6COM2000 Advanced Artificial Intelligence

Regression vs Classification. Introduction to Naive Bayes for NLP

Olga Tveretina

Overview

Regression vs Classification

What is Regression? What is Classification?

Understanding the difference between regression and classification is useful for developing models and solving problems.

- Regression algorithms predict continuous value from the provided input.
  - The method of least squares
  - Linear regression
  - Logistic regression
- A procedure in which a function separates the data into discrete values, i.e., multiple classes of datasets using independent features, is called classification.
  - Naive Bayes
  - Binary classification

Applications

Examples of applications of regression and classification:

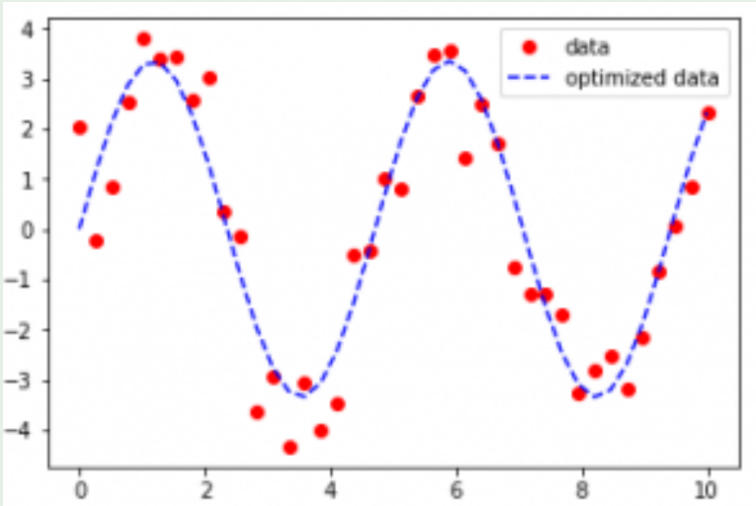
- Regression:
  - Predicting stock prices
  - Sales forecasting
- Classification
  - Email spam filtering
  - Image recognition

Curve Fitting and Regression

Curve Fitting

1. Curve fitting examines the relationship between one or more predictors (independent variables) and a response variable (dependent variable), with the goal of defining a "best fit" model of the relationship.

Example (Curve Fitting Example)



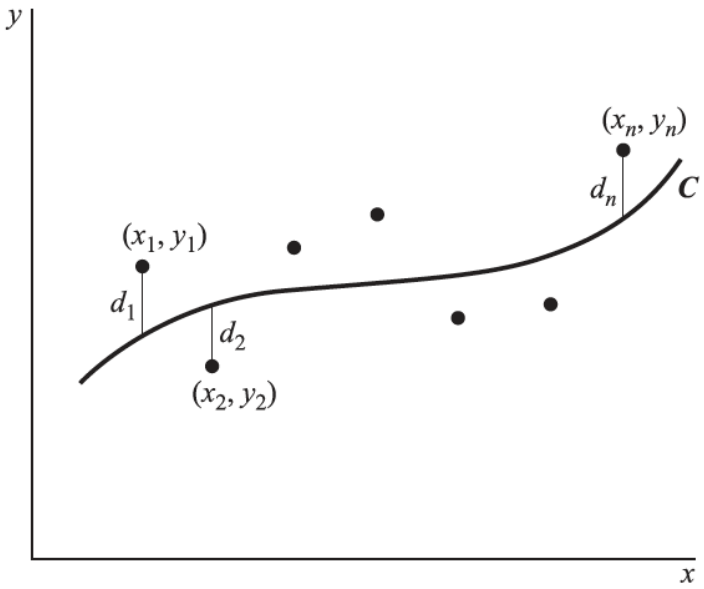
Source: https://www.geeksforgeeks.org

What Is a Regression?

