

## 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

Referring to hard-margin case, soft-margin SVM can be simplified in such form:

$$\min_{\mathbf{w} \in \mathbb{R}^d, \boldsymbol{\xi} \in \mathbb{R}_{\geq 0}^n} \max_{\boldsymbol{\alpha} \in \mathbb{R}_+^n, \boldsymbol{\beta} \in \mathbb{R}_+^n} \sum_{i \in [n]} \alpha_i (1 - \xi_i - y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) + \sum_{i \in [n]} -\beta_i \xi_i + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + C \sum_{i \in [n]} \xi_i$$

with constraints:

$$y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)} \geq 1 - \xi_i, \quad \xi_i \geq 0$$

and thus we have the dual problem:

$$D(\boldsymbol{\alpha}) = \min_{\mathbf{w} \in \mathbb{R}^d, \boldsymbol{\xi} \in \mathbb{R}_{\geq 0}^n} \sum_{i \in [n]} \alpha_i (1 - \xi_i - y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) + \sum_{i \in [n]} -\beta_i \xi_i + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + C \sum_{i \in [n]} \xi_i$$

For the gradient of the problem with respect to  $\mathbf{w}$ :

$$\begin{aligned} \nabla_{\mathbf{w}} \sum_{i \in [n]} \alpha_i (1 - \xi_i - y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) + \sum_{i \in [n]} -\beta_i \xi_i + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + C \sum_{i \in [n]} \xi_i &= 0 \\ \Rightarrow \mathbf{w} &= \sum_{i \in [n]} \alpha_i y^{(i)} \mathbf{x}^{(i)} \end{aligned}$$

For the gradient of the problem with respect to  $\boldsymbol{\xi}$ :

$$\begin{aligned} \nabla_{\boldsymbol{\xi}} \sum_{i \in [n]} \alpha_i (1 - \xi_i - y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) + \sum_{i \in [n]} -\beta_i \xi_i + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + C \sum_{i \in [n]} \xi_i &= 0 \\ \Rightarrow \beta_i &= C - \alpha_i \end{aligned}$$

Therefore, the dual form of soft-margin SVM will be:

$$\max_{\alpha_i, j \in (0, C)} \sum_{i \in [n]} \alpha_i - \frac{1}{2} \sum_{i, j \in [n]} \alpha_i \alpha_j y^{(i)} y^{(j)} \mathbf{x}^{(i)\top} \mathbf{x}^{(j)}$$

## 2 SVM, RBF Kernel and Nearest Neighbor: 6pts

1.

$$\begin{aligned}\hat{\mathbf{w}} &= \sum_{i \in [n]} \hat{\alpha}_i y^{(i)} \mathbf{x}^{(i)} \\ \Rightarrow f(\mathbf{x}) &= \hat{\mathbf{w}}^\top \mathbf{x} = \sum_{i \in [n]} \hat{\alpha}_i y_i \mathbf{x}_i^\top \mathbf{x}\end{aligned}$$

2.

$$f_\sigma(\mathbf{x}) = \hat{\mathbf{w}}^\top \kappa(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i \in [n]} \hat{\alpha}_i y_i \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right)$$

3.

$$\begin{aligned}f_\sigma(\mathbf{x}) &= \hat{\mathbf{w}}^\top \kappa(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i \in S} \hat{\alpha}_i y_i \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right) \\ &= \sum_{i \in T} \hat{\alpha}_i y_i \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right) + \sum_{i \in S \setminus T} \hat{\alpha}_i y_i \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right) \\ &= \sum_{i \in T} \hat{\alpha}_i y_i \exp\left(-\frac{\rho^2}{2\sigma^2}\right) + \sum_{i \in S \setminus T} \hat{\alpha}_i y_i \exp\left(-\frac{c}{2\sigma^2}\right) \\ \Rightarrow \frac{f_\sigma(\mathbf{x})}{\exp(-\rho^2/2\sigma^2)} &= \sum_{i \in T} \hat{\alpha}_i y_i + \sum_{i \in S \setminus T} \hat{\alpha}_i y_i \exp\left(\frac{\rho^2 - \|\mathbf{x}_i - \mathbf{x}\|_2^2}{2\sigma^2}\right)\end{aligned}$$

Since:

$$\forall i \in S \setminus T, \rho^2 < \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$$

we have the limit:

$$\lim_{\sigma \rightarrow 0} \frac{\rho^2 - \|\mathbf{x}_i - \mathbf{x}\|_2^2}{2\sigma^2} \rightarrow -\infty$$

hence:

$$\begin{aligned}\lim_{\sigma \rightarrow 0} \frac{f_\sigma(\mathbf{x})}{\exp(-\rho^2/2\sigma^2)} &= \lim_{\sigma \rightarrow 0} \sum_{i \in T} \hat{\alpha}_i y_i + \sum_{i \in S \setminus T} \hat{\alpha}_i y_i \exp\left(\frac{\rho^2 - \|\mathbf{x}_i - \mathbf{x}\|_2^2}{2\sigma^2}\right) \\ &= \sum_{i \in T} \hat{\alpha}_i y_i + 0 = \sum_{i \in T} \hat{\alpha}_i y_i\end{aligned}$$

