UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

TFG del Grado en Ingeniería Informática

Simulador árboles de decisión

Presentado por Daniel Drefs Fernandes
en Universidad de Burgos — 15 de abril de 2024
Tutor: Carlos López Nozal, Ismael Ramos Pérez

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

D. nombre tutor, profesor del departamento de nombre departamento, área de nombre área.

Expone:

Que el alumno D. Daniel Drefs Fernandes, con DNI dni, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado título de TFG.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 15 de abril de 2024

Vº. Bº. del Tutor:	Vº. Bº. del co-tutor:
D. nombre tutor	D. nombre co-tutor

Resumen

En este primer apartado se hace una **breve** presentación del tema que se aborda en el proyecto.

Descriptores

Palabras separadas por comas que identifiquen el contenido del proyecto Ej: servidor web, buscador de vuelos, android ...

Abstract

A **brief** presentation of the topic addressed in the project.

Keywords

keywords separated by commas.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

Decision trees are powerful tools in machine learning and data mining, serving as intuitive and interpretable models for decision-making. Two notable algorithms, IDE3 and C4.5, have significantly contributed to their development and widespread use. IDE3, or Iterative Dichotomiser 3, laid the foundation for decision tree learning by recursively partitioning data based on attribute values, aiming to maximize information gain at each step. C4.5 improved upon IDE3 by handling continuous attributes, missing values, and pruning techniques, enhancing the robustness and accuracy of decision trees. These algorithms play a crucial role in various fields, enabling efficient classification and prediction tasks while offering insights into decision-making processes.

That is why this project aims to build a simulator in form of a web application that teaches these algorithms in a way that is easy to understand.

2. Objetivos del proyecto

The primary objective of this project has been to create an interactive and informative web application focused on educating users about decision tree algorithms such as IDE3 and C4.5. In relation to that, it also teaches the concept of entropy and how it is connected to decision trees. The users are provided with explanations, color-coded and dynamic step-by-step visualizations which make the concepts easier to understand. The server uses SVG images to display the algorithm's progress and entropy functions.

3. Conceptos teóricos

In the following, all the theoretical concepts relevant for the understanding of the project will be explained.

Algunos conceptos teóricos de LaTeX [1].

3.1. Decision Trees

A decision tree [4] is a versatile supervised learning algorithm used for classification and regression tasks. Its goal is to predict the value of a variable based on previously processed input. Its hierarchical structure includes a root node and several internal and leaf nodes.

It starts at the root, which represents the feature that best separates the underlying dataset based on a certain criterion. An example would be information gain, which will be explained in a later section. From there branches extend to internal nodes, also called decision nodes. These internal nodes also represent features along with a decision rule that tells us how to further split the data. These features are continuously evaluated until homogenous subsets are created by the leaf nodes. These represent all the possible outcomes of the dataset with each one corresponding to a class label.

Decision tree learning utilizes a divide and conquer approach, iteratively finding the best split points until all or most of the input data is classified.

An example of a decision tree which evaluates whether a person is going to be an astronaut or not:

[1] Créditos a los proyectos de Álvaro López Cantero: Configurador de Presupuestos y Roberto Izquierdo Amo: PLQuiz

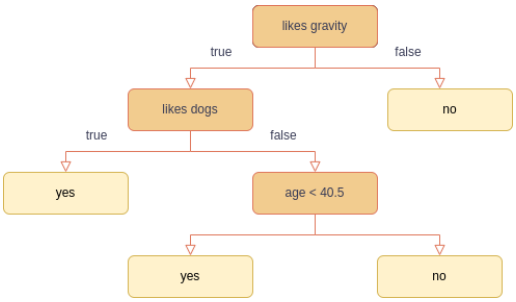

Figura 3.1: Decision tree example

Subsecciones

Además de secciones tenemos subsecciones.

Subsubsecciones

Y subsecciones.

3.2. Entropy

In the context of decision tree algorithms, entropy [3] can be viewed as a measure of impurity in a dataset. It can also be described as a measure of disorder, uncertainty or the expected surprise of a classification, but going forward, the term impurity will be used to avoid confusion. In datasets with binary classes, where variables can only have two possible outcome values, the entropy value lies between 0 and 1, inclusive. The higher the entropy, the more impure the dataset is. In a binary-class dataset, a node that has an equal distribution of, e.g., 5 instances belonging to one class and the other 5 instances belonging to the other class, would have an entropy value of 1. Inversely, a node that has all its instances belong to only one class would have an entropy value of 0, making´it a pure node. The value is calculated using the following formula:

$$E(X) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

$E(X)$ is the entropy of dataset X, n describes the number of classes in the dataset, and p_i the proportion of instances in class i or, in other words, the probability of an instance belonging to class i.

