應用機器學習

Brian Chan 陳醒凡

課程目標

1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

今天課堂概要

Decision Tree & Random Forest

1. The fundamental concepts of decision trees
2. The mathematics behind the decision tree learning algorithm
3. Information gain and impurity measures
4. Classification trees

CLASSIFICATION

Classification, like regression, is a predictive task, but one in which the outcome takes only values across discrete categories; classification problems are very common (arguably just as or perhaps even more common than regression problems!

Observed Data be $D_i = (x_i, y_i), i = 1, \ldots, N.$

Examples:

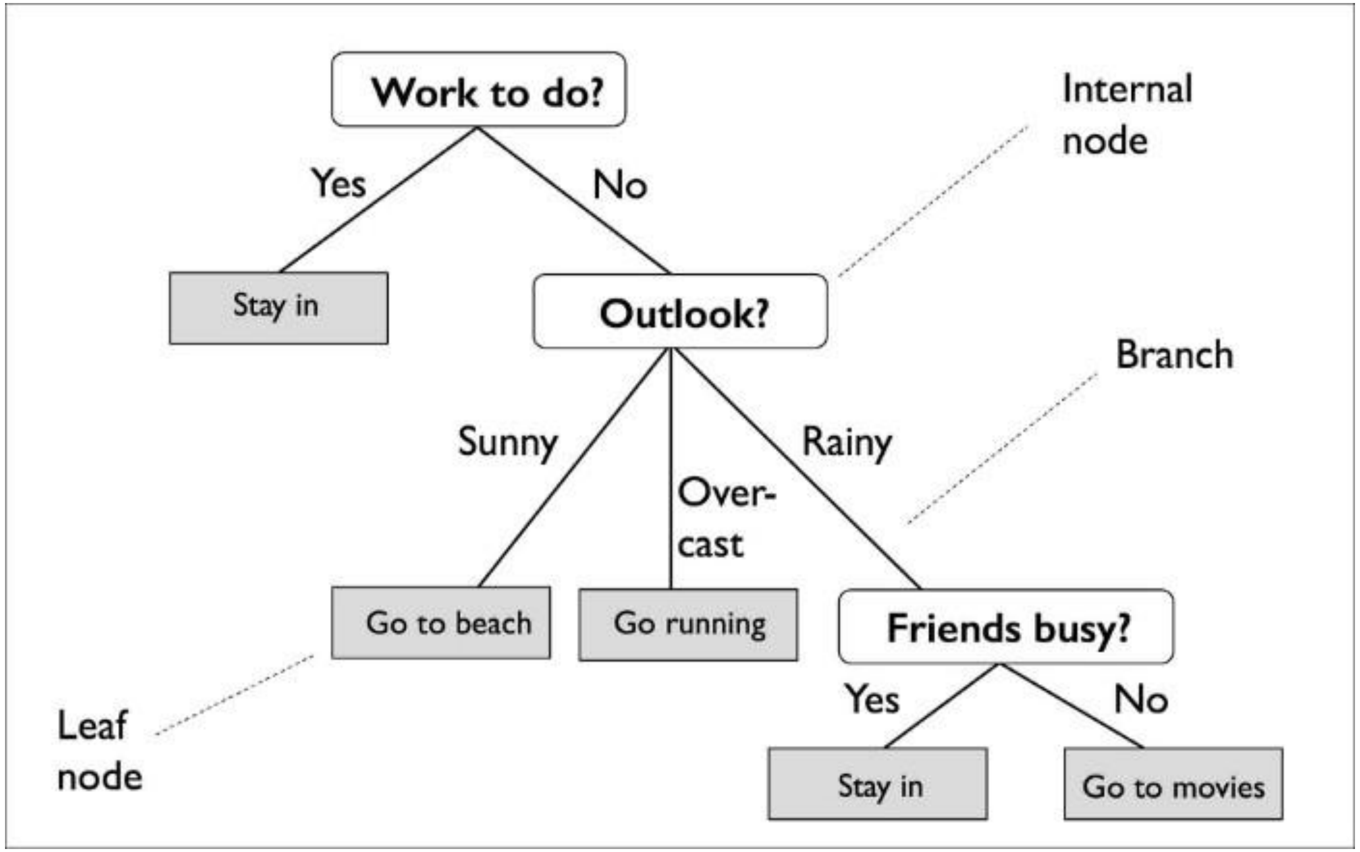
–Predicting whether a patient will develop breast cancer or remain healthy, given genetic information