### 3 Decision Tree and Adaboost: 12 pts

1.

$$\begin{aligned} I(\mathcal{D}) &= - \sum_{y \in -1, 1} p(y|\mathcal{D}) \log_2 p(y|\mathcal{D}) \\ &= -p(1|\mathcal{D}) \log_2 p(1|\mathcal{D}) - p(-1|\mathcal{D}) \log_2 p(-1|\mathcal{D}) \\ &= -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \end{aligned}$$

2. rule:  $f_1 : \text{sign}(x_1 - 4.5)$

information gain with 3 green 1 blue v.s. 2 blue:

$$\begin{aligned} IG(\mathcal{D}, f_1) &= I(\mathcal{D}) - \sum_{j=1}^2 \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j) \\ &= 1 - \frac{4}{6} I(\mathcal{D}_1) - \frac{2}{6} I(\mathcal{D}_2) = 1 - \frac{2}{3} \left( -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) - \frac{1}{3} (-\log_2 1) \\ &= \frac{1}{2} \log_2 3 - \frac{1}{3} \end{aligned}$$

3. rule (for the left node holding  $\mathcal{D}_1$ ):  $f_2 : \text{sign}(1.5 - x_2)$

In the left node, information gain with 3 green v.s. 1 blue:

$$IG(\mathcal{D}_1, f_2) = I(\mathcal{D}_1) - I(\mathcal{D}_1|f_2)$$

Since all samples are in perfect split in  $\mathcal{D}_1$  with  $f_2$ , we have:  $I(\mathcal{D}_1|f_2) = 0$ . Thus,

$$IG(\mathcal{D}_1, f_2) = I(\mathcal{D}_1) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = \frac{3}{4} \log_2 3 - 2$$

For the right node holding  $\mathcal{D}_2$ , since the first split has already been perfect for all samples, there's no need for the second split and thus the information gain of the second split equals to 0.

4. For iteration  $t = 1$ ,  $f_1 : x_j \geq 4.5$ :

$$\begin{aligned} \gamma_1 &= \left[ \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \right]^\top \\ \epsilon_1 &= \frac{\sum_{i: y^{(i)} \neq f_1(\mathbf{x}^{(i)})} \gamma_i^1}{\sum_i \gamma_i^1} = \frac{1}{6} \end{aligned}$$

$$z_1 = \sum_{i=1}^6 \gamma_1^{(i)} y^{(i)} f_1(\mathbf{x}^{(i)}) = \frac{5}{6} - \frac{1}{6} = \frac{2}{3}$$

$$\alpha_1 = \frac{1}{2} \ln \frac{1 + z_1}{1 - z_1} = \frac{1}{2} \ln 5$$

$$\gamma_2^{(i)} = \frac{\gamma_1^{(i)} \exp(-\alpha_1 y^{(i)} f_1(\mathbf{x}^{(i)}))}{Z_t}$$

with  $\gamma_1^{(i)} \exp(-\alpha_1 y^{(i)} f_1(\mathbf{x}^{(i)})) = 5^{-\frac{1}{2}}$  for  $i = 1, 3, 4, 5, 6$  and  $5^{\frac{1}{2}}$  for  $i = 2$ , while  $5^{\frac{1}{2}} = 5 \cdot 5^{-\frac{1}{2}}$ . Thus,

$$\gamma_2 = [\frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}]^\top$$

For iteration  $t = 2$ ,  $f_2 : x_j < 1.5$ :

$$\epsilon_2 = \frac{\sum_{i: y^{(i)} \neq f_2(\mathbf{x}^{(i)})} \gamma_i^2}{\sum_i \gamma_i^2} = \frac{2 \cdot \frac{1}{10}}{1} = \frac{1}{5}$$

$$z_2 = \sum_{i=1}^6 \gamma_2^{(i)} y^{(i)} f_2(\mathbf{x}^{(i)}) = \frac{1}{2} + 3 \cdot \frac{1}{10} - 2 \cdot \frac{1}{10} = \frac{3}{5}$$

$$\alpha_2 = \frac{1}{2} \ln \frac{1 + z_2}{1 - z_2} = \frac{1}{2} \ln 4$$

5. Final classifier:

$$f(\mathbf{x}) = \text{sign}(\alpha_1 f_1(\mathbf{x}) + \alpha_2 f_2(\mathbf{x}))$$

## 4 Learning Theory: 14pts

## 5 Coding: SVM, 4pts