This following graph [6] that describes the entropy function in relation to the composition of a node shows really well how entropy is used to determine how impure a node is:

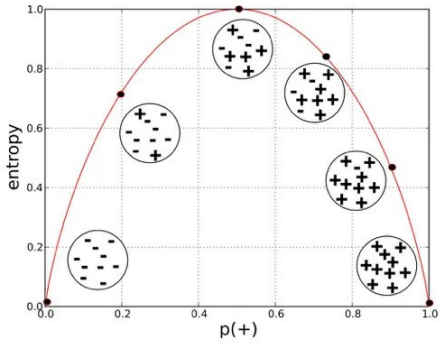

Figura 3.2: Descriptive graph of the entropy function

In a dataset with binary classes, the x-axis describes the amount of instances in a dataset belonging to the positive class while the y-axis measures their entropy values. It records the lowest values at the extremes where there are only or no positive instances recorded in a given dataset while the highest impurity is reached when the numbers of positive and negative instances are equal. However, it must be mentioned that the value of entropy can reach values higher than 1 if the dataset possesses more than 2 class labels.

Conditional Entropy

In the context of decision trees, the conditional entropy [8] $E(Y \mid X)$measures the uncertainty in the target variable Y, which is usually the class labels, given a certain value of the attribute X that is used for splitting a node. It can be described as the weighted sum of $E(Y \mid X = x)$ for every possible value x of X:

$$E(Y \mid X) = \sum_{x \in X} p(x) E(Y \mid X = x)$$

The formula sums up the conditional entropies for all the possible values x of attribute X. As the weight, it uses $p(x)$, which is the proportion between the number of instances that have the attribute value x and the size of the dataset. Calculating the conditional entropies for the individual values x of

attribute X is possible with the following formula:

$$E(Y \mid X = x) = -\sum_{y \in Y} p(y \mid x) log_2(p(y \mid x))$$

$p(y \mid x)$ represents the proportion of instances in the dataset with attribute value x that also have the class value y. It sums over all the possible values of the target variable Y. It can be used synonymous with the formula for $E(X)$ that was previously shown in 3.2

In the end, this conditional entropy calculation evaluates the effectiveness of splitting a node based on the particular attribute X. The lower the conditional entropy, the lower the impurity of the data after splitting the node based on the particular attribute the conditional entropy was calculated for.

3.3. Referencias

Las referencias se incluyen en el texto usando cite [**?**]. Para citar webs, artículos o libros [5], si se desean citar más de uno en el mismo lugar [1, 5].

3.4. Imágenes

Se pueden incluir imágenes con los comandos standard de LaTeX, pero esta plantilla dispone de comandos propios como por ejemplo el siguiente:

Figura 3.3: Autómata para una expresión vacía

3.5. Listas de items

Existen tres posibilidades:

- primer item.
- segundo item.

1. primer item.
2. segundo item.

Primer item más información sobre el primer item.

Segundo item más información sobre el segundo item.

-

3.6. Tablas

Igualmente se pueden usar los comandos específicos de LaTeXo bien usar alguno de los comandos de la plantilla.

Herramientas	App AngularJS	API REST	BD	Memoria
HTML5	X			
CSS3	X			
BOOTSTRAP	X			
JavaScript	X			
AngularJS	X			
Bower	X			
PHP		X		
Karma + Jasmine	X			
Slim framework		X		
Idiorm		X		
Composer		X		
JSON	X	X		
PhpStorm	X	X		
MySQL			X	
PhpMyAdmin			X	
Git + BitBucket	X	X	X	X
MikTEX				X
TEXMaker				X
Astah				X
Balsamiq Mockups	X			
VersionOne	X	X	X	X

Tabla 3.1: Herramientas y tecnologías utilizadas en cada parte del proyecto

4. Técnicas y herramientas

Esta parte de la memoria tiene como objetivo presentar las técnicas metodológicas y las herramientas de desarrollo que se han utilizado para llevar a cabo el proyecto. Si se han estudiado diferentes alternativas de metodologías, herramientas, bibliotecas se puede hacer un resumen de los aspectos más destacados de cada alternativa, incluyendo comparativas entre las distintas opciones y una justificación de las elecciones realizadas. No se pretende que este apartado se convierta en un capítulo de un libro dedicado a cada una de las alternativas, sino comentar los aspectos más destacados de cada opción, con un repaso somero a los fundamentos esenciales y referencias bibliográficas para que el lector pueda ampliar su conocimiento sobre el tema.

