COMP 333 — Week 10 ML Example 1

Machine Learning Example 1

The tutorial

Decision Tree

`https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/`

for Example 1 is about *decision trees*
one of the most used ML methods
for classification.

Decision trees form the basis of *random forests*
which is often the *"go-to"* method in ML.

There are also variations of decision trees
that perform both regression and classification
called CART (Classification And Regression Tree).

The article is an easy read
that gives you an introduction to how an algorithm
might learn to make decisions.

You do not need to know the theory behind decision trees.

You do not need to know the algorithm for decision trees.

You should take away from the article
the section **Coding a decision tree**
which shows Python scikit-learn in action
using the iris dataset.

Decision Trees

In a decision tree, each node of the tree
encodes one decision as a logical condition
based on the value of one feature.

Different nodes make different decisions
and may use different features.

As you follow the path from the root to a leaf,
you AND together the conditions at the nodes on the path.
At the leaf node is the classification result.

Supervised Learning

A decision tree is an example of *supervised learning.*
The dataset has a target variable, in this case *Play?*,
and the dataset contains values for the target variable,
in this case *Yes* or *No.*

The decision tree is learning to predict the target value
of an observation from the values of the features
Weather, Temperature, Humidity, Wind.

A decision tree is an example of *classification.*

In the first example, it builds a *binary* classifier
to decide one of two outcomes for each day:
Play, or not Play.

For the iris dataset, the decision tree is predicting
the target variable *class*
which has three values: *Iris Setosa, Iris Versicolour, Iris Virginica*.

For an observation, the decision tree predicts one of the three values.

It is an example of a *multi-class* classifier,
to distinguish it from a binary classifier,
because the target variable has more than two values,
and a single value is the output of the classifier.

Each of the values *Iris Setosa, Iris Versicolour, Iris Virginica*
for the target variable is called a *label*.

Sometimes, you want the classifier to predict more than one label
as a result, for example, to indicate a hybrid plant you might output
the subset $\{Iris\ Setosa, Iris\ Virginica\}$ containing two labels
to predict that the plant is a hybrid of *Iris Setosa* cross *Iris Virginica*.

Such a classifier is called a *multi-label* classifier.