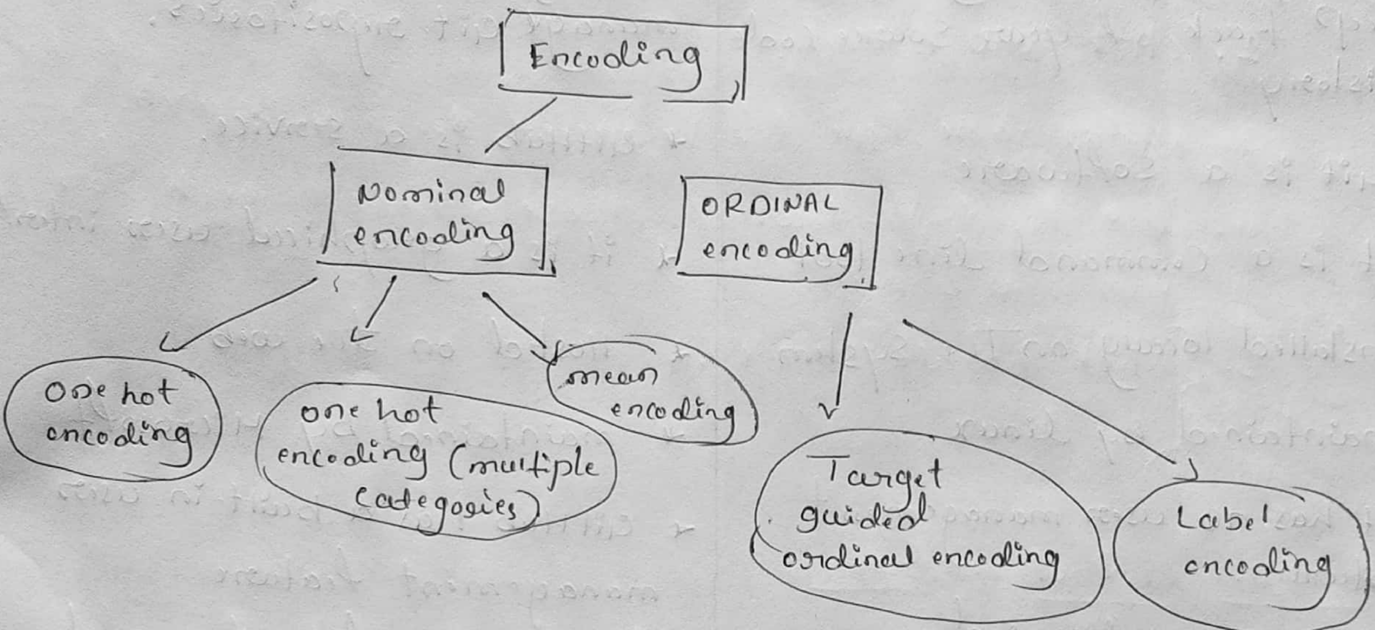
Encoding in ML

Encoding is a technique of converting categorical variables into numerical values so that it could be easily fitted to a machine learning model.



ONE HOT ENCODING

This method is applied to nominal categorical variables

example

Suppose we have a column contaning 3 categorical variables

Then one hot encoding 3 columns will be created each for a categorical variable

Red

blue ⟹

Green

Red	Blue	Green
1	0	0
0	1	0
0	0	1

Difference between Git and GitHub

Git	GitHub
* Git is a version control system that lets you manage and keep track of your source code history	* GitHub is a cloud-based hosting service that lets you manage Git repositories.
* Git is a software	* GitHub is a service.
* it is a command line tool	* it is a graphical user interface.
* installed locally on the system	* Hosted on the web
* maintained by linux	* maintained by Microsoft
* Git has no user management feature	* GitHub has a built-in user management feature.
* open-source licensed	* includes a free-tier and pay-for-use-tier.
* Git provides a Desktop interface named Git Gui	* provides a Desktop interface named GitHub Desktop.

Repository

- Repository in GIT Contain an collection of files of various different versions of a project
- These files are imported from the repository into the local server of the user for further updations.
- users Can also Create a new repository or delete an existing repository.

- we can own repositories individually or we can share ownership of repositories with other people in an organization.
- we can restrict who has access to a repository by choosing the repository's visibility.
- GitHub repository is a repository over the cloud. This means, whatever the data is available on Local repository can be uploaded to Remote Repository on GitHub.

