FLATIRON SCHOOL

Modeling Patterns with Regression Algorithms.

DS

hosted by KASH

CRITICAL QUESTION

How can we **construct and apply algorithms** that leverage **regression fitness** and **continuous distributions** in order to learn patterns in data?

CRITICAL QUESTION

How can we utilize **linear regression** and **logistic regression** algorithms to predict upon new data in unique ways?

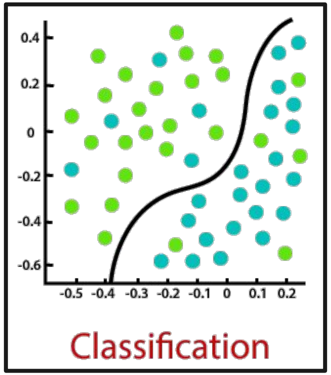
A REFRESHER ON CLASSICAL MACHINE LEARNING

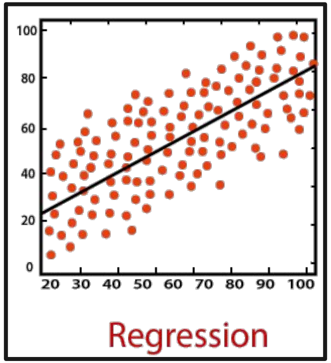
With classification-based prediction, we are allowed to predict labels based on categorical values such as species, survivability, or flower type.

However, there also exists **regression**.

With regression, we predict based on **continuous labels**, such as **age**, **price**, **temperature**, or **square footage**.

CLASSIFIERS AND REGRESSORS: THE BIG DIFFERENCE





As such, many models exist to help fit algorithmically to patterns in data when predicting continuously is mandated.

While classifiers tend to be much more dynamically based on a *mix of programmatic and mathematical construct*, regressors tend to be **far more mathematical** in nature, **based upon the expected distribution of data that the algorithm can fit to**.

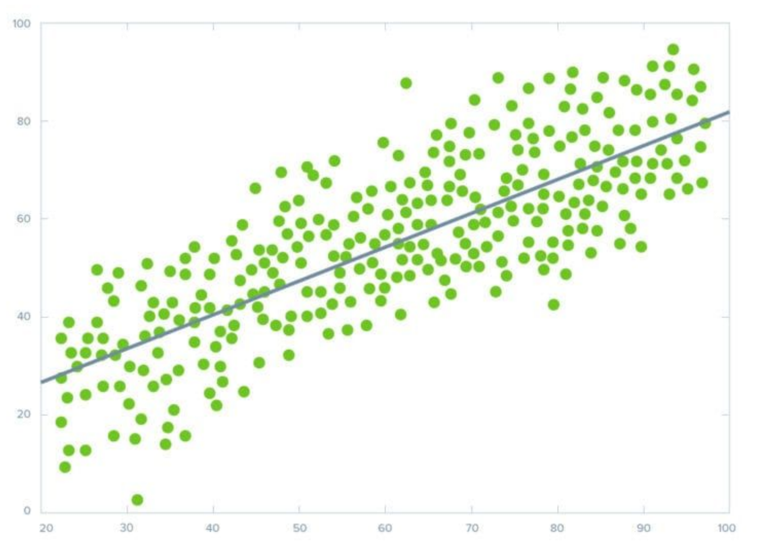
WRITTEN IDEATION

WHAT **TYPES OF DATA** ARE BEST SUITED FOR **REGRESSION ANALYSIS**? CONSIDER DATA FROM A DOMAIN-OF-INTEREST.

AN INTRODUCTION TO REGRESSION-BASED ALGORITHMS

Regression-based algorithms – or **"regressors"**, as they're often called – are machine learning models that **mathematically predict continuous data from input training data**.

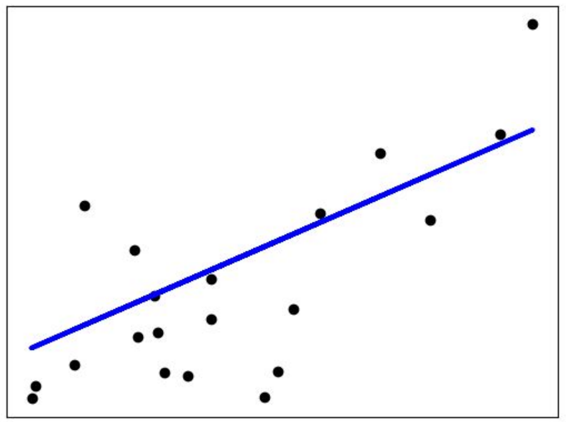
Other than their unique prediction cases, **they often work very similarly to classifiers** in terms of the general structure of how they "fit" to data before "predicting" off of data.