4.1. Bootstrap

Bootstrap [7] is a front-end framework which is known for providing useful and easy-to-use HTML and CSS templates like tables, buttons, forms and many others. It also comes with JavaScript components like modal dialogues and dropdown menus.

One of Bootstrap's key features is its grid system which lets users divide their web page's contents into rows and columns:

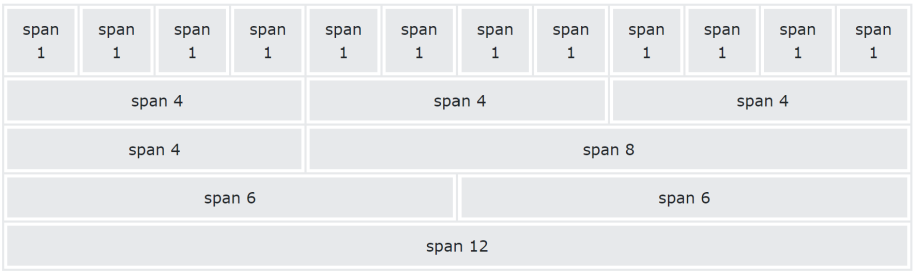

Figura 4.1: The Bootstrap grid system

Each row possesses 12 columns which the user can freely utilize to organize the contents that are to be displayed. Bootstrap also automatically puts the device on which the web page is displayed into one of six different size categories, based on the device's screen width. This, in combination with the grid system, allows the developer to make their website responsive to different screens, ranging from large desktop monitors to smartphones.

In addition to Bootstrap's ability to create responsive websites, the wide variety of CSS classes used for common HTML elements like buttons offer further simplicity to the web design aspect of creating a page. With this, maintaining a visual consistency throughout the project is made easier, too.

In this project, the newest version of Bootstrap at present, Bootstrap 5, is used. The main advantages are its usage of vanilla JavaScript for its components instead of relying on jQuery, new components for better customization and simplified CSS which reduces file size and loading times for the created pages.

4.2. D3.js

"D3-[2] stands for "data-driven documents.ªnd perfectly describes the free, open-source JavaScript library. "Documentsrefers to the Document Object Model (DOM). D3.js lets the user bind data to its elements. The library works like a toolbox that uses a variety of discrete modules which, e.g., allow selection and transition operations. It binds these modules together so all the necessary tools are at hand, ready to be applied.

D3.js does not invent new data presentation formats, instead, it makes use of web standards like SVG to display contents. Incorporating these standards, the library also allows the use of external stylesheets which can be employed to change the graphics' visual representations.

A major feature of D3.js is its ability to dynamically change the displayed contents. Whether that change is triggered by user interactions or a change in underlying data, the library's data join concept allows separate operations for entering, updating and exiting existing DOM elements based on a given set of data. Besides filtering and sorting, it lets you control what happens to your contents in many ways when changes happen and update your website accordingly.

5. Aspectos relevantes del desarrollo del proyecto

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros3, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

6. Trabajos relacionados

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

7. Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliografía

[1] Zachary J Bortolot and Randolph H Wynne. Estimating forest biomass using small footprint lidar data: An individual tree-based approach that incorporates training data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(6):342–360, 2005.

[2] Mike Bostock and Inc. Observable. What is d3? https://d3js.org/what-is-d3, 2024. [Internet; visitado 15-abril-2024].

[3] Shailey Dash. Decision trees explained — entropy, information gain, gini index, ccp pruning. https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c#:~:text=In%20the%20context%20of%20Decision,only%20pass%20or%20only%20fail., 2022. [Internet; visitado 29-marzo-2024].

[4] IBM. What is a decision tree? https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes., -. [Internet; visitado 15-marzo-2024].

[5] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

[6] Sam T. Entropy: How decision trees make decisions. https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8, 2019. [Internet; visitado 29-marzo-2024].

[7] W3Schools. Bootstrap 5 tutorial. https://www.w3schools.com/bootstrap5/index.php, 2024. [Internet; visitado 15-abril-2024].

[8] Wikipedia. Conditional entropy. https://en.wikipedia.org/wiki/Conditional_entropy, 2024. [Internet; visitado 29-marzo-2024].