0 Instructions

Homework is due Thursday, February 20, 2024 at 23:59pm Central Time. Please refer to `https://courses.grainger.illinois.edu/cs446/sp2024/homework/hw/index.html` for course policy on homeworks and submission instructions.

1 Soft-margin SVM: 4pts

Rewriting constraints of the primal form as,

$$y_i(w^T x_i) - 1 + \xi_i \geq 0 \implies 1 - \xi_i - y_i(w^T x_i) \leq 0$$

$$\xi_i \geq 0 \implies -\xi_i \leq 0$$

Thus we can form a Lagrangian for the primal as,

$$L(w, \alpha, \beta, \xi) = \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i(1 - \xi_i - y_i(w^T x_i)) + \sum_{i=1}^{N} \beta_i(-\xi_i)$$

To find the dual we need to find,

$$\min_{w,\xi} L(w, \alpha, \beta, \xi)$$

Rewriting as,

$$\min_{\xi}(\min_{w} L(w, \alpha, \beta, \xi))$$

We find optimal w^* first to solve for optimal ξ^*. Let's minimize with respect to w,

$$\nabla_w L(w, \alpha, \beta, \xi) = 0$$

Note we get the same form of w^* as seen in the derivation of the dual of the hard-margin SVM, since the other terms are with respect to ξ, β, α which are independent of w. Hence we can say from the derivation in class that,

$$w^* = \sum_{i=1}^{N} \alpha_i y_i x_i$$

Plugging this value back in the Lagrangian and minimizing with respect to ξ,

$$\nabla_{\xi} L(w^*, \alpha, \beta, \xi) = C - \alpha_i - \beta_i = 0$$

We see that the minimization is independent of the value of ξ. Hence, we only get constraints for our multipliers,

$$\alpha_i + \beta_i = C$$

Lastly, we need to minimize with respect to our bias term which is an implicit term in our given equation. We can decompose it as follows

$$y_i(w^T x_i) \equiv y_i(w^T x_i + b)$$

Then,

$$\nabla_b L(w^*, \alpha, \beta, \xi^*) = \sum_{i=1}^{N} \alpha_i y_i = 0$$

Now we can write out our dual program,

$$\max_{\alpha \geq 0, \beta \geq 0} D(\alpha, \beta) = \frac{1}{2}(w^*)^T w^* + C \sum_{i=1}^{N} \xi_i^* + \sum_{i=1}^{N} \alpha_i (1 - \xi_i^* - y_i((w^*)^T x_i)) + \sum_{i=1}^{N} \beta_i (-\xi_i^*)$$

Since our dual is independent of the value of ξ, we can set $\xi_i = 0 \, \forall i$. Hence we get rid of β and obtain,

$$\max_{\alpha \geq 0} D(\alpha) = \frac{1}{2}(w^*)^T w^* + \sum_{i=1}^{N} \alpha_i (1 - y_i((w^*)^T x_i))$$

When we plug $w^* = \sum_{i=1}^{N} \alpha_i y_i x_i$, we should get the same (given in the problem) hard-margin SVM dual form,

$$\max_{\alpha \geq 0} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

Since our function is in terms of α we need to write the correct constraints for it. Observe that,

$$\alpha_i = C - \beta_i, \quad \alpha_i, \beta_i \geq 0$$

Then,

$$C - \beta_i \leq C \implies \alpha_i \leq C$$

Hence our dual program for soft-margin SVM is,

$$\boxed{\max_{C \geq \alpha \geq 0} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i^T x_j), \quad C \geq \alpha_i \geq 0 \, \forall i \in [N], \quad \sum_{i=1}^{N} \alpha_i y_i = 0.}$$

2 SVM, RBF Kernel and Nearest Neighbor: 6pts

1. For hard-margin SVM we know that,

$$\hat{w} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i$$

The transpose form is,

$$\hat{w}^T = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i^T$$

Given new x then the prediction on x is,

$$\boxed{f(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i^T x}$$

2. Given the optimal $\hat{\alpha}$, we can use the equation from part 1 to get the prediction on $\mathbf{x}$ using the RBF Kernel,

$$f_\sigma(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i^T x$$

Using kernel trick to replace expression for $x_i^T x$,

$$x_i^T x = \kappa(x_i, x) = \exp(-\frac{||x_i - x||_2^2}{2\sigma^2})$$

Hence,

$$\boxed{f_\sigma(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i \exp(-\frac{||x_i - x||_2^2}{2\sigma^2})}$$

