0 Instructions

Homework is due Tuesday, 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

The Lagrangian form of the soft-margin SVM is given by

$$L(\omega, b, \xi, \alpha, \beta) = \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\omega^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i$$

The dual form of the problem is then given by

$$D(\alpha, \beta) = \min_{\omega,b,\xi} L(\omega, b, \xi, \alpha, \beta)$$

Because the problem is convex, we know that the maximum of the dual is the minimum of the primal. The solution to the dual occurs when the gradients of the Lagrangian are 0, i.e.

$$\begin{aligned}
\nabla_\omega L(\omega, b, \xi, \alpha, \beta) &= \omega - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \\
\nabla_b L(\omega, b, \xi, \alpha, \beta) &= -\sum_{i=1}^{n}\alpha_i y_i = 0 \\
\nabla_\xi L(\omega, b, \xi, \alpha, \beta) &= C - \alpha_i - \beta_i = 0
\end{aligned}$$

Substitue these back into the Lagrangian, we have

$$\begin{aligned}
D(\alpha, \beta) &= \frac{1}{2}||\sum_{i=1}^{n}\alpha_i y_i x_i||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\sum_{j=1}^{n}\alpha_j y_j x_j^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j
\end{aligned}$$

subject to

$$\begin{aligned} \alpha_i &\geq 0 \\ \beta_i &\geq 0 \\ \alpha_i + \beta_i &= C \\ \sum_{i=1}^{n} \alpha_i y_i &= 0 \end{aligned}$$

After eliminating β_i, we have the following constraints,

$$\begin{aligned} 0 \leq \alpha_i &\leq C \\ \sum_{i=1}^{n} \alpha_i y_i &= 0 \end{aligned}$$

2 SVM, RBF Kernel and Nearest Neighbor: 6pts

2.1

$$\begin{aligned} \hat{\omega} &= \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \\ f(x) &= (\sum_{i=1}^{N} \hat{\alpha}_i y_i x_i)^T x \end{aligned}$$

2.2

$$\begin{aligned} \hat{\omega} &= \sum_{i=1}^{N} \hat{\alpha}_i y_i \phi(x_i) \\ f(x) &= (\sum_{i=1}^{N} \hat{\alpha}_i y_i \phi(x_i))^T \phi(x) \\ &= \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x_i, x) \end{aligned}$$

2.3

$$\begin{aligned} \lim_{\delta \to 0} \frac{\sum_{i=1}^{N} \hat{\alpha} y_i e^{-\frac{||x_i - x||^2}{2\delta^2}}}{e^{-\frac{\rho^2}{2\delta^2}}} &= \lim_{\delta \to 0} \sum_{i=1}^{S} \hat{\alpha} y_i e^{-\frac{||x_i - x||^2 - \rho^2}{2\delta^2}} \\ &= \lim_{\delta \to 0} \sum_{i=1}^{T} \hat{\alpha} y_i e^{-\frac{||x_i - x||^2 - \rho^2}{2\delta^2}} + \lim_{\delta \to 0} \sum_{i=1}^{S/T} \hat{\alpha} y_i e^{-\frac{||x_i - x||^2 - \rho^2}{2\delta^2}} \\ &= \sum_{i=1}^{T} \hat{\alpha} y_i + 0 \end{aligned}$$

3 Decision Tree and Adaboost: 12 pts

3.1

From the definition of entropy, we have

$$\begin{aligned} I(D) &= -\sum_{c=1}^{C} p(c|D) \log_2 p(c|D) \\ &= -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2} \\ &= 1 \end{aligned}$$

3.2

First consider split on x_1. Split points between two adjacent points can be considered as same, so we have possible split points at $\tau = 2, 3, 4, 5, 6$. For each τ, we split the dataset following the rule $y = 1$ if $x \geq \tau$, otherwise $y = -1$.

$I(D|f_{\tau=2}) = 0.809, I(D|f_{\tau=3}) = 1, I(D|f_{\tau=4}) = 0.918, I(D|f_{\tau=5}) = 0.541, I(D|f_{\tau=6}) = 0.809$

So the best split for x_1 is at $\tau = 5$.

Similarly, we have possible split points at $\tau = 2, 3, 4, 5, 6$ for x_2, following the rule $y = -1$ if $x \geq \tau$, otherwise $y = 1$.

