Machine Learning, 2021 Spring
Homework 5 Solution

Due on 12:59 MAY 17, 2021

Problem 1

Definition 1 (leave-one-out cross-validation) *Select each training example in turn as the single example to be held-out, train the classifier on the basis of all the remaining training examples, test the resulting classifier on the held-out example, and count the errors.*

Let the superscript '$-i$' denote the parameters we would obtain by finding the SVM classifier f without the ith training example. Define the *leave-one-out CV error* as

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\boldsymbol{x}_i; \boldsymbol{w}^{-i}, b^{-i})),$$

where $\mathcal{L}$ is the zero-one loss. Prove that [1.5pts]

$$\textit{leave-one-out CV error} \leq \frac{\text{number of support vectors}}{\text{n}}. \tag{1}$$

Solution:
According to this problem, we assume that there are k support vectors, and there are $n-k$ non-support vectors. For non-support vectors, the *leave-one-out CV error* is 0, so we only need to consider the *leave-one-out CV error* of support vectors.
For the support vectors, we can get

$$\mathcal{L}(y_i, f(\boldsymbol{x}_i; \boldsymbol{w}^{-i}, b^{-i})) \leq 1, \quad 1 \leq i \leq k$$

In this way,

$$\sum_{i=1}^{n}\mathcal{L}(y_i, f(\boldsymbol{x}_i; \boldsymbol{w}^{-i}, b^{-i})) \leq \text{k}.$$

Therefore,

$$\textit{leave-one-out CV error} \leq \frac{\text{number of support vectors}}{\text{n}}.$$

Problem 2

The ℓ_1-norm SVM can be formulated as follows

$$\begin{aligned} \min_{(\boldsymbol{w},b)} \quad & \|\boldsymbol{w}\|_1 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad i = 1, \cdots, n. \end{aligned} \tag{2}$$

Please derive the equivalent linear programming formulation of (2). [1.5pts]

Solution:

$$\begin{aligned} \min_{(\boldsymbol{w},b)} \quad & \sum_{i=1}^{n}(\boldsymbol{w}_i^+ + \boldsymbol{w}_i^-) \\ \text{s.t.} \quad & y_i((\boldsymbol{w}^+ - \boldsymbol{w}^-)^T\boldsymbol{x}_i + b) \geq 1, \quad i = 1, \cdots, n. \\ & \boldsymbol{w}^+ \geq 0, \quad \boldsymbol{w}^- \geq 0 \end{aligned} \tag{3}$$

Problem 3

For the example in page 14 of Lecture 13, given

$$\boldsymbol{x} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix},$$

please provide the soft-margin SVM model of this problem. Derive the associated Lagrangian and the dual problem of it. [3pts]

(**Hint: the dual problem is a quadratic programming problem.**)

Solution:
The soft-margin SVM is

$$\begin{aligned} \min_{(\boldsymbol{w},b,\xi)} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{m}\xi^i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \cdots, n. \end{aligned} \tag{4}$$

And we can get

$$\begin{aligned} -b &\geq 1 - \xi_1 \\ -2w_1 - 2w_2 - b &\geq 1 - \xi_2 \\ 2w_1 + b &\geq 1 - \xi_3 \\ 3w_1 + b &\geq 1 - \xi_4 \end{aligned} \tag{5}$$

And we can get the soft-margin SVM model

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & \boldsymbol{u}^T\boldsymbol{Q}\boldsymbol{u} + \boldsymbol{p}^T\boldsymbol{u} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} \geq \boldsymbol{c}. \end{aligned} \tag{6}$$

The associated Lagrangian equation is

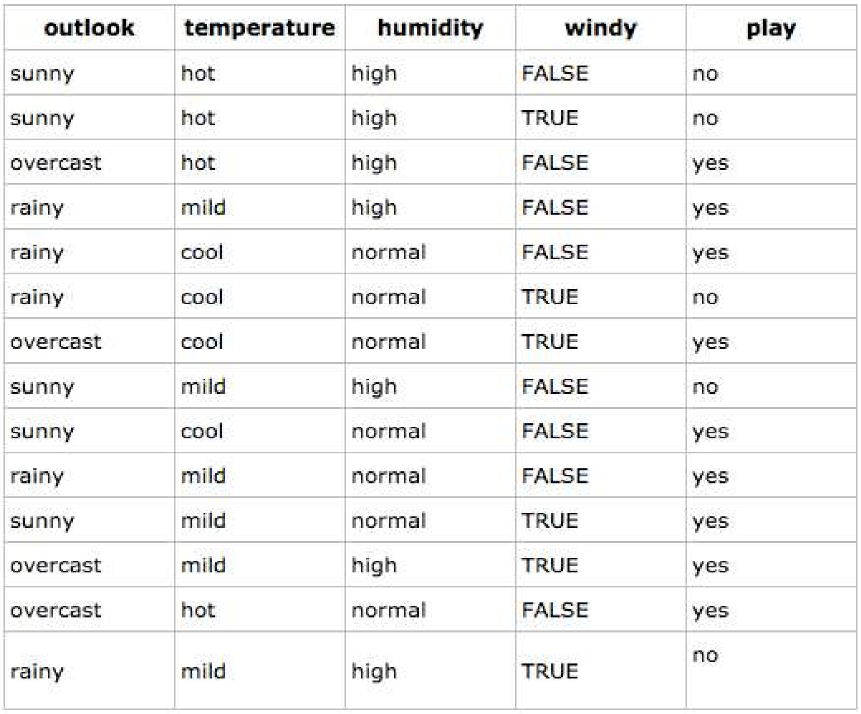
$$\begin{aligned} \mathcal{L} = \frac{1}{2}(w_1^2 + w_2^2) + c\sum_{i=1}^{4} + \lambda_1(1 - \xi_1 + b) + \lambda_2(1 - \xi_2 + b + 2w_1 + 2w_2) \\ + \lambda_3(1 - \xi_3 - b - 2w_1) + \lambda_4(1 - \xi_4 - b - 3w_1) + \sum_{i=1}^{4}\mu_i\xi_i \end{aligned} \tag{7}$$