3. To prove the given equation we can start from the L.H.S. and derive the R.H.S. Let's rewrite the L.H.S using part ii,

$$\lim_{\sigma \to 0} \frac{f_\sigma(x)}{\exp(-\frac{\rho^2}{2\sigma^2})} = \lim_{\sigma \to 0} \frac{\sum_{i=1}^{N} \hat{\alpha}_i y_i \exp(-\frac{||x_i - x||_2^2}{2\sigma^2})}{\exp(-\frac{\rho^2}{2\sigma^2})}$$

Simplify by using laws of exponents,

$$\lim_{\sigma \to 0} \sum_{i=1}^{N} \hat{\alpha}_i y_i \exp(\frac{\rho^2 - ||x_i - x||_2^2}{2\sigma^2})$$

We are given that $S \subset \{1, 2, 3...n\}$ and we also observe that $T \subset S$.
Using this we can rewrite our summation as,

$$\lim_{\sigma \to 0} \sum_{i \in S} \hat{\alpha}_i y_i \exp(\frac{\rho^2 - ||x_i - x||_2^2}{2\sigma^2})$$

We can now split up our summation,

$$\lim_{\sigma \to 0} \sum_{i \in T} \hat{\alpha}_i y_i \exp(\frac{\rho^2 - ||x_i - x||_2^2}{2\sigma^2}) + \lim_{\sigma \to 0} \sum_{i \in S \setminus T} \hat{\alpha}_i y_i \exp(\frac{\rho^2 - ||x_i - x||_2^2}{2\sigma^2})$$

Let's look at the definition of ρ,

$$\rho = \min_{i \in S} ||x - x_i||_2$$

Since l_2 norms are lower bounded by 0 ($|.|_2 \geq 0$), we can equivalently state,

$$\rho^2 = \min_{i \in S} ||x - x_i||_2^2$$

Then we can redefine our set T as,

$$T = \{i \in S \, : ||x - x_i||_2^2 = \rho^2\}$$

Consider any arbitrary $j \in T$. Then this must hold,

$$||x - x_j||_2^2 = \rho^2 \implies \rho^2 - ||x - x_j||_2^2 = 0$$

Now consider an arbitrary $k \in S \setminus T$. Then this must hold,

$$||x - x_k||_2^2 > \min_{i \in S} ||x - x_i||_2^2 = \rho^2 \implies \rho^2 - ||x - x_j||_2^2 < 0$$

Rewriting our summation as,

$$\sum_{i \in T} \hat{\alpha}_i y_i \exp(\lim_{\sigma \to 0} \frac{0}{2\sigma^2}) + \sum_{i \in S \setminus T} \hat{\alpha}_i y_i \exp(\lim_{\sigma \to 0} \frac{\alpha}{2\sigma^2}), \quad \alpha < 0$$

Let's evaluate the limits separately,

$$\lim_{\sigma \to 0} \frac{0}{2\sigma^2} = 0$$

$$\lim_{\sigma \to 0} \frac{\alpha}{2\sigma^2} = \lim_{\kappa \to -\infty} \kappa, \quad \alpha < 0$$

Hence our summation is,

$$\sum_{i \in T} \hat{\alpha}_i y_i \exp(0) + \sum_{i \in S \setminus T} \hat{\alpha}_i y_i \exp(\lim_{\kappa \to -\infty} \kappa)$$

Since α_i, y_i are constants with respect to κ then,

$$\alpha_i y_i \exp(\lim_{\kappa \to -\infty} \kappa) = 0$$

Hence we obtain,

$$\sum_{i \in T} \hat{\alpha}_i y_i$$

Thus we have shown that,

$$\boxed{\lim_{\sigma \to 0} \frac{f_\sigma(x)}{\exp(-\frac{\rho^2}{2\sigma^2})} = \sum_{i \in T} \hat{\alpha}_i y_i}$$

3 Decision Tree and Adaboost: 12 pts

1. $\mathcal{D}$ contains three data points with label +1 and three data points with label −1. Hence, we calculate the sample entropy as follows,

$$I(\mathcal{D}) = -\sum_{c=1}^{C} p(c|\mathcal{D}) \log_2(p(c|\mathcal{D}))$$

$$I(\mathcal{D}) = -(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})$$

$$I(\mathcal{D}) = -\log_2(\frac{1}{2})$$

$$\boxed{I(\mathcal{D}) = 1}$$

2. The formula for information gain is,

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - \sum_{j=1}^{N} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j)$$

