CSE 250B: Homework 4 Solutions

1. To classify a point x, we evaluate the three linear functions and pick the one with the highest value.

   The region where class 1 beats class 2 is:

$$w_1 \cdot x + b_1 > w_2 \cdot x + b_2 \;\Leftrightarrow\; (w_1 - w_2) \cdot x + (b_1 - b_2) > 0 \;\Leftrightarrow\; x_2 > 1$$

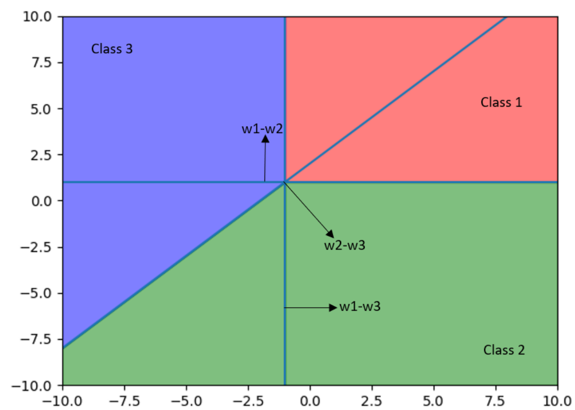
   The region where class 1 beats class 3 is:

$$w_1 \cdot x + b_1 > w_3 \cdot x + b_3 \;\Leftrightarrow\; (w_1 - w_3) \cdot x + (b_1 - b_3) > 0 \;\Leftrightarrow\; x_1 > -1$$

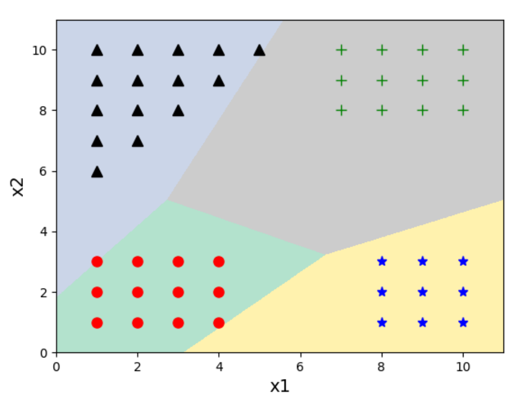
   The region where class 2 beats class 3 is:

$$w_2 \cdot x + b_2 > w_3 \cdot x + b_3 \;\Leftrightarrow\; (w_2 - w_3) \cdot x + (b_2 - b_3) > 0 \;\Leftrightarrow\; x_1 - x_2 > -2$$

   So class 1 is predicted in the intersection of the first two regions, etc. This is summarized in the figure below.

