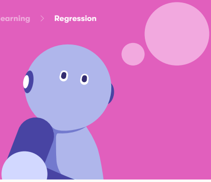


III. Regression

Our main learning objective in this section is another nice example of supervised learning methods, and almost as simple as the nearest neighbor classifier too: linear regression. We'll introduce its close cousin, logistic regression as well.

Note

The difference between classification and regression

There is a small but important difference in the kind of predictions that we should produce in different scenarios. While for example the nearest neighbor classifier chooses a class label for any item out of a given set of alternatives (like spam/ham, or 0,1,2,...,9), linear regression produces a numerical prediction that is not constrained to be an integer (a whole number as opposed to something like 3.14). So linear regression is better suited in situations where the output variable can be any number like the price of a product, the distance to an obstacle, the box-office revenue of the next Star Wars movie, and so on.

The basic idea in linear regression is to add up the effects of each of the feature variables to produce the predicted value. The technical term for the adding up process is *linear combination*. The idea is very straightforward, and it can be illustrated by your shopping bill.

Note

Thinking of linear regression as a shopping bill

Suppose you go to the grocery store and buy 2.5kg potatoes, 1.0kg carrots, and two bottles of milk. If the price of potatoes is 2€ per kg, the price of carrots is 4€ per kg, and a bottle of milk costs 3€, then the bill, calculated by the cashier, totals 2.5 × 2€ + 1.0 × 4€ + 2 × 3€ = 15€. In linear regression, the amount of potatoes, carrots, and milk are the inputs in the data. The output is the cost of your shopping, which clearly depends on both the price and how much of each product you buy.

The word linear means that the increase in the output when one input feature is increased by some fixed amount is always the same. In other words, whenever you add, say, two kilos of carrots into your shopping basket, the bill goes up 8€. When you add another two kilos, the bill goes up another 8€, and if you add half as much, 1kg, the bill goes up exactly half as much, 4€.

Key terminology

Coefficients or weights

In linear regression terminology, the prices of the different products would be called coefficients or weights (this may appear confusing since we measured the amount of potatoes and carrots by weight, but not let yourself be tricked by this). One of the main advantages of linear regression is its easy interpretability: the learned weights may in fact be more interesting than the predictions of the outputs.

For example, when we use linear regression to predict the life expectancy, the weight of smoking (cigarettes per day) is about minus half a year, meaning that smoking one cigarette more per day takes you on the average half a year closer to termination. Likewise, the weight of vegetable consumption (handful of vegetables per day) has weight plus one year, so eating a handful of greens gives every day gives you on the average one more year.

Unanswered

Exercise 16: Linear regression

Suppose that an extensive study is carried out, and it is found that in a particular country, the life expectancy (the average number of years that people live) among non-smoking women who don't eat any vegetables is 80 years. Suppose further that on the average, men live 5 years less. Also take the numbers mentioned above: every cigarette per day reduces the life expectancy by half a year, and a handful of veggies per day increases it by one year.

Calculate the life expectancies for the following example cases:

For example, the first case is a male (subtract 5 years), smokes 8 cigarettes per day (subtract 8 × 0.5 = 4 years), and eats two handfuls of veggies per day (add 2 × 1 = 2 years), so the predicted life expectancy is 80 - 5 - 4 + 2 = 73 years.

Gender	**Smoking (cigarettes per day)**	**Vegetables (handfuls per day)**	**Life expectancy (years)**
male	8	2	73
male	0	6	**A**
female	16	1	**B**
female	0	4	**C**

Your task: Enter the correct value as an integer (whole number) for the missing sections A, B, and C above.

A	Answer
B	Answer
C	Answer

Submit

In the above exercise, the life expectancy of non-smoking, veggie-hating women, 80 years, was the starting point for the calculation. The technical term for the starting point is the **intercept**. We will return to this below when we discuss how to learn linear regression models from data.

Learning linear regression

Above, we discussed how predictions are obtained from linear regression when both the weights and the input features are known. So we are given the inputs and the weight, and we can produce the predicted output.

When we are given the inputs and the outputs for a number of items, we can find the weights such that the predicted output matches the actual output as well as possible. This is the task solved by machine learning.

Note

Example

Continuing the shopping analogy, suppose we were given the contents of a number of shopping baskets and the total bill for each of them, and we were asked to figure out the price of each of the products (potatoes, carrots, and so on). From one basket, say 1kg of sirloin steak, 2kg of carrots, and a bottle of Chianti, even if we knew that the total bill is 35€, we couldn't determine the prices because there are many sets of prices that will yield the same total bill. With many baskets, however, we will usually be able to solve the problem.

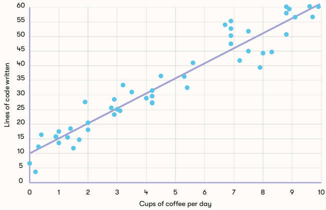
But the problem is made harder by the fact that in the real world, the actual output isn't always fully determined by the input, because of various factors that introduce uncertainty or "noise" into the process. You can think of shopping at a bazaar where the prices for any given product may vary from time to time, or a restaurant where the final damage includes a variable amount of tip. In such situations, we can estimate the prices but only with some limited accuracy.

Finding the weights that optimize the match between the predicted and the actual outputs in the training data is a classical statistical problem dating back to the 1800s, and it can be easily solved even for massive data sets.

We will not go into the details of the actual weight-finding algorithms, such as the classical least squares technique, simple as they are. However, you can get a feel of finding trends in data in the following exercises.

Visualizing linear regression

A good way to get a feel for what linear regression can tell us is to draw a chart containing our data and our regression results. As a simple toy example our data set has one variable, the number of cups of coffee an employee drinks per day, and the number of lines of code written per day by that employee as the output. This is not a real data set as obviously there are other factors having an effect on the productivity of an employee other than coffee that interact in complex ways. The increase in productivity by increasing the amount of coffee will also hold only to a certain point after which the jitters distract too much.



When we present our data in the chart above as points where one point represents one employee, we can see that there is obviously a trend that drinking more coffee results in more lines of code being written (recall that this is completely made-up data). From this data set we can learn the coefficient, or the weight, related to coffee consumption, and by eye we can already say that it seems to be somewhere close to five, since for each cup of coffee consumed the number of lines programmed seems to go up roughly by five. For example, employees who drink around two cups of coffee per day seem to produce around 20 lines of code per day, and similarly at four cups of coffee, the amount of lines produced is around 30.

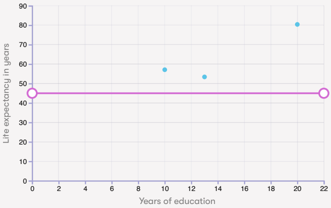
It can also be noted that employees who do not drink coffee at all also produce code, and is shown by the graph to be about ten lines. This number is the intercept term that we mentioned earlier. The intercept is another parameter in the model just like the weights are, that can be learned from the data. Just as in the life expectancy example it can be thought of as the starting point of our calculations before we have added in the effects of the input variable, or variables if we have more than one, be it coffee cups in this example, or cigarettes and vegetables in the previous one.

The line in the chart represents our predicted outcome, where we have estimated the intercept and the coefficient by using an actual linear regression technique called least squares. This line can be used to predict the number of lines produced when the input is the number of cups of coffee. Note that we can obtain a prediction even if we allow only partial cups (like half, 1/4 cups, and so on).

Unanswered

Exercise 17: Life expectancy and education (part 1 of 2)

Let's study the link between the total number of years spent in school (including everything between preschool and university) and life expectancy. Here is data from three different countries displayed in a figure represented by dots:



We have one country where the average number of years in school is 10 and life expectancy is 57 years, another country