Let our split rule be,

$$x_1 \leq 4 \implies Choose - 1.$$

Based on this rule we will have the following two datasets,

$$\mathcal{D}_1 = \{[1,2]^T, [2,1]^T, [3,4]^T, [4,6]^T\}$$

$$\mathcal{D}_2 = \{[5,3]^T, [6,5]^T\}$$

For $\mathcal{D}_1$ the points $[1,2]^T, [3,4]^T, [4,6]^T$ are green whereas the point $[2,1]^T$ is blue. Similarly for $\mathcal{D}_2$ the points $[5,3]^T, [6,5]^T$ are blue whereas there is no such data point that is green.
Using this information, let's now calculate the respective sample entropies for our new datasets,

$$I(\mathcal{D}_1) = -\sum_{c=1}^{C} p(c|\mathcal{D}_1) \log_2(p(c|\mathcal{D}_1)) = -(\frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4})) \approx 0.811$$

$$I(\mathcal{D}_2) = -\sum_{c=1}^{C} p(c|\mathcal{D}_2) \log_2(p(c|\mathcal{D}_2)) = -(\frac{2}{2}\log_2(\frac{2}{2}) + 0) = 0$$

Let's now calculate the information gain,

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - \frac{|\mathcal{D}_1|}{|\mathcal{D}|}I(\mathcal{D}_1) - \frac{|\mathcal{D}_2|}{|\mathcal{D}|}I(\mathcal{D}_2)$$

$$IG(\mathcal{D}, f) = 1 - \frac{4}{6}(0.811) - \frac{2}{6}(0) \approx 0.459$$

Hence the max information gain we obtain is,

$$\boxed{IG(\mathcal{D}, f) \approx 0.459}$$

3. Now we have two datasets $\mathcal{D}_1, \mathcal{D}_2$ and for each we need a split rule. We can define the two split rules as follows,

$$\mathcal{D}_1 : \quad x_2 \leq 1 \implies Choose + 1$$

$$\mathcal{D}_2 : \quad x_1 \geq 5 \implies Choose + 1$$

Based on this our two datasets get split as follows,

$$\mathcal{D}_{1,1} = \{[1,2]^T, [3,4]^T, [4,6]^T\}$$

$$\mathcal{D}_{1,2} = \{[2,1]^T\}$$

$$\mathcal{D}_{2,1} = \{[5,3]^T, [6,5]^T\}$$

$$\mathcal{D}_{2,2} = \{\}$$

Here datasets $\mathcal{D}_{1,1}, \mathcal{D}_{2,2}$ should have datapoints that should be classified as -1 and the datasets $\mathcal{D}_{1,2}, \mathcal{D}_{2,1}$ have datapoints that should be classified as $+1$. Let's calculate the sample entropies for our 4 new datasets,

$$I(\mathcal{D}_{1,1}) = -(\frac{3}{3}\log_2(\frac{3}{3}) + 0) = 0$$

$$I(\mathcal{D}_{1,2}) = -(\frac{1}{1}\log_2(\frac{1}{1}) + 0) = 0$$

$$I(\mathcal{D}_{2,1}) = -(\frac{2}{2}\log_2(\frac{2}{2}) + 0) = 0$$

$$I(\mathcal{D}_{2,2}) = 0\log_2 0 = 0$$

Let's now calculate the information gains,

$$IG(\mathcal{D}_1, f) = I(\mathcal{D}_1) - \frac{|\mathcal{D}_{1,1}|}{|\mathcal{D}_1|} I(\mathcal{D}_{1,1}) - \frac{|\mathcal{D}_{1,2}|}{|\mathcal{D}_1|} I(\mathcal{D}_{1,2})$$

$$IG(\mathcal{D}_1, f) \approx 0.811 - \frac{3}{4} \times 0 - \frac{1}{4} \times 0 \approx 0.811$$

$$IG(\mathcal{D}_2, f) = I(\mathcal{D}_2) - \frac{|\mathcal{D}_{2,1}|}{|\mathcal{D}_2|} I(\mathcal{D}_{2,1}) - \frac{|\mathcal{D}_{2,2}|}{|\mathcal{D}_2|} I(\mathcal{D}_{2,2})$$

$$IG(\mathcal{D}_2, f) = 0$$

Hence the maximum information gains we obtain are,

$$\boxed{IG(\mathcal{D}_1, f) \approx 0.811, \quad IG(\mathcal{D}_2, f) = 0}$$

4. Let's start by writing down the formulas for the sample weight γ_t, weight error rate ϵ_t, weight of the decision stump α_t and the decision stump f_t.