1. A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables.
2. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.
3. It does this by essentially fitting a best-fit line and seeing how the data is dispersed around this line.
4. Regression captures the correlation between variables observed in a data set and quantifies whether those correlations are statistically significant or not.
5. The two basic types of regression are simple linear regression and multiple linear regression.

The Method of Least Squares

1. **Simple regression or ordinary least squares (OLS):** linear regression is the most common form of this technique
2. Linear regression establishes the linear relationship between two variables based on a line of best fit



Source: [3]

- A measure of the goodness of fit of the curve C to the set of data is provided by the quantity $d_1^2 + d_2^2 + \cdots + d_n^2$.
- If this is small, the fit is good, if it is large, the fit is bad.

The Method of Least Squares (cont.)

Definition

Of all curves in a given family of curves approximating a set of n data points, a curve having the property that $d_1^2 + d_2^2 + \cdots + d_n^2$ a minimum is called a best-fitting curve in the family.

The least squares line approximating the set of points $(x_1, y_1), \ldots, (x_n, y_n)$ has the equation (This equation is derived by minimizing the sum of the squares of the vertical deviations from each data point to the line (hence, "least squares"))

$$y = a + bx$$

where the constants a and b are determined by solving simultaneously the equations

$$b = \frac{\sum[(x - \bar{x}) \times (y - \bar{y})]}{\sum(x - \bar{x})^2}$$

The Method of Least Squares: Steps

1. Find the means of the dependent and independent variables.
2. Calculate the difference between each value and the mean value for both the dependent and the independent variable.
3. Find the product of multiplying these two differences together.
4. The final step is to calculate the intercept, which we can do using the initial regression equation with the values of test score and time spent set as their respective means, along with our newly calculated coefficient.

The Method of Least Squares: Example

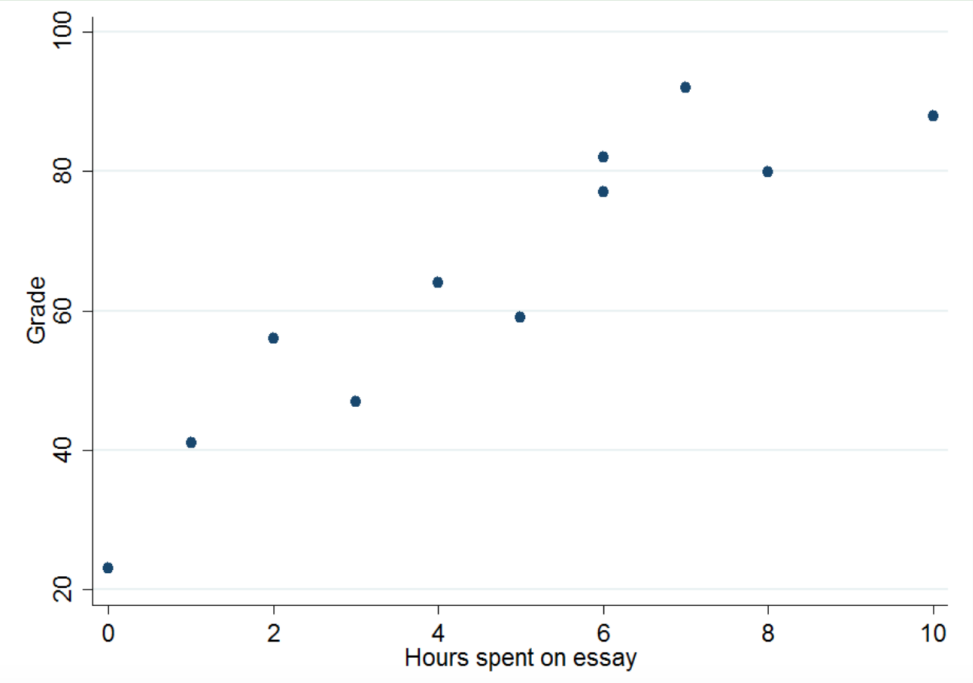
Example

1. A teacher is asked to work out how time spent writing an essay affects essay grades
2. It is easy to look at a graph of time spent writing essays and essay grades say "Hey, people who spend more time on their essays are getting better grades."
3. What is much harder (and realistically, pretty impossible) to do by eye is to try and predict what score someone will get in an essay based on how long they spent on it

https://www.technologynetworks.com/

The Method of Least Squares: Example (cont.)