–Predicting whether or not a user will like a new product, based on user covariates and a history of his/her previous ratings

–Predicting the region of Italy in which a brand of olive oil was made, based on its chemical composition

–Predicting the next elected president, based on various social, political, and historical measurements

DECISION TREE



DECISION TREE

A **decision tree** is a supervised machine learning model used to predict a target by learning decision rules from features. As the name suggests, we can think of this model as breaking down our data by making a decision based on asking a series of questions.

A decision tree is constructed by **recursive partitioning** — starting from the root node (known as the first **parent**), each node can be split into left and right **child** nodes. These nodes can then be further split and they themselves become parent nodes of their resulting children nodes.

DECISION TREE

Maximizing Information Gain

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

Binary case

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

IMPURITY MEASURE

Gini measure

$$I_G(t) = 1 - p_L^2 - p_R^2$$

Entropy

$$I_E(t) = -p_L \times \log(p_L) - p_R \times \log(p_R)\,,$$

where $p_L = p(\text{left}|t)$ and $p_R = p(\text{right}|t)$.

EXAMPLE-GINI IMPURITY

A (40,40) -> (30,10) & (10,30)

B (40,40) -> (20,40) & (20, 0)

IG = (1-0.5^2-0.5^2)=0.5

A: IG_left = (1-(3/4)^2-(1/4)^2)=0.375

A: IG_right = (1-(1/4)^2-(3/4)^2)=0.375

A: IG_G = 0.5 – 4/8*0.375 – 4/8*0.375=0.125

B: IG_left = (1-(2/6)^2-(4/6)^2)=4/9

B: IG_right = (1-(1)^2-(0)^2)=0

B: IG_G = 0.5 –6/8*4/9=0.16

Case B gives higher IG_G, so it should be chosen.

Gini impurity of partition A	
left	right
$\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)=0.375$	$\left(1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2\right)=0.375$
$0.5-\frac{4}{8}*0.375-\frac{4}{8}*0.375=0.125$	

Gini impurity of partition B	
left	right
$\left(1-\left(\frac{2}{6}\right)^2-\left(\frac{4}{6}\right)^2\right)=\frac{4}{9}$	$(1-(1)^2-(0)^2)=0$
$0.5-\frac{6}{8}*\frac{4}{9}-0=0.16$	

GINI OR ENTROPY?

- "Gini" will tend to find the largest class, and "entropy" tends to find groups of classes that make up ~50% of the data.

- Some studies show this doesn't matter – these differ less than 2% of the time.

- Entropy might be a little slower to compute (because it makes use of the logarithm).

https://www.garysieling.com/blog/sklearn-gini-vs-entropy-criteria

https://www.unine.ch/files/live/sites/imi/files/shared/documents/papers/Gini_index_fulltext.pdf

RANDOM FOREST

The random forest algorithm can be summarized in four simple steps:

1. Draw a random bootstrap sample of size n
(randomly choose n samples from the training set with replacement).

2. Grow a decision tree from the bootstrap sample. At each node:

- a. Randomly select d features without replacement.
- b. Split the node using the feature that provides the best split according to the objective function, for instance, maximizing the information gain.

3. Repeat the steps 1-2 k times.

4. Aggregate the prediction by each tree to assign the class label by majority vote.

ADVANTAGES & DISADVANTAGES

Advantages

1. Higher accuracy (remedy overfitting)
2. Handle thousands of input variables without variable selection
3. Indicate the variables that are important in the classification task

Disadvantages

1. Computationally expensive
2. Not easy to interpret (hard to visualize the model or understand why it predicted something)

REFERENCE

Some illustration to the parameters of the code

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://blog.csdn.net/u012102306/article/details/52228516

Decision Tree:

https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054

今天課堂概要

Decision Tree & Random Forest

1. The fundamental concepts of decision trees
2. The mathematics behind the decision tree learning algorithm
3. Information gain and impurity measures
4. Classification trees

下一課...

Dimension reduction method