$$\epsilon_t = \frac{\sum_{i:y^{(i)} \neq f_t(x^{(i)})}^{N} \gamma_t^{(i)}}{\sum_{i=1}^{N} \gamma_t^{(i)}}$$

$$\alpha_t = \frac{1}{2} \ln(\frac{1 - \epsilon_t}{\epsilon_t})$$

Update rule for γ,

$$\gamma_{t+1}^{(i)} = \frac{\gamma_t^{(i)} \exp(-\alpha_t \cdot y^{(i)} f_t(x^{(i)}))}{Z_t}, \quad \gamma_1^{(i)} = \frac{1}{N} \forall i.$$

For $t = 1$,
The value for γ_1 for $N = 6$,

$$\gamma_1 = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]^T$$

This is because the dataset is initially uniformly distributed, where all the datapoints have the same weight. We can use the split rule from part ii, to formulate f_1 using the decision stump formula,

$$f_1(x^{(i)}) = -sign(4 - x_1^{(i)}) \equiv (x_1^{(i)} < 5 \implies Choose - 1.)$$

Calculating our other weights,

$$\epsilon_1 = \frac{\frac{1}{6}}{6 \times \frac{1}{6}} = \frac{1}{6}$$

$$\alpha_1 = \frac{1}{2}\ln(\frac{1-\frac{1}{6}}{\frac{1}{6}}) = \frac{1}{2}\ln(5) \approx 0.805$$

For $t = 2$,

$$\gamma_2^{(i)} = \begin{cases} \gamma_1^{(i)} exp(-\alpha_1) = \frac{1}{6} exp(-0.805) \approx 0.0745 & \text{if } i \neq 2 \\ \gamma_1^{(i)} exp(\alpha_1) = \frac{1}{6} exp(0.805) \approx 0.373 & \text{if } i = 2 \end{cases}$$

Normalizing by using Z_2,

$$Z_2 \approx 5 \times 0.0745 + 0.373 \approx 0.745$$

$$\gamma_2 = [\frac{0.0745}{0.745}, \frac{0.373}{0.745}, \frac{0.0745}{0.745}, \frac{0.0745}{0.745}, \frac{0.0745}{0.745}, \frac{0.0745}{0.745}]^T = [\frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}]^T$$

Now the second data point $x^{(2)}$ is weighted unevenly with respect to (5 times the weight) other data points. Hence we can choose a split rule, f_2 as follows,

$$f_2(x^{(i)}) = sign(1 - x_2^{(i)}) \equiv (x_2^{(i)} < 2 \implies Choose\ +1.)$$

Calculating our new weights,

$$\epsilon_2 = \frac{\frac{1}{10} + \frac{1}{10}}{\frac{1}{2} + 5 \times \frac{1}{10}} = \frac{1}{5}$$

$$\alpha_2 = \frac{1}{2}\ln(\frac{1-\frac{1}{5}}{\frac{1}{5}}) = \frac{1}{2}\ln(4) \approx 0.693$$

Hence our final answers are,

$$\boxed{f_1(x^{(i)}) = (x_1^{(i)} < 5 \implies Choose\ -1.) \quad \gamma_1 = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]^T \quad \epsilon_1 = \frac{1}{6} \quad \alpha_1 \approx 0.805}$$

$$\boxed{f_2(x^{(i)}) = (x_2^{(i)} < 2 \implies Choose\ +1.) \quad \gamma_2 = [\frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}]^T \quad \epsilon_2 = \frac{1}{5} \quad \alpha_2 \approx 0.693}$$