Example



	Hours spent on essay	Grade
	6	82
	10	88
	2	56
	4	64
	6	77
	7	92
	0	23
	1	41
	8	80
	5	59
	3	47
Mean	4.72	64.45

(Example by Andrew Lee)

The Method of Least Squares: Example (cont.)

Example

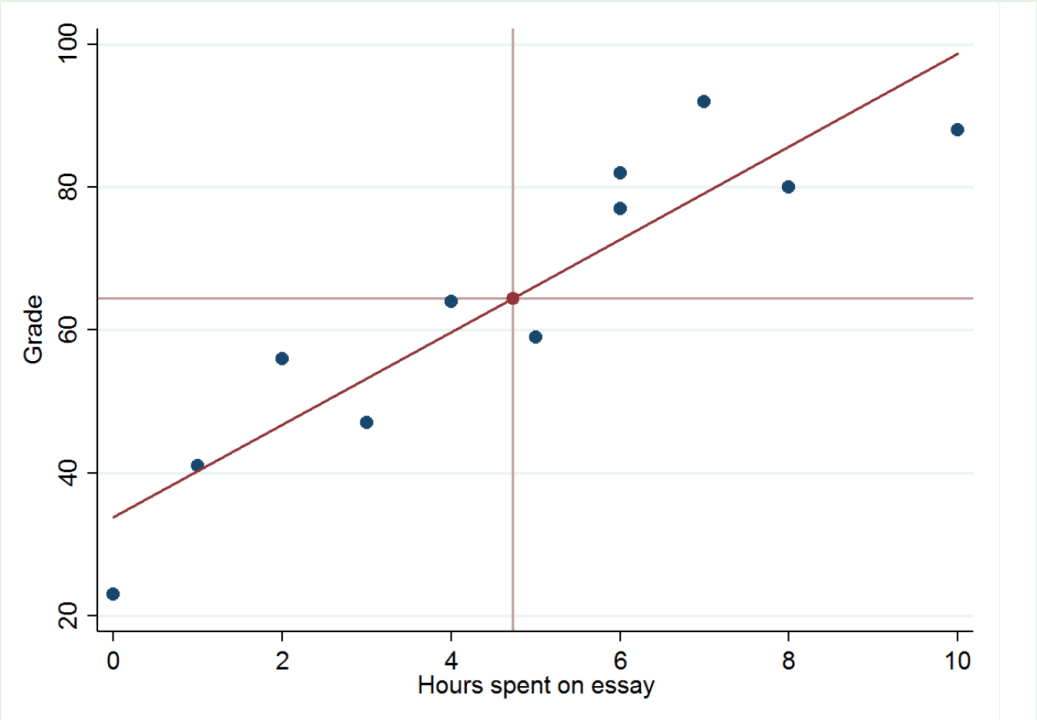
Hours spent on essay	Grade	Hours spent – Average Hours Spent $(x - \bar{x})$	Grade – Average Grade $(y - \bar{y})$	$(x - \bar{x}) \times (y - \bar{y})$
6	82	1.27	17.55	22.33
10	88	5.27	23.55	124.15
2	56	-2.73	-8.45	23.06
4	64	-0.73	-0.45	0.33
6	77	1.27	12.55	15.97
7	92	2.27	27.55	62.60
0	23	-4.73	-41.45	195.97
1	41	-3.73	-23.45	87.42
8	80	3.27	15.55	50.88
5	59	0.27	-5.45	-1.49
3	47	-1.73	-17.45	30.15

(Example by Andrew Lee, https://www.technologynetworks.com/)

The Method of Least Squares: Example (cont.)

