Classification and regression problems

An introduction

Paulo S. A. Sousa

Prediction problems

In general, in a predicton problem, we have:

- The predictors: $X_1, X_2, \cdots, X_n$
- The outcome variable, Y.
- And we want to predict Y with the predictors.

Prediction problems

We have dataset:

X_1	X_2	$\cdots$	X_n	Y
value_{11}	value_{12}	$\cdots$	value_{1n}	y_1
value_{21}	value_{22}	$\cdots$	value_{2n}	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
value_{k1}	value_{k2}	$\cdots$	value_{kn}	y_k

Prediction problems

We can imagine that a function f can be found such that:
$Y = f\left(X_1, X_2, \cdots, X_n\right).$

- Unfortunately, we *only* have the data!
- Our goal is to find a function that fits the data, and
  - More importantly, a function that can be used to predict new cases.

Prediction problems

- Since out goal is predicting:
  - We do not need to find out the “perfect” function.
  - That is enough to have a function that produces *good* predictions.
  - Is such an undertaking possible?
    - Yes, in very many practical applications, these methods have proven effective.

Prediction problems

- If the underlying mathematical process behind the data is constantly changing over time:
  - It will be hard and sometimes impossible to produce good predictions.
- We will focus on situations that do not change over time or change very little.

Classification problems

- In many important problems, the outcome variable, Y, is categorical.
- By categorical, we mean that the variable assume values *only* from a finite set of discrete values.

Classification problems

- These values that the outcome variable can assume are called **classes**.
- If the outcome variable has only *two* values, then the problem is called a **binary classification problem**.

Classification problems

Some examples of binary classification problems:

- Predicting whether a credit card transaction is fradulent or not.
- Predicting whether a student will pass an exam or not.
- Predicting whether the sales of next year will be larger than a certain threshold or not.
- Predicting whether a marriage will be a happy one or not.

Classification problems

More examples of classification problems:

- **Fraud detection:** Classification models can be used to detect fraudulent transactions, identify potentially fraudulent customers, and prevent financial losses.
- **Customer segmentation:** Classification models can be used to segment customers into different groups based on their demographics, behavior patterns, and purchasing history.

Classification problems

More examples of classification problems:

- **Credit scoring:** Classification models can be used to predict the likelihood of loan default and assess credit risk based on factors such as credit history, income, and employment status.
- **Churn prediction:** Classification models can be used to predict which customers are likely to stop using a product or service, and to develop strategies for retaining those customers.

Classification problems

More examples of classification problems:

- **Spam filtering:** Classification models can be used to classify email messages as spam or legitimate, and to prevent unwanted messages from reaching users' inboxes.
- **Sentiment analysis:** Classification models can be used to classify customer feedback and social media posts as positive, negative, or neutral, and to monitor brand reputation.

Classification problems

More examples of classification problems:

- **Product categorization:** Classification models can be used to automatically categorize products based on their attributes and characteristics, and to improve search and navigation on e-commerce websites.
- **Customer retention:** Classification models can be used to identify customers who are at risk of leaving a company, and to develop targeted retention strategies.

Classification problems

More examples of classification problems:

- **Image classification:** Classification models can be used to classify images based on their content, and to automate tasks such as object recognition and facial recognition.
- **Anomaly detection:** Classification models can be used to identify unusual patterns or outliers in data, and to detect anomalies such as network intrusions or equipment failures.

Classification problems

Let us go to this online playground, to get some intuition about how classification prediction models work:
Classifier Playground

Classification problems

From the examples we have tried in the classifier playground, we can draw the following conclusions:

- The classifier tries to separate the classes via a mathematical formula.
- The more flexible is the mathematical separator, the better potentially the classifier will be.

Classification problems

- However, simple linear separators may be just perfect for some problems.

Regression problems

- While in classification problems, the outcome variable is *categorical*, in regression problems, the outcome variable is *numerical*.
- Consequently, in classification problems, we try to predict a *class*, whereas in regression problems, we try to predict a *number*.

Regression problems

Some examples of regression problems:

- **Sales forecasting:** One of the most common regression problems in business management is predicting future sales based on historical sales data, market trends, and other factors such as seasonality and promotions.

  Price optimization: Regression can be used to determine the optimal price for a product or service by analyzing the relationship between price and demand.

Regression problems

More examples of regression problems:

- **Customer lifetime value:** Regression models can be used to predict the future value of a customer based on their past purchase history, demographic data, and other factors.
- **Customer satisfaction:** Regression can be used to identify the key drivers of customer satisfaction and how they impact overall customer experience.

Regression problems

More examples of regression problems:

- **Employee turnover:** Regression models can be used to predict which employees are most likely to leave a company based on factors such as job satisfaction, compensation, and tenure.
- **Inventory management:** Regression can be used to forecast demand for products and optimize inventory levels to minimize stockouts and overstocking.

Regression problems

More examples of regression problems:

- **Market share analysis:** Regression models can be used to analyze the relationship between market share and factors such as price, advertising, and product quality.
- **Credit risk assessment:** Regression can be used to predict the likelihood of default on loans and other forms of credit based on factors such as credit history, income, and employment status.

Regression problems

More examples of regression problems:

- **Market demand analysis:** Regression can be used to analyze the relationship between market demand and factors such as price, income, and consumer preferences.
- **Supply chain optimization:** Regression can be used to optimize supply chain operations by forecasting demand, identifying bottlenecks, and optimizing production and distribution schedules.

Regression problems

More examples of regression problems:

- **Return on investment (ROI) analysis:** Regression can be used to analyze the relationship between marketing or advertising expenditures and sales revenue, and to calculate the ROI of various marketing campaigns.
- **Resource allocation:** Regression can be used to determine the optimal allocation of resources such as marketing budgets or production capacity based on factors such as market demand and resource availability.

Regression problems

More examples of regression problems:

- **Risk analysis:** Regression can be used to identify and quantify risks in business operations and to develop strategies for mitigating those risks.
- **Employee performance analysis:** Regression can be used to analyze the relationship between employee performance and factors such as training, job satisfaction, and compensation.

Regression problems

More examples of regression problems:

- **Energy consumption analysis:** Regression can be used to analyze the relationship between energy consumption and factors such as temperature, time of day, and occupancy, and to optimize energy usage in buildings and other facilities.
- **Quality control:** Regression can be used to identify factors that affect product quality and to develop strategies for improving quality control processes.