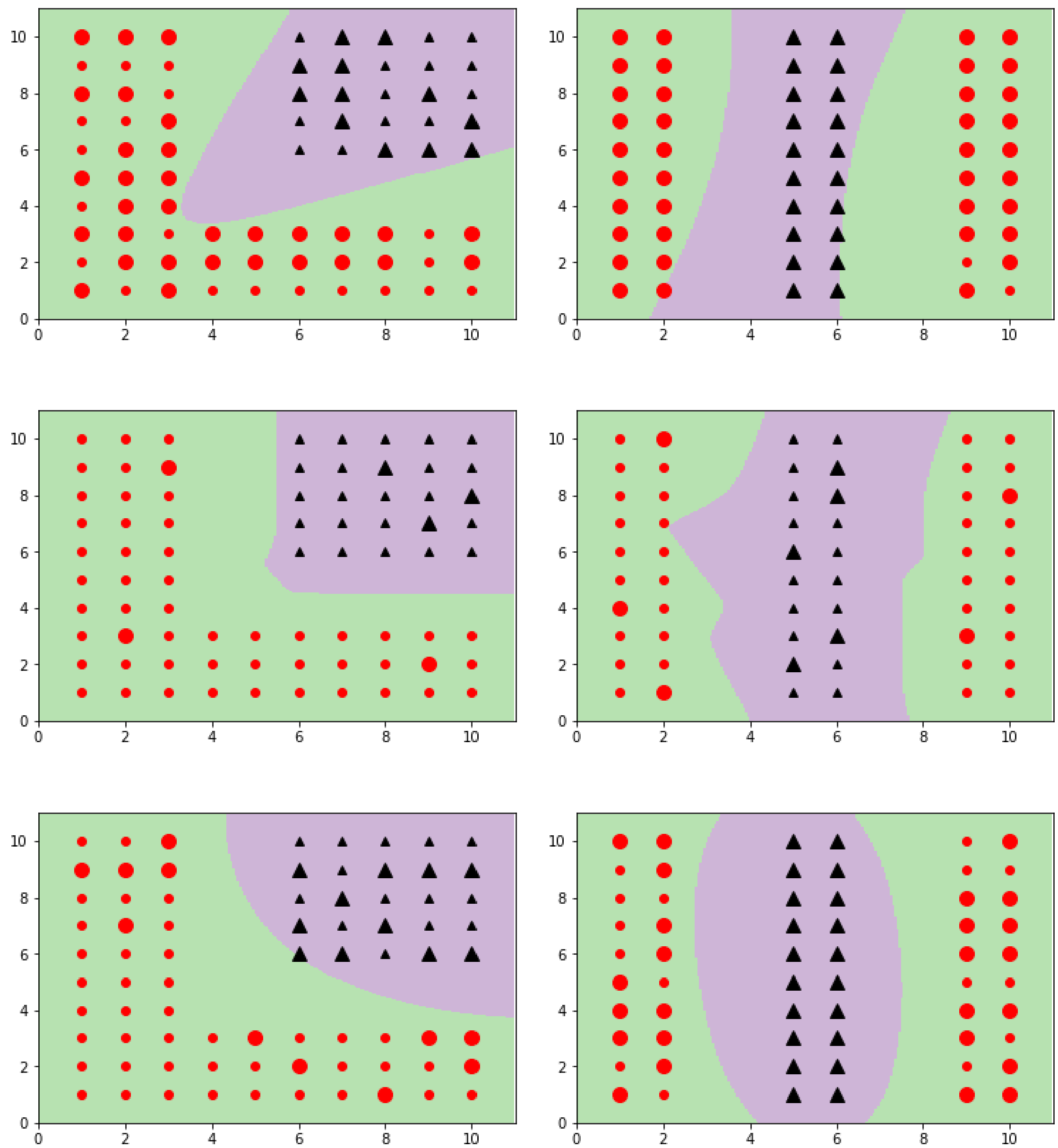


2. *Multiclass Perceptron.*



3. *Kernel Perceptron.*

   Left: `data1`, Right: `data2`. The three rows correspond to the quadratic kernel, the RBF kernel with $\sigma = 1.0$, and the RBF kernel with $\sigma = 10.0$, respectively.



4. (a) The MNIST data is not linearly separable.

C	train error (%)	test error (%)
0.01	14.94	15.44
0.1	10.92	11.62
1.0	10.97	11.90
10.0	11.51	12.02
100.0	11.80	12.88

   (b) Using a quadratic kernel with $C = 1.0$, we get training error 0.0% and test error 1.94%. The number of support vectors is 8652.

5. *Pointwise product of positive semidefinite matrices.*

(a) Because X and Y are independent, $\mathbb{E}(Z) = \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = 0$.

$$\begin{aligned}
\mathrm{Cov}(Z_i, Z_j) &= \mathrm{Cov}(X_iY_i, X_jY_j) \\
&= \mathbb{E}(X_iX_jY_iY_j) - \mathbb{E}(X_iY_i)\mathbb{E}(X_jY_j) \\
&= \mathbb{E}(X_iX_j)\mathbb{E}(Y_iY_j) - \mathbb{E}(X_i)\mathbb{E}(Y_i)\mathbb{E}(X_j)\mathbb{E}(Y_j) \\
&= \mathbb{E}(X_iX_j)\mathbb{E}(Y_iY_j) \\
&= \mathbb{E}((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)))\mathbb{E}((Y_i - \mathbb{E}(Y_i))(Y_j - \mathbb{E}(Y_j))) \\
&= \mathrm{Cov}(X_i, X_j)\mathrm{Cov}(Y_i, Y_j) \\
&= M(i,j)N(i,j)
\end{aligned}$$

So, the covariance matrix of Z is the pointwise product of M and N.