$I(D|f_{\tau=2}) = 0.809, I(D|f_{\tau=3}) = 1, I(D|f_{\tau=4}) = 0.918, I(D|f_{\tau=5}) = 1, I(D|f_{\tau=6}) = 0.809$

So the best split for x_2 is at $\tau = 2$ and $\tau = 6$. Since the information gain split on x_1 is higher, we choose to split on x_1 at $\tau = 5$.

$$IG(D, f) = I(D) - I(D|f) = 0.459$$

3.3

After split on $\tau = 5$, from the figure we can know that only split on x_2 at $\tau = 2$ can further reduce the entropy. So the best split for the second level is at $\tau = 2$, following the rule $y = -1$ if $x \geq \tau$, otherwise $y = 1$.

$$\begin{aligned} I(D|f_{x_1=5,x_2=2}) &= \frac{1}{6}0 + \frac{1}{2}0 + \frac{1}{3}0 = 0 \\ IG(D|f_{x_1=5}, f_{x_2=2}) &= I(D|f_{x_1=5}) - I(D|f_{x_1=5,x_2=2}) = 0.541 \end{aligned}$$

So, after the second split, the further information gain is 0.541. The total information gain is $0.459 + 0.541 = 1$.

3.4

Follow the Adaboost routine, at $t = 1$, we have,

$$\begin{aligned}
\gamma_1^i &= \frac{1}{6} \\
f_1(x) &= \begin{cases} 1, & x_1 \geq 5 \\ -1, & x_1 < 5 \end{cases} \\
z_1 &= \sum_{i=1}^{6} \frac{1}{6} y^i f_1(x_i) = \frac{4}{6} \\
\alpha_1 &= \frac{1}{2} \ln \frac{1+z_1}{1-z_1} = \frac{1}{2} \ln 5 \\
\gamma_2^2 &= \frac{1}{6Z_1} e^{-\frac{1}{2}\ln 5(-1)} \\
\gamma_2^i &= \frac{1}{6Z_1} e^{-\frac{1}{2}\ln 5}, i \neq 2
\end{aligned}$$

After the normalization, we have $\gamma_2^2 = \frac{1}{2}$, $\gamma_2^i = \frac{1}{10}, i \neq 2$. Thus, at $t = 2$, we have,

$$\begin{aligned}
f_2(x) &= \begin{cases} -1, & x_2 \geq 4 \\ 1, & x_2 < 4 \end{cases} \\
z_2 &= \sum_{i=1}^{6} \gamma_2^i y^i f_2(x_i) = \frac{3}{5} \\
\alpha_2 &= \frac{1}{2} \ln \frac{1+z_2}{1-z_2} = 0.6931
\end{aligned}$$

3.5

The final classifier is given by

$$f(x) = sign(0.8047 * f_1(x) + 0.6931 * f_2(x))$$

$$where \quad f_1(x) = \begin{cases} 1, & x_1 \geq 5 \\ -1, & x_1 < 5 \end{cases}$$

$$f_2(x) = \begin{cases} -1, & x_2 \geq 4 \\ 1, & x_2 < 4 \end{cases}$$

$$\begin{aligned}
f(x_1) &= sign(0.8047 * (-1) + 0.6931 * (1)) = -1 \quad \text{correct} \\
f(x_2) &= sign(0.8047 * (-1) + 0.6931 * (1)) = -1 \quad \text{incorrect} \\
f(x_3) &= sign(0.8047 * (-1) + 0.6931 * (-1)) = -1 \quad \text{correct} \\
f(x_4) &= sign(0.8047 * (-1) + 0.6931 * (-1)) = -1 \quad \text{correct} \\
f(x_5) &= sign(0.8047 * (1) + 0.6931 * (1)) = 1 \quad \text{correct} \\
f(x_6) &= sign(0.8047 * (1) + 0.6931 * (-1)) = 1 \quad \text{correct}
\end{aligned}$$

4 Learning Theory: 14pts

4.1

From Hoeffding's inequality, with probability at least 0.95, we have

$$\epsilon_\mu(h) \leq \epsilon_{\mathcal{D}}(h) + \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

Since, we want to have $\epsilon_\mu(h) - \epsilon_{\mathcal{D}}(h) \leq 0.05$, we have,

$$\begin{aligned}\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} &\leq 0.05 \\ n &\geq 200 \log 40 \\ n &\geq 738\end{aligned}$$

4.2

4.2.1

$\mathsf{VC}(\mathcal{F}_{\text{affine}}) = 3$.

We can image this function as a line in 2D space, and the objective is to seperate the points into two classes. Points above the line are labeled as 1, and points below the line are labeled as -1. For 3 points not in the same line, we can always find a line to seperate them.

On the other hand, for any dataset containing 4 points, if there are 3 points in a line, they cannot be seperated. Otherwise, there are at least 4 points forming in a XOR pattern so that we cannot seperate as well. So the VC dimension of $\mathcal{F}_{affine}$ is 3.