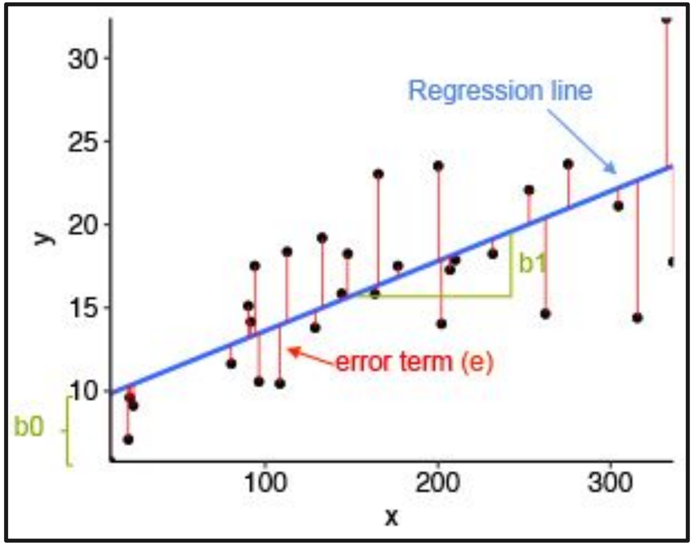
THE LINEAR REGRESSION MODEL

The **linear regression** model is the simplest of all regression algorithms, and in fact is one that you've likely learned more in elementary algebra – it's the **mathematical line-of-best-fit that can be imposed on a linearly-distributed spread of data**.

The strength of a linear regressor is its relative **simplicity** and **reliability** for predicting future values assuming a linear distribution – other types of regression are often more complex and follow more abstract distributions.



PARAMETRIZATION FOR A LINEAR REGRESSOR



The hyperparameters for regression algorithms often have to deal with the **fitness and proximity of the line/curve-of-best-fit to the surrounding data**.

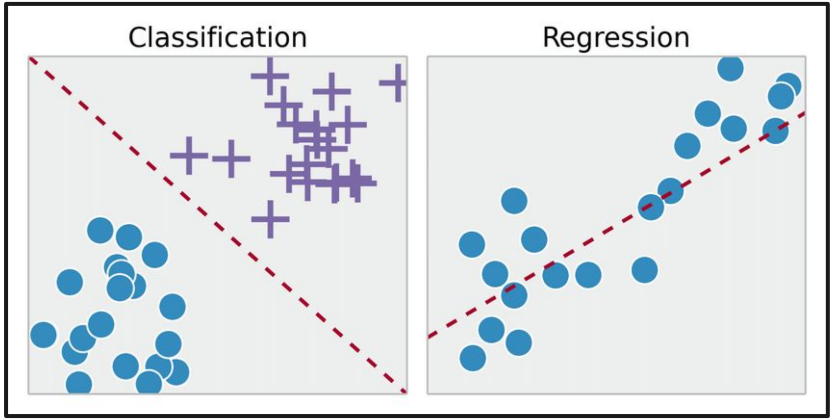
As such, some popular hyperparameters include:

- "***normalize***", a normalization parameter.
- **"*alpha*"**, a penalty term.
- **"*solver*"**, an auto-solver for linear equations.
- **"*fit_intercept*"**, a fitness function.

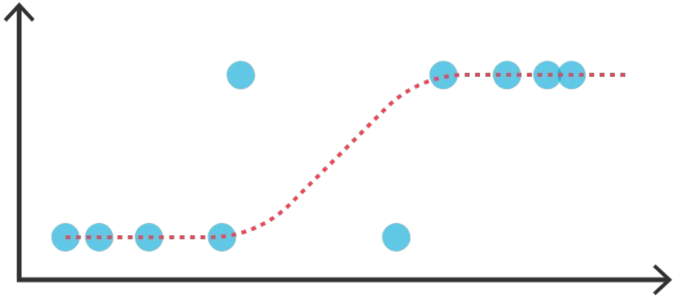
CLASSIFIERS AND REGRESSORS: NOT MUCH OF A DIFFERENCE

By now, you may have suspected that under-the-hood, there's not that much of a difference in **how classifier algorithms and regression algorithms function**.