(b) Since covariance matrices are always positive semidefinite, it follows that Q is PSD.

6. *Closure properties of kernels.*
In each case, we will establish that $k(x, x')$ is a kernel function by invoking Mercer's condition. That is, we will show that for any finite set of points $x_1, \ldots, x_m \in \mathcal{X}$, the $m \times m$ matrix K given by

$$K_{ij} = k(x_i, x_j)$$

is positive semidefinite.

(a) Pick any $x_1, \ldots, x_m \in \mathcal{X}$ and define matrix K as above. Also define $m \times m$ matrices $K^{(1)}$ and $K^{(2)}$ by

$$K^{(1)}_{ij} = k_1(x_i, x_j), \quad K^{(2)}_{ij} = k_2(x_i, x_j).$$

Since k_1 and k_2 are kernel functions, we know that $K^{(1)}$ and $K^{(2)}$ are PSD. And since the set of PSD matrices is closed under addition and under multiplication by a nonnegative scalar, it follows that $K = \alpha_1 K^{(1)} + \alpha_2 K^{(2)}$ is also PSD.

(b) Define $K, K^{(1)}, K^{(2)}$ as above. This time K is the pointwise product of $K^{(1)}$ and $K^{(2)}$; by the previous problem, K is PSD.

7. Let $\Phi(x) = (x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{d-1}x_d)$, where all pairs of coordinates are included. Then

$$\begin{aligned}
\Phi(x) \cdot \Phi(z) &= \sum_{i=1}^{d} x_i^2 z_i^2 + 2\sum_{i \neq j} x_i z_i x_j z_j \\
&= (x_1z_1 + x_2z_2 + \ldots + x_dz_d)^2 \;=\; (x \cdot z)^2 \;=\; k(x,z).
\end{aligned}$$

8. *Monotone disjunctions.*

(a) There are as many disjunctions as there are subsets of features, so $|\mathcal{H}| = 2^d$.

(b) The true error of h can be bounded thus, with probability at least $1 - \delta$:

$|H| = \sum_{i}^{k} {}^{d}C_{i}$

$$\mathrm{err}(h) \leq \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta} = \frac{1}{n}\left(d \ln 2 + \ln \frac{1}{\delta}\right).$$

(c) $|\mathcal{H}_k| \leq d^k$, so we get

$$\mathrm{err}(h) \leq \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta} = \frac{1}{n}\left(k \ln d + \ln \frac{1}{\delta}\right).$$

9. By the central limit theorem, $\widehat{p}$ follows roughly a $N(3/4, 1/1600)$ distribution. With 95% probability, $\widehat{p}$ will fall within 2 standard deviations of its mean, that is, in the interval $[0.7, 0.8]$.

10. *VC dimension.*

(a) The class $\mathcal{H}$ of intervals on the real line shatters any set of two distinct points: it can realize all four labelings of these points. But it cannot shatter any set of three points, because it cannot label the middle one 0 while making the other two 1. Therefore $\mathrm{VC}(\mathcal{H}) = 2$.

(b) The class $\mathcal{H}$ of axis-aligned rectangles in the plane shatters the set $\{(0,1),(0,-1),(1,0),(-1,0)\}$: all 16 labelings can be realized. But it cannot shatter any set of five points. To see this, pick any $x_1,\ldots,x_5 \in \mathbb{R}^2$. One of them must lie in the bounding box of the other four points; say x_5 lies in the bounding box of x_1,x_2,x_3,x_4. Then we cannot realize the labeling $y_1 = y_2 = y_3 = y_4 = 1$ and $y_5 = 0$. Thus $\mathrm{VC}(\mathcal{H}) = 4$.