DATA SCIENCE CLUB

Soccer Analysis Workshop

General Schedule

1. Presentation: ~10 min
2. Individual working time: ~40 min
3. Presentation: ~10 min

Presentation Topics

1. Explaining Logistic Regression
2. Explaining Data
3. Explaining any Extra technologies
4. Intro to Notebook

Logistic Regression

Logistic regression is a linear model used to determine probability of an outcome given specifically curated input data points. it essentially creates a distribution over your data points for the probability of that data point belonging to a specific class. This can result in a linear bound on the distribution to perform actual classification(usually is probability of 0.5)

Data: Inputs are continuous vectors of length M. Outputs are discrete.

$$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N} \text{ where } \mathbf{x} \in \mathbb{R}^M \text{ and } y \in \{0, 1\}$$

Model: Logistic function applied to dot product of parameters with input vector.

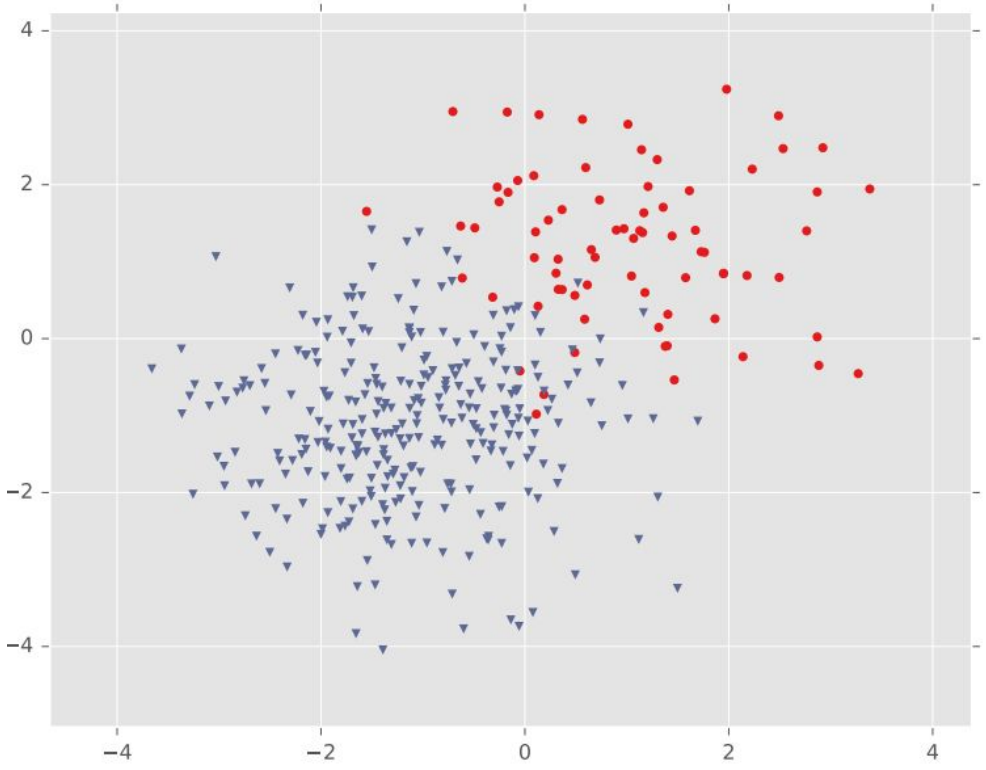
$$p_{\boldsymbol{\theta}}(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x})}$$

Learning: finds the parameters that minimize some objective function. $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} J(\boldsymbol{\theta})$

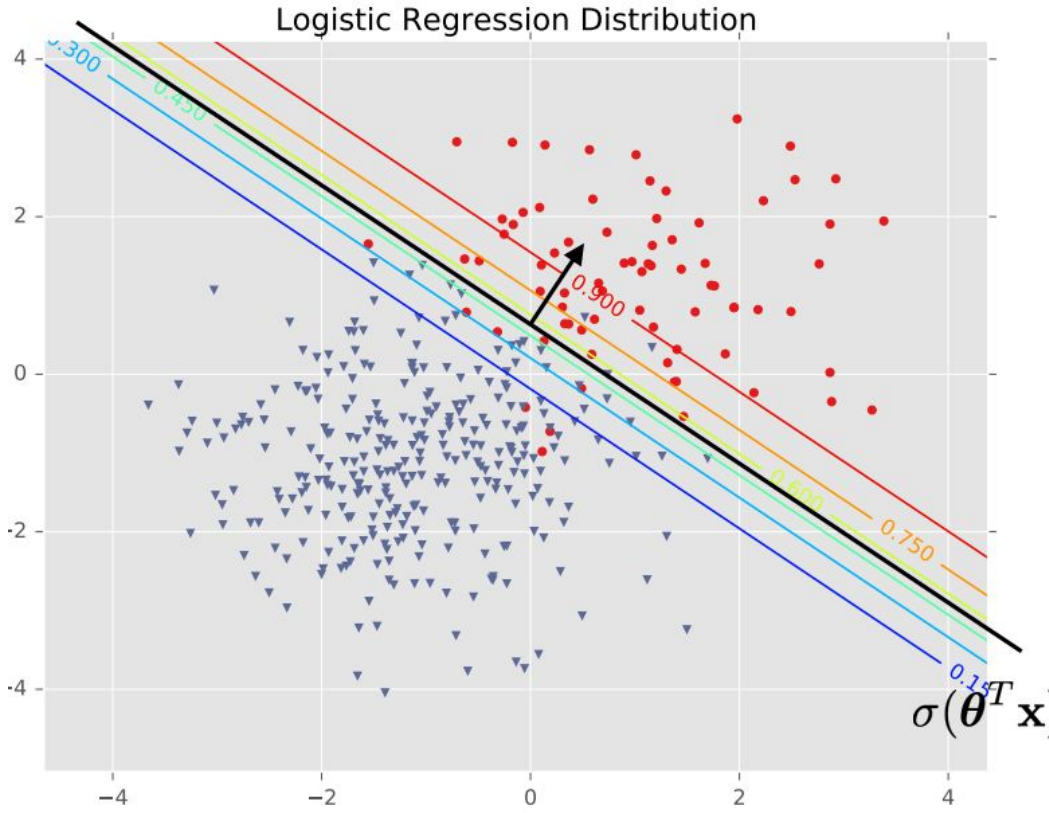
Prediction: Output is the most probable class.