5. Define our final classifier $f(x)$ as,

$$f(x) = sign(\frac{\sum_{t=1}^{2} \alpha_t f_t(x)}{\sum_{t=1}^{2} |\alpha_t|})$$

$$f(x) = sign(\frac{0.805 f_1(x) + 0.693 f_2(x)}{0.805 + 0.693})$$

$$f(x) = sign(0.537 f_1(x) + 0.463 f_2(x))$$

Let's classify our data points using this classifier f,

$$\boxed{f([1,2]^T) = sign(-1) = -1 \implies Correct!}$$

$$\boxed{f([2,1]^T) = sign(-0.537 + 0.463) = -1 \implies Incorrect :(}$$

$$\boxed{f([3,4]^T) = sign(-1) = -1 \implies Correct!}$$

$$\boxed{f([4,6]^T) = sign(-1) = -1 \implies Correct!}$$

$$\boxed{f([5,3]^T) = sign(0.537 - 0.463) = 1 \implies Correct!}$$

$$\boxed{f([6,5]^T) = sign(0.537 - 0.463) = 1 \implies Correct!}$$

4 Learning Theory: 14pts

1. To solve this, we make use of Hoeffding's Inequality.
   It states that for independent RVs $Z_1, Z_2, ...Z_n$ where $Z_i \in \{a, b\}$, $\hat{Z}_n = \frac{1}{n}\sum_{i=1}^{N} Z_i$, $\exists\epsilon$ such that,
   $$Pr(|\hat{Z}_n - \mathbb{E}[\hat{Z}_n]| \geq \epsilon) \leq 2\exp(-\frac{2n\epsilon^2}{(b-a)^2})$$
   This also gives us a lower bound,
   $$Pr(|\hat{Z}_n - \mathbb{E}[\hat{Z}_n]| \leq \epsilon) \geq 1 - 2\exp(-\frac{2n\epsilon^2}{(b-a)^2})$$
   Now we can apply this to our problem to determine n.
   Here $a = 0, b = 1, \epsilon = 0.05$. Hence,
   $$Pr(|p - \hat{p}| \leq 0.05) \geq 1 - 2\exp(-\frac{2n(0.05)^2}{(1-0)^2}) = 0.95$$
   Solving for n,
   $$1 - 2 \times exp(\frac{-n}{200}) = 0.95 \implies \exp(\frac{-n}{200}) = 0.025$$
   $$n = ln(0.025) \times -200 \approx 737.7$$
   $$\boxed{n = 738}$$

2. For all parts please assume $\mathcal{Y} = \{-1, 1\}$, replacing what's given in the questions.
   (a)
   For $\mathcal{X} = \mathbb{R}$, I conjecture that,
   $$VC(\mathcal{F}_{affine}) = 2$$
   To prove that, I can find a dataset $\mathcal{D} = \{4, 6\}$, $|\mathcal{D}| = 2$. that can be shattered by $\mathcal{F}$.
   We can approach this with a case-by-case basis.
   Case 1:
   $$y_1 = +1, y_2 = +1$$
   Let,
   $$w = 1, w_0 = -4$$
   Then,
   $$\mathbb{1}\{1 \times 4 - 4\} = +1$$
   $$\mathbb{1}\{1 \times 6 - 4\} = +1$$

Case 2:

$$y_1 = -1, y_2 = -1$$

Let,

$$w = -1, w_0 = 3$$

Then,

$$\mathbb{1}\{-1 \times 4 + 3\} = -1$$

$$\mathbb{1}\{-1 \times 6 + 3\} = -1$$

Case 3:

$$y_1 = -1, y_2 = +1$$

Let,

$$w = 1, w_0 = -5$$

Then,

$$\mathbb{1}\{1 \times 4 - 5\} = -1$$

$$\mathbb{1}\{1 \times 6 - 5\} = +1$$

Case 4:

$$y_1 = +1, y_2 = -1$$

Let,

$$w = -1, w_0 = 5$$

Then,

$$\mathbb{1}\{-1 \times 4 + 5\} = +1$$

$$\mathbb{1}\{-1 \times 6 + 5\} = -1$$

So we have shown that $\mathcal{D}$, $|\mathcal{D}| = 2$ is shattered by $\mathcal{F}$.
Now consider $\hat{\mathcal{D}}$, $|\hat{\mathcal{D}}| = 3$. Let our points be,

$$\hat{\mathcal{D}} = \{\alpha, \beta, \gamma\}, \quad \alpha \leq \beta \leq \gamma.$$

To show that $\hat{\mathcal{D}}$ is valid we can consider all 2^3 label permutations. However, we note for 2 such label permutations,

$$\{y_\alpha, y_\beta, y_\gamma\} = \{\{1, -1, 1\}, \{-1, 1, -1\}\}$$

There exist no pairs $(w, w_0) \in \mathbb{R}^2$ to attain these two label permutations. This comes from the notion that the three points are collinear (lie on the same line) and any classifier that tries to classify both α and γ both as the same label is forced to classify

