DSAA5002 Data Mining & Data Science

(2025 FALL)

Question 1. Table 1 summarizes a data set with two attributes A and B, and two class labels + and -.

A	B	Label
T	F	+
T	T	+
T	T	+
T	T	-
T	T	+
F	T	-
F	F	-
F	F	-
T	T	-
F	F	+

Table 1. A Data Set for Question 1

(a) According to the Gini index criterion, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in Gini index.

(b) Use the remaining attribute to split in the second level of the tree. Draw the two-level decision tree. Mark the class label in each leaf node.

(c) Show the confusion matrix based on the induced decision tree, and calculate the accuracy, precision, recall and F1 measure. (Note that precision, recall and F1 measure are defined with respect to the + class)

Question 2. Derive an expression for the number of training examples sufficient to ensure that every hypothesis will have true error no worse than $(1 + \gamma)\ errorD(h)$.You can use the general Chernoff bounds to derive such a result.

Chernoff bounds: Suppose $X_1, \ldots, X_m$ are the outcomes of m independent coin flips (Bernoulli trials), where the probability of heads on any single trial is $Pr[X_i = 1] = p$ and the probability of tails is $Pr[Xi = 0] = 1 - p$. Define $S = X_1 + X_2 + \ldots + X_m$ to be the sum of the outcomes of these m trials. The expected value of S/m is $E[S/m] = p$. The Chernoff bounds govern the probability that S/m will differ from p by some factor $0 \leq \gamma \leq 1$.

$$\Pr[S/m > (1+\gamma)p] \leq e^{-mp\gamma^2/3}$$

$$\Pr[S/m < (1-\gamma)p] \leq e^{-mp\gamma^2/2}$$

Question 3. Why perceptron algorithm cannot be used to represent XOR functions?

Question 4. Figure 1 shows margin of linear separable data set in a two-dimensional space, which is defined between two margin hyperplanes. Prove margin equals $\frac{2}{\|w\|}$

(2025.10.14 Update: Replace $\frac{2}{\|w\|^2}$ with $\frac{2}{\|w\|}$)

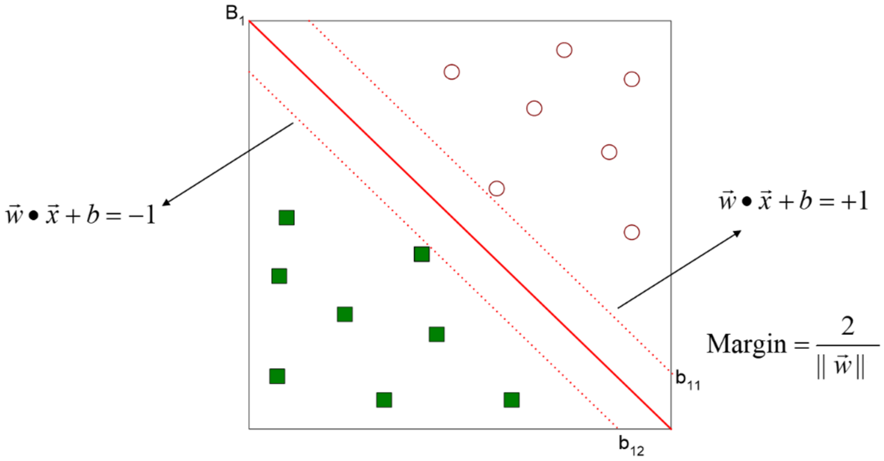
B1
$\vec{w} \bullet \vec{x} + b = -1$
$\vec{w} \bullet \vec{x} + b = +1$
Margin $= \frac{2}{\|\vec{w}\|}$
b11
b12

Figure 1 Margin between margin hyperplanes