Local repository - is a Git repository that is stored on computer.

Remote repository - is a Git repository that is stored on some remote computer.

- it is usually used by teams as a central repository into which everyone pushes the changes from his local repository and from which everyone pulls changes to his local repository.

Branch in GitHub

Branches allow you to develop features, fix bugs, or safely experiment with new ideas in a contained area of your repository.

- always create a branch from an existing branch.

Creating branch in Github

1. At the top of the app, click current branch and then in the list of branches, click the branch that you want to base your new branch on.

2. click New Branch.

3. under name, type the name of the new branch.

4. use the drop-down to choose a base branch for your new branch.

5. click Create branch.

Version Control System

- A version control system (VCS) refers to the method used to save a file's versions for further reference.
- version control, also known as source control, is the practice of tracking and managing changes to software code.
- version control systems are software tools that helps software teams manage changes to source code over time.
- version control software keeps track of every modification to the code in a special kind of database.

Difference between Classification & Regression in ML

Classification	Regression
The mapping function is used for mapping values to predefined classes.	• mapping function is used for the mapping of values to continuous output.
• Involves prediction of Discrete values	• Continuous values.
• Nature of the predicted data is unordered	• Nature of the predicted data is ordered.
• method of calculation by measuring accuracy	• by measuring root mean square error
example algorithms - Decision tree, logistic regression etc.	• Regression tree (Random forest), Linear regression etc.

Nominal Data

- Nominal data is "labeled" or "named" data which can be divided into various groups that do not overlap.
- Data is not measured or evaluated in this case, it is just assigned to multiple groups.
- In Latin nomenclature "Nomen" means - Name.
- Nominal data does present a similarity between the various

items but details regarding this similarity might 06 not be disclosed.

- In some cases nominal data is also called categorical data.
- if binary data represents "two-valued" data, nominal data represents "multi-valued" data and it can't be quantitative.

Characteristics

- Inconclusive mean value
- qualitative Data property
- Alphabetical
- conclusive mode
- Not quantifiable
- Absence of orders.

Categorical Data

- categorical data is the statistical data comparing categorical variables of data that are converted into categories.
- example = grouped data.
- categorical data could be derived from qualitative data analysis that are countable or from quatitative data analysis grouped within given intervals.

- when we consider data analysis, it is referred to use The term "categorical data", which is applied to data sets.

examples of categorical data

- Birth date
- Travel method to school
- School postcode etc.

- In general, categorical data has values and observations which can be sorted into categories or groups.

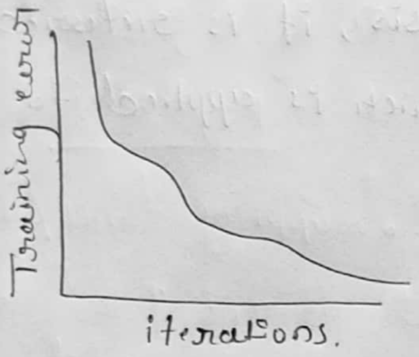
The best way to represent these data is bar graphs and pie charts.

Categorical data are further classified into two types

- Nominal data
- ordinal Data.

Gradient Descent

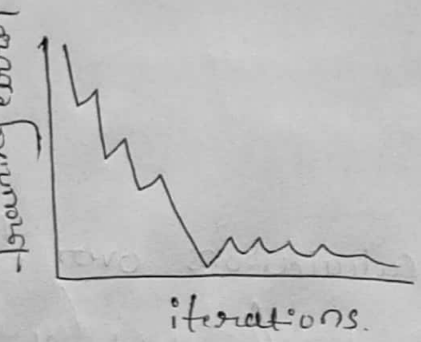
- Gradient Descent involves calculations over the full training set at each step as a result of which it is very slow on very large training data.
- it becomes very computationally expensive to do GD.
- this is great for convex or relatively smooth errors manifolds.



Stochastic Gradient Descent

→ SGD tries to solve the main problem in Batch Gradient descent which is the usage of whole training data to calculate gradients as each step.

- SGD is stochastic in nature.
- it picks up a "random" instance of training data at each step and then computes the gradient making it much fastor as there is much lower data to manipulate at a single time, unlike Batch GD.


training error
iterations.