β as the same label too and vice versa.
Hence we have shown that,

$$VC(\mathcal{F}_{affine}) < 3$$

Thus,

$$\boxed{VC(\mathcal{F}_{affine}) = 2}$$

(b)
For $\mathcal{X} = \mathbb{R}^k$, I conjecture that,

$$VC(\mathcal{F}^k_{affine}) = k + 1$$

To prove this, I can find a dataset $\mathcal{D}$ such that ($|\mathcal{D}| = k + 1$)

$$\mathcal{D} = \{x^{(1)}, x^{(2)}, ..., x^{(k+1)}\} \quad x^{(i)} \in \mathbb{R}^k,\ x^{(i)} = \mathbf{e}_i\ \forall i < k + 1, x^{(k+1)} = [0]$$

Here, $\mathbf{e}_i \in \mathbb{R}^k$ is the i^{th} standard basis vector where all entries are 0 except the i^{th} entry which is 1, and [0] is the zero-vector in $\mathbb{R}^k$.
Now, consider the data matrix $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} (x^{(1)})^T & 1 \\ (x^{(2)})^T & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ (x^{(k)})^T & 1 \\ ([0])^T & 1 \end{bmatrix} \quad \mathbf{X} \in \mathbb{R}^{(k+1)\times(k+1)}.$$

We define our column vector $w^* \in \mathbb{R}^{k+1}$,

$$w^* = \begin{bmatrix} w \\ w_0 \end{bmatrix} \quad w \in \mathbb{R}^k,\ w_0 \in \mathbb{R}$$

Hence we can define our label class $y \in \{-1, 1\}^{k+1}$,

$$\mathbb{1}\{\mathbf{X}w^*\} = y$$

Which can be thought of as a row-wise indicator function operation on the result of $\mathbf{X}w^*$.
Now, to show that we can find a (w, w_0) pair for all 2^{k+1} label permutations we can solve the system of equations for $\hat{y} \in \mathbb{R}^{k+1}$ such that $\mathbb{1}\{\hat{y}\} = y$ to determine $\hat{w}$, i.e,

$$\mathbf{X}\hat{w} = \hat{y}$$

We can directly solve for,

$$\hat{w}_0 = \hat{y}_{k+1}$$

Consequently,

$$\hat{w}_i + \hat{w}_0 = \hat{y}_i \implies \hat{w}_i = \hat{y}_i - \hat{w}_0 = \hat{y}_i - \hat{y}_{k+1} \, \forall i \leq k.$$

Hence,

$$\hat{w} = \begin{bmatrix} \hat{y}_1 - \hat{y}_{k+1} \\ \hat{y}_2 - \hat{y}_{k+1} \\ \vdots \\ \hat{y}_k - \hat{y}_{k+1} \\ \hat{y}_{k+1} \end{bmatrix}$$

As a consequence, we can find any corresponding $\hat{w}$ for a given $\hat{y}$ which satisfies our label permutation since $\hat{w} \in \mathbb{R}^{k+1}$.
Hence, we have shown that $\mathcal{D}$ can be shattered by $\mathcal{F}$.
Now consider $\bar{\mathcal{D}}$, $|\bar{\mathcal{D}}| = k+2$. We have the data matrix $\bar{X}$,

$$\bar{X} = \begin{bmatrix} (x^{(1)})^T & 1 \\ (x^{(2)})^T & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ (x^{(k+1)})^T & 1 \\ (x^{(k+2)})^T & 1 \end{bmatrix} \quad \bar{X} \in \mathbb{R}^{(k+2)\times(k+1)}.$$

For $\bar{w} \in \mathbb{R}^{k+1}$ we try to solve the system,

$$\bar{X}\bar{w} = \bar{y}, \quad \bar{y} \in \mathbb{R}^{k+2}$$

Observe that the rows of $\bar{X}$ can be thought of as vectors in $\mathbb{R}^{k+1}$ where the last coordinate is 1. Since we have $k+2$ such vectors in $\mathbb{R}^{k+1}$ then we can have at most $k+1$ independent vectors (by definition of basis dimensionality) $\implies \geq 1$ dependent row vector/s in $\bar{X}$.
Suppose we have an arbitrary row vector $x^{(j)}$ that is the linear combination of other row vectors in $\bar{X}$, i.e,