As a matter-of-fact, **even the predictive functionality of both types of models are highly similar**: most classifiers work by drawing a decision boundary "line" between groups of data, while most regressors work by drawing a best fit "line" across groups of data!



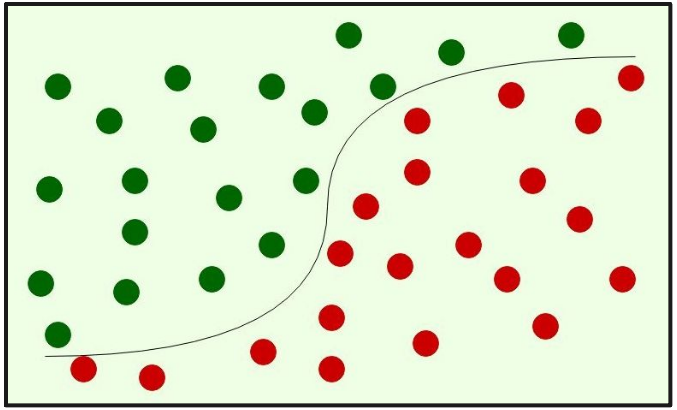
THE LOGISTIC REGRESSION CLASSIFIER



One special type of regression-like algorithm takes advantage of the **symmetry between regressors and classifiers** in order to **exploit continuous regression for discrete classification purposes**... sounds crazy, right?

Logistic regressors depend on the discretized nature of some continuous data in order to create an **S-like minimization-maximization function that essentially converts continuous data into a discrete domain**.

PARAMETRIZATION FOR A LOGISTIC REGRESSION MODEL



Thankfully, despite it being used for classification purposes, **the logistic regressor is still a regression algorithm at heart** and thus is **parametrized by similar regression conditions** as other regression functions.

This means that we can use similar hyperparameters:

- **"*solver*"**, an auto-solver for linear equations.
- **"*penalty*"**, a penalty term.
- **"*C*"**, a distribution flexibility function.

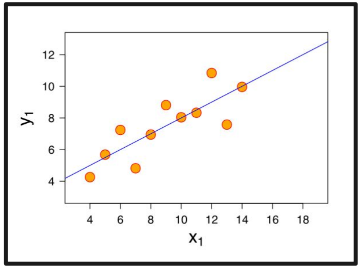
CODING WALKTHROUGH

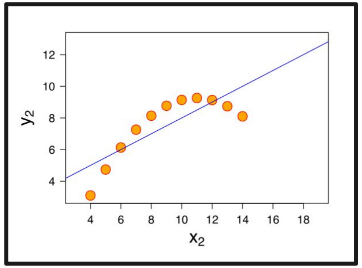
Basic Predictive Analyses and Hyperparameter Tuning with Two Unique Regressors

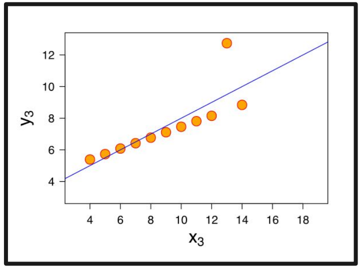
A FOOTNOTE ON EVALUATION AND ERROR FOR REGRESSORS

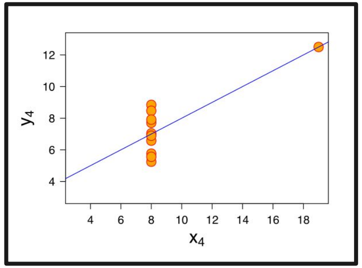
One important comment is that **regression functions do not abide by the same type of errors and accuracy metrics as classifiers**, since now we don't care as much about discrete correctness as much as **continuous proximity and distance to expected labels**.

Error metrics for regressors are generally more mathematical, such as **mean-squared error** or **mean-averaged error** and depend on metrics like the **coefficient-of-determination (R^2)** for overall distributive fitness.









THANK YOU FOR YOUR TIME!