4.2.2

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

Rewrite the definition of $\mathcal{F}^k_{\text{affine}}$ as $\mathcal{F}^k_{\text{affine}} = \{\mathbb{I}\{\begin{bmatrix} x^T & 1 \end{bmatrix} \begin{bmatrix} \omega \\ \omega_0 \end{bmatrix}\}\}$. The first matrix $\mathbf{X}$ is of $\mathbb{R}^{n\times(k+1)}$, the second matrix $\mathbf{W}$ is of $\mathbb{R}^{(k+1)\times 1}$. For $n = k+1$, we know that the first matrix is of full rank, so we can always get a solution to the problem $\mathbf{XW} = \mathbf{Y}$, where $\mathbf{Y}$ is arbitrary labels for points.

On the other hand, for $n = k+2$, i.e. there are $k+2$ points in $\mathbb{R}^{k+1}$ space, thus there exists a point $x_j = \sum_{i\neq j} a_i x_i$, where not all a_i are 0. Assign label for x_j as 1 and $sign(a_i)$ for other

y_is, we know that $y_i = sign(a_i)$ is achieved by $sign(W^T x_i)$. If $a_i = 0$, the label is arbitrary. Then we have $sign(w^T x_j) = sign(\sum_{i \neq j} a_i w^T x_i)$. The signs of $w^T x_i$ is the same as a_i, thus we get $sign(w^T x_j) = y_j > 0$, which is Contradictory to our assumption. So we know that the VC dimension of $\mathcal{F}^k_{\text{affine}}$ is $k + 1$.

4.2.3

$\mathsf{VC}(\mathcal{F}^k_{\cos}) = infinite.$

5 Coding: SVM, 4pts

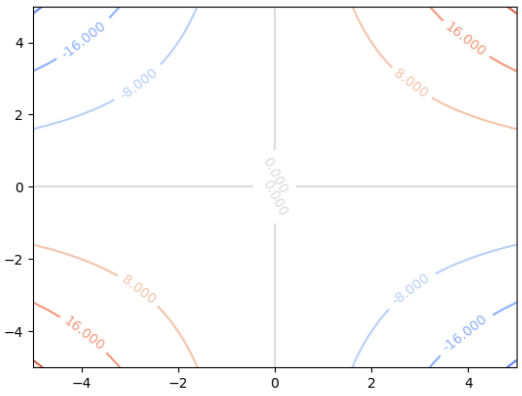

Figure 1: Results of Polynomial kernel degree of 2

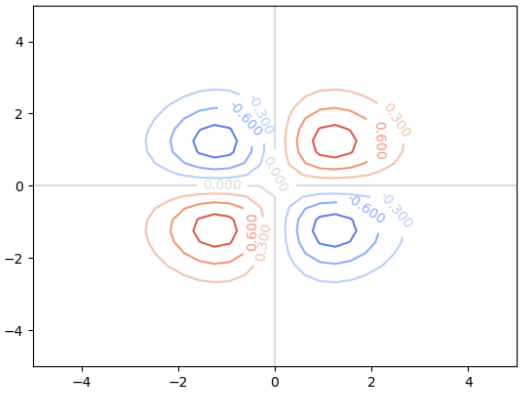

Figure 2: Results of RBF kernel with $\gamma = 1$

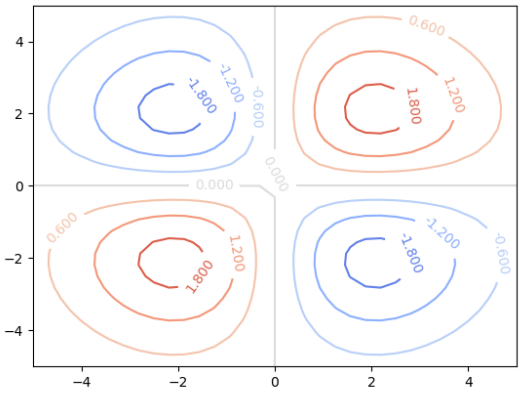

Figure 3: Results of RBF kernel with $\gamma = 2$

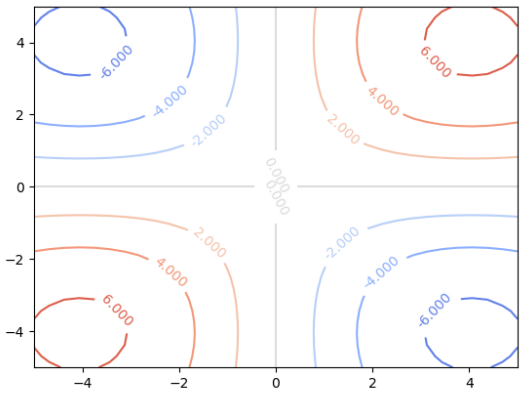

Figure 4: Results of RBF kernel with $\gamma = 4$