$$x^{(j)} = \sum_{\{i:i=1, i\neq j\}}^{k+2} \alpha_i x^{(i)},$$

Since $\bar{w} \in \mathbb{R}^{k+1}$ we can write our $\alpha - i$ in terms of $\bar{w}$,

$$\alpha_i = \bar{w}^T x^{(i)}$$

Note that we can always expand our functions and definitions to make the bias term w_0 implicit. Hence we assume from now on that we have expanded the function and the bias term is implicit (alerady considered). Then, we can fix $y_j = -1$ and $y_i = sign(\alpha_i)$. Thus ($\bar{W}$ is adjusted version of $\bar{w}$ for bias),

$$\bar{W}^T x^{(j)} \equiv \sum_{\{i:i=1, i\neq j\}}^{k+2} (\alpha_i \bar{W}^T x^{(i)})$$

Substitute expression for α_i,

$$\sum_{\{i:i=1, i\neq j\}}^{k+2} (\bar{W}^T x^{(i)})^2 \implies \bar{W}^T x^{(j)} \geq 0$$

This means that,

$$\mathbb{1}\{\bar{W}^T x^{(j)}\} = +1 \neq y_j$$

This is a contradiction since we have chosen our original value of y_j to be -1. This implies that we have found a label permutation that cannot be classified correctly by any $\bar{W}$ which means that,

$$VC(\mathcal{F}^k_{affine}) < k + 2$$

Hence,

$$\boxed{VC(\mathcal{F}^k_{affine}) = k + 1}$$

(c)
Since a cosine is just a shifted sine, if we can show and prove what the $VC\{\mathcal{F}_{sin}\}$ is then we can state the same for $VC\{\mathcal{F}_{cos}\}$.

Let's choose a dataset $\mathcal{D} := \{(2\pi 10^{-i}, y_i)\}_{i=1}^n$ and define the parameter ω,

$$\omega = \frac{1}{2}\left(1 + \sum_{i=1}^{n} \frac{1 - y_i}{2} 10^i\right).$$

Case 1:

We look at data points in $\mathcal{D}$ that have negative labels. The parameter is now,

$$\omega = \frac{1}{2}(1 + \sum_{\{i:y_i=-1\}} 10^i)$$

We note that, for any point $x_j = 2\pi 10^{-j}$ in the data set such that $y_j = -1$, the term 10^j appears in the sum. This means that,

$$\omega x_j = \pi 10^{-j}\left(1 + \sum_{\{i:y_i=-1\}} 10^i\right)$$

$$= \pi 10^{-j}\left(1 + 10^j + \sum_{\{i:y_i=-1, i\neq j\}} 10^i\right)$$

$$= \pi\left(10^{-j} + 1 + \sum_{\{i:y_i=-1, i<j\}} 10^{i-j} + \sum_{\{i:y_i=-1, i>j\}} 10^{i-j}\right)$$

This means that 10^{i-j} is even for $i > j$ and can be written as $2\alpha_i$ for some $\alpha_i \in \mathbb{N}$. Define $\alpha \in \mathbb{N}$ (which is also even) as,

$$\alpha = \left(\sum_{\{i:y_i=-1, i>j\}} 2\alpha_i\right)$$

Consider the $i < j$ sum,

$$\sum_{\{i:y_i=-1, i<j\}} 10^{i-j}$$

It's upper-bounded as follows,

$$\sum_{\{i:y_i=-1, i<j\}} 10^{i-j} < \sum_{k=-1}^{-\infty} 10^k = \frac{1}{9}$$

Then,

$$\omega x_j = \pi(1 + 10^{-j} + \sum_{\{i:y_i=-1, i<j\}} 10^{i-j} + 2\alpha)$$

$$\omega x_j = \pi(1 + 10^{-j} + \sum_{\{i:y_i=-1,i<j\}} 10^{i-j}) + 2\pi\alpha$$

Let γ be,

$$\gamma = (10^{-j} + \sum_{\{i:y_i=-1,i<j\}} 10^{i-j})$$

Then,

$$\omega x_j = \pi(1 + \gamma) + 2\pi\alpha$$

Hence,

$$\pi < (\gamma + 1)\pi < 2\pi$$

This means that,

$$sin(\omega x_j) < 0$$

Hence, the classifier is able to predict all negative labels correctly.

