MA2823: Introduction to Machine Learning **Homework 8**

Due December 1, 2017.

Turn in your homework

— as a PDF file

— named `HW08_<LastName>_<FirstName>.pdf` (no accents)

— at http://tinyurl.com/ma2823-2017-hw

Consider the following data set, containing 10 plants for which the length and width of sepals have been measured. The plants belong either to the *Iris virginica* (+) or the *Iris versicolor* (-) species.

label	+	+	+	+	+	+	-	-	-	-
sepal length (cm)	6.7	6.7	6.3	6.5	6.2	5.9	6.1	6.4	6.6	6.8
sepal width (cm)	3.3	3.0	2.5	3.0	3.4	3.0	2.8	2.9	3.0	2.8

Question 1

Consider the first feature (sepal length). List the possible splitting values for this feature.

Solution: The possible splitting points are: <5.9, 5.9, 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8.

Question 2

Compute the Gini Index for splitting along the first feature, at each of these points.

Solution:

$$GI(j,s) = \frac{2}{n}\left(\frac{n_l^- n_l^+}{n_l} + \frac{n_r^- n_r^+}{n_r}\right)$$
$$= \frac{2}{n}\left(\frac{(n_l - n_l^+)n_l^+}{n_l} + \frac{(n - n^+ - n_l + n_l^+)(n^+ - n_l^+)}{n - n_l}\right)$$

n =10 and n^+=6.

s	<5.9	5.9	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8
n_l	0	1	2	3	4	5	6	7	9	10
n_l^+	0	1	1	2	3	3	4	4	6	6
GI	–	0.444	0.475	0.476	0.450	0.480	0.467	0.476	**0.400**	–

Question 3

Compute the Gini Index for splitting along the second feature (sepal width), at each possible splitting point.

Solution:

s	<2.5	2.5	2.8	2.9	3.0	3.3	3.4
n_l	0	1	3	4	8	9	10
n_l^+	0	1	1	1	4	5	6
GI	–	0.444	0.419	**0.317**	0.400	0.444	–

Question 4

Draw the first node of a decision tree built on this data, using the Gini Index. Show how many instances of each class are on each of the branches.

Solution: The smallest value of GI is 0.317, obtained for j=1 (sepal width) and s=2.9, hence we'll split along that criterion.