$$\hat{y} = \underset{y \in \{0,1\}}{\operatorname{argmax}} p_{\boldsymbol{\theta}}(y|\mathbf{x})$$

Example Visual

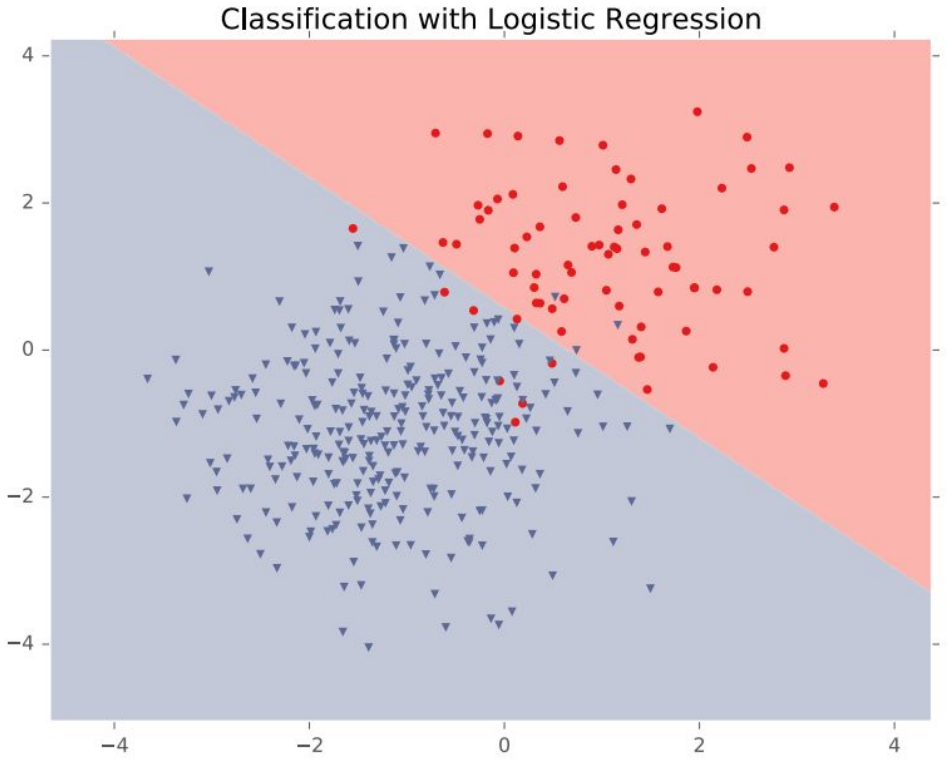


Example Visual



$$\sigma(\boldsymbol{\theta}^T \mathbf{x}) = 0.5$$

Example Visual



Our Dataset

- +25,000 matches
- +10,000 players
- 11 European Countries with their lead championship
- Seasons 2008 to 2016
- Players and Teams' attributes* sourced from EA Sports' FIFA video game series, including the weekly updates
- Team line up with squad formation (X, Y coordinates)
- Betting odds from up to 10 providers
- Link: https://www.kaggle.com/datasets/hugomathien/soccer/data

SQLite

SQLite is a lightweight and self-contained relational database management system. Unlike other client-server databases, SQLite operates as a zero-configuration, and file-based system, making it easy to setup and deploy. The main way they are interacted with is through commands called queries that request specific data from the database

Pandas/Numpy

Pandas and NumPy are both very common tools in data science. Pandas simplifies data manipulation with easy-to-use data structures, like DataFrames, while NumPy provides efficient numerical operations through vectorization. Both tools enable data handling, analysis, and computation in Python with minimal code.

Sklearn

Sklearn is an open-source machine learning library in Python. It provides simple and efficient tools for data analysis and modeling. Sklearn offers a comprehensive set of functionalities for various tasks, including classification, regression, clustering, dimensionality reduction, and model selection.

Notebook/Goal for today

The goal of this is workshop is to help you ease into thinking about how to leverage data to get as much insight and information for prediction, and letting you challenge yourself as much as you'd like. We also try to encourage you to explore many things by looking through the documentation.

This workshop is going to focus on more simplistic Data analysis workflow (ie: working more with the data itself and less complicated models)

Dealing with NaN values

As you'll see all over in this data set, NaN values creep up in annoying ways. Especially since some data points in some tables that are not NaN will reference something in another table that is NaN .There are multiple ways to deal with it, ranging from more involved methods like completely purging the dataset of any datapoint that either is or has any relation to a NaN, to less involved methods such filling in NaN data points with dummy values to avoid those errors.

Choosing features

Much like dealing with NaN values, choosing which features of the data to investigate is a matter of circumstance and choice. Sometimes the insight a feature provides isn’t worth the hassle of wrangling it together from all the data points. It all depends on what you want to get out of the dataset

Link to Notebook

https://drive.google.com/file/d/1T_EQYVQl1cQZzG8pJDlCo2lRrdDnTS9R/view?usp=sharing