Case 2: For the positive label case we have,

$$\omega x_j = 2\pi 10^{-j} \times \frac{1}{2}(1 + \sum_{\{i:y_i=-1,i\neq j\}} 10^{i})$$

$$\omega x_j = \pi \left(10^{-j} + \sum_{\{i:y_i=-1,i<j\}} 10^{i-j} + \sum_{\{i:y_i=-1,i>j\}} 10^{i-j}\right)$$

We can use same γ and α from Case 1,

$$\omega x_j = \pi\gamma + 2\pi\alpha$$

Then,

$$sin(\omega x_j) > 0$$

Since,

$$0 < \pi\gamma < \pi$$

Hence, all positive labels are also correctly classified by our particular ω value. Since $n \to +\infty$ for our original dataset $\mathcal{D}$ then, we have shown that our $\mathcal{F}$ can shatter a dataset of any size $n \in \mathbb{N}$. This means that,

$$VC\{\mathcal{F}_{sin}\} = +\infty$$

By using the transformation formula,

$$sin(\omega x + \frac{\pi}{2}) = sin(\omega x)cos(\frac{\pi}{2}) + sin(\frac{\pi}{2})cos(\omega x) = cos(\omega x)$$

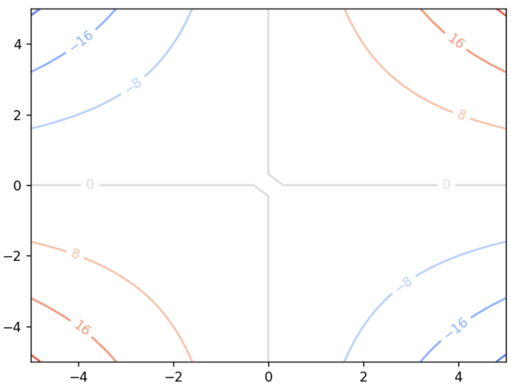
This shows that,

$$\boxed{VC\{\mathcal{F}_{cos}\} = +\infty}$$

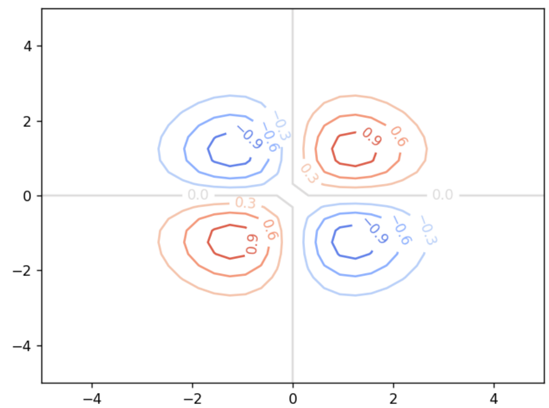
5 Coding: SVM, 24pts

3.

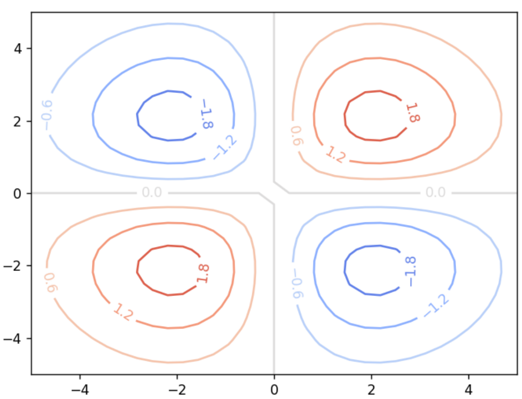
Table 1: Four Contour Plots



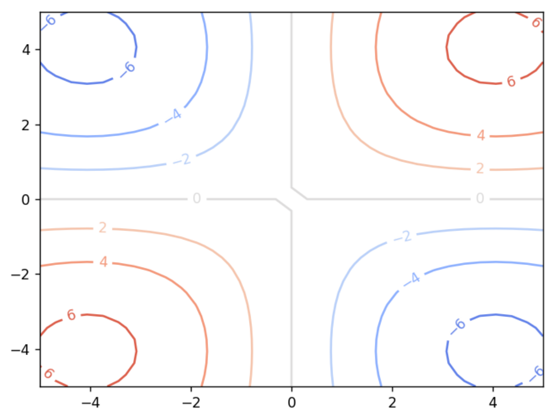
Plot 1: Polynomial Kernel: $Degree = 2$



Plot 2: RBF Kernel: $\sigma = 1$



Plot 3: RBF Kernel: $\sigma = 2$



Plot 4: RBF Kernel: $\sigma = 4$