double sum is over all possible pairs of weights  $w_i, w_j$ .

(c) Weight sharing: Let  $\{k_\alpha\}_{\alpha=1}^K$  represent the “means” of the  $K$  clusters. Each weight  $w_i$  “belongs” to the cluster with the closest mean. We let  $S_\alpha$  denote the set of weights that belong to cluster  $k_\alpha$ . We can use the measure in (b) to enforce that the weights in each cluster are similar. We can sum over the clusters to get the overall spread. Now comes the choice of the  $k_\alpha$ . We should choose the  $k_\alpha$  that minimize this spread. the final penalty term then becomes

$$\Omega_R(\mathbf{w}) = \min_{k_1, k_2, \dots, k_K} \sum_{\alpha=1}^K \sum_{w_i \in S_\alpha} \sum_{w_j \in S_\alpha} (w_i - w_j)^2$$

(d) Such terms are regularization terms because they restrict the initial learning model. (b) and (c) encourage smoothness by effectively only allowing 1 or  $K$  “different” weights. (a) on the other hand is not necessarily encouraging smoothness. One could get very complicated functions with binary weights.

(e) Symmetry: Invariance: Let  $Z = \{\mathbf{z}_i\}_{i=1}^M$  be an *arbitrary* set chosen to represent the input space. Let  $K$  be any integer  $> 0$ . We enforce invariance on these points by adding the following type of penalty term

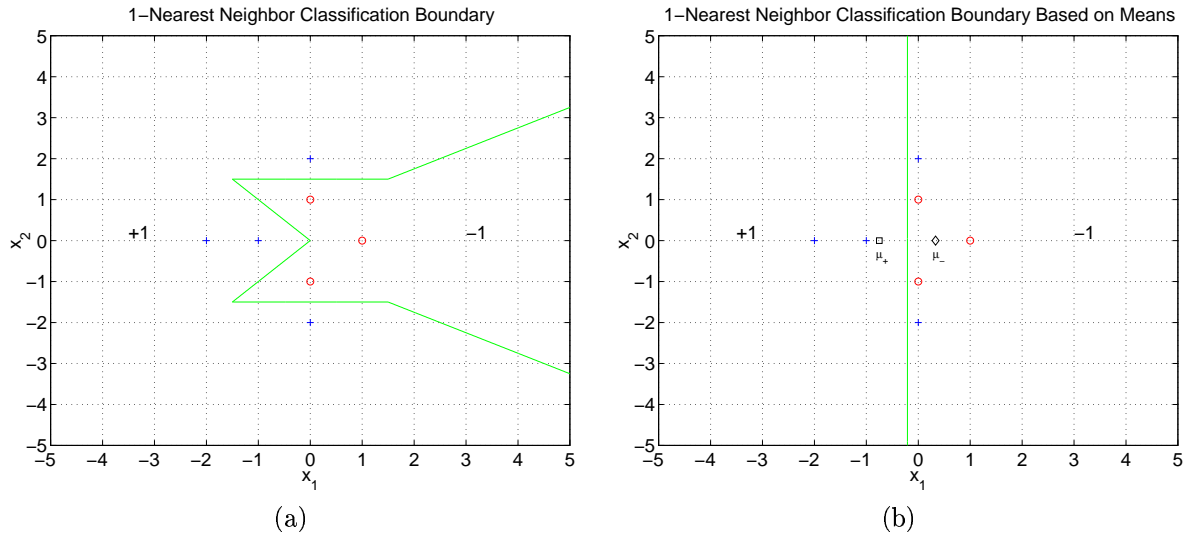
$$\Omega_R(\mathbf{w}) = \sum_{i=1}^M \sum_{k=1}^K \left( g(\hat{L}^k \mathbf{z}_i; \mathbf{w}) - g(\mathbf{z}_i; \mathbf{w}) \right)^2$$

where  $\hat{L}^k$  represents composition of  $\hat{L}$  with itself  $k$  times. Specializing to the case where  $\hat{L} = -\mathbf{I}$ , which corresponds to symmetry, we see that we only need go up to  $K = 1$ , hence the penalty term becomes

$$\Omega_R(\mathbf{w}) = \sum_{i=1}^M (g(-\mathbf{z}_i; \mathbf{w}) - g(\mathbf{z}_i; \mathbf{w}))^2$$

(f) Cross Validation.

## 2 (350pts) Nearest Neighbor Classifiers



(c) 2 Points are misclassified using only the means ( $\mu_+$ ,  $\mu_-$ ) to construct the decision rule.

## 3 (350pts) Radial Basis Functions

(a) Using exact interpolation, it suffices to find a parameter  $\mathbf{w}$  such that  $g(\mathbf{x}_m) = y_m$  for all the data points. Let  $A_{mn} = e^{-\frac{1}{2}\|\mathbf{x}_m - \mathbf{x}_n\|^2}$ . Then we need  $\sum_{n=1}^7 A_{mn} w_n = y_m$ , or that  $\mathbf{A}\mathbf{w} = \mathbf{y}$  which gives  $\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$ . Performing the calculation in Matlab, we find that

$$\mathbf{w}^T = [0.9924, -4.1069, -4.1069, 3.8459, 3.1528, 3.1528, -0.7849]$$

(b)