According to the Lagrangian equation and KKT conditions, we can know the dual problem

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \sum_{i=1}^{4}\lambda_i - \frac{1}{2}(8\lambda_2^2 4\lambda_3^2 + 9\lambda_4^2 - 8\lambda_2\lambda_3 - 12\lambda_2\lambda_4 + 12\lambda_3\lambda_4) \\ \text{s.t.} \quad & 0 \leq \xi_i \leq c, \quad \lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 = 0, \quad i = 1, 2, 3, 4. \end{aligned} \tag{8}$$

Problem 4

Complete the decision trees on the following example by both ID3 and CART methods (refer to Lecture 14 for more details). [4pts]



outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Solution:

1. The decision tree by ID3 method According to Page 19-21 of Lecture 14, we can know that
   $E(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.9403.$
   $E(S, \text{outlook=sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971.$
   $E(S, \text{outlook=overcast}) = 0.$
   $E(S, \text{outlook=rainy}) = 0.971.$
   $\text{Gain(outlook)} = E(S) - \left(\frac{5}{14} \times 0.971 + 0 + \frac{5}{14} \times 0.971\right) = 0.247$

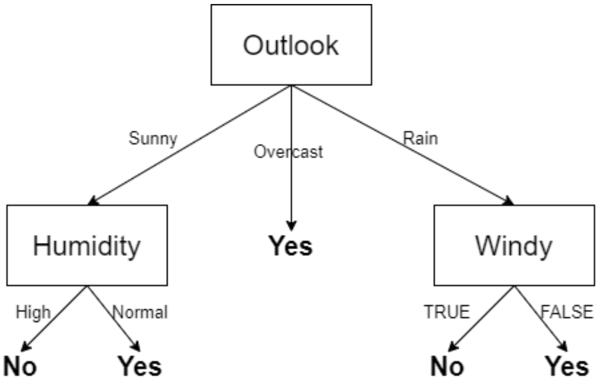
   According to Page 19-21 of Lecture 14, we can know that the gain of outlook is the biggest, then we choose sunny as node.
   $E(\text{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971.$

   $E(\text{sunny,temperature=hot}) = 0.$
   $E(\text{sunny,temperature=mild}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1.$
   $E(\text{sunny,temperature=cool}) = 0.$
   $\text{Gain(temperature)} = E(\text{sunny}) - \frac{1}{2} \times 1 = 0.571.$

   $E(\text{sunny,humidity=high}) = 0.$
   $E(\text{sunny,humidity=normal}) = 0.$
   $\text{Gain(humidity)} = E(\text{sunny}) = 0.971.$

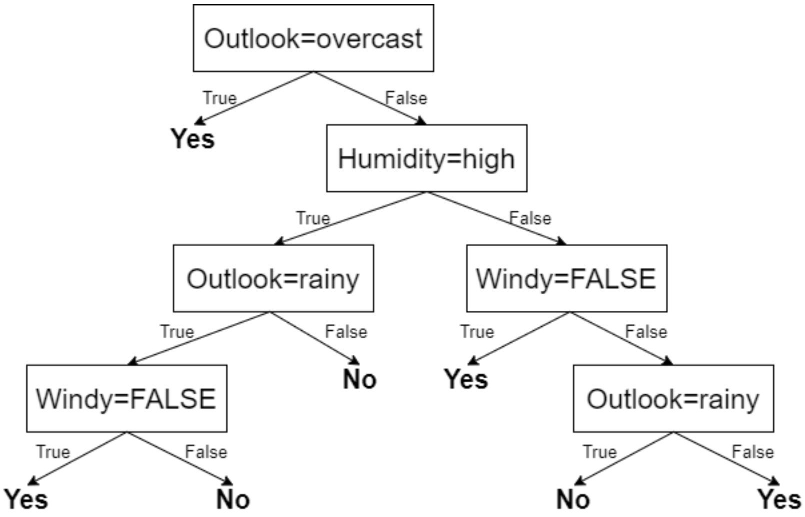
   $E(\text{sunny,windy=TRUE}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1.$
   $E(\text{sunny,windy=FALSE}) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918.$
   $\text{Gain(windy)} = E(\text{sunny}) - \frac{1}{2} \times 1 - \frac{1}{3} \times 0.918 = 0.165.$

After comparison, we choose humidity as next node. The calculation steps for Overcast and Rain are similar as those above and we will not show them again. Therefore, we can get the following decision tree.



2. The decision tree by CART method

   Similar to ID3, for CART method, we choose Gini index and choose the smallest one as the node separating condition. Since the decision tree by CART method is a binary tree, so we get the following decision tree.


Outlook=overcast
True
False
Yes
Humidity=high
True
False
Outlook=rainy
Windy=FALSE
True
False
No
True
False
Yes
Windy=FALSE
Outlook=rainy
True
False
Yes
No
True
False
No
Yes