Example

This line should cross the means of both the time spent on the essay and the mean grade received:



(Example by Andrew Lee, https://www.technologynetworks.com/)

The Method of Least Squares: Example (cont.)

Example

To calculate a least squares regression line (the equation $y = a + bx$), follow these steps:

1. Calculate the means of x and y.
2. Calculate the deviations of each x and y from their means.
3. Multiply each x deviation by the corresponding y deviation and sum them all up to get the covariance.
4. Square each x deviation, then sum them all to get the variance of x.
5. Calculate the slope as the covariance divided by the variance.
6. Calculate the y-intercept (a) as the mean of y minus 'b' times the mean of x:

$$a = \overline{y} - b\overline{x}$$

Limitations of the Least Squares Method

Even though the least-squares method is considered the best method to find the line of best fit, it has a few limitations. They are:

1. This method exhibits only the relationship between the two variables. All other causes and effects are not taken into consideration.
2. This method is unreliable when data is not evenly distributed.
3. This method is very sensitive to outliers. In fact, this can skew the results of the least-squares analysis.

Bayes' Theorem

Bayes' Theorem

1. Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes, is a formula for determining conditional probability.
2. The theorem has become a useful element in the implementation of machine learning.

Theorem

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

$P(A)$ = the probability of A occurring; $P(B)$ = the probability of B occurring; $P(A|B)$ = the probability of A given B; $P(B|A)$ = the probability of B given A; $P(AB)$ = the probability of both A and B occurring.

Deriving the Bayes' Theorem Formula

Bayes' Theorem follows simply from the axioms of conditional probability, which is the probability of an event given that another event occurred.

$$P(AB) = P(A)P(B|A)$$

$$P(AB) = P(B)P(A|B)$$

That is,

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

Example of Bayes' Theorem I
https://www.mathsisfun.com/data/bayes-theorem.html

Example (Fire)

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then find the probability of dangerous Fire when there is Smoke:

$$P(Fire|Smoke) = \frac{P(Fire) \cdot P(Smoke|Fire)}{P(Smoke)} = \frac{0.01 \cdot 0.9}{0.1} = 0.09$$

Example of Bayes' Theorem II
https://www.mathsisfun.com/data/bayes-theorem.html

Example (Picnic Day)

You are planning a picnic today, but the morning is cloudy

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

$$P(Rain|Cloud) = \frac{P(Rain) \cdot P(Cloud|Rain)}{P(Cloud)} = \frac{0.1 \cdot 0.5}{0.4} = 0.125$$

Naive Bayes

Naive Bayes Introduction

To build a classifier, we will first start by creating conditional probabilities given the following table:

Positive tweets:

I am happy because I am learning AI
I am happy, not sad

Negative tweets:

I am sad, I am not learning AI
I am sad, not happy

Naive Bayes Introduction (cont.)

To build a classifier, we will first start by creating the following table of probabilities (approx.):

Words	**Positive**	**Negative**
I	3	3
am	3	3
happy	2	1
because	1	0
learning	1	1
AI	1	1
sad	1	2
not	1	2
	13	**12**

Words	**Positive**	**Negative**
I	0.24	0.25
am	0.24	0.25
happy	0.15	0.08
because	0.08	0
learning	0.08	0.08
AI	0.08	0.08
sad	0.08	0.17
not	0.08	0.17
sum $\approx$	1	1

Naive Bayes Introduction (cont.)

I am happy today, I am learning

Words	Pos	Neg
I	0.24	0.25
am	0.24	0.25
happy	0.15	0.08
because	0.08	0
learning	0.08	0.08
AI	0.08	0.08
sad	0.08	0.17
not	0.08	0.17
sum $\approx$	1	1

$$\prod_{i=1}^{n} \frac{P(\omega_i|pos)}{P(\omega_i|neg} = \frac{0.24}{0.25} \times \cdots \times \frac{0.08}{0.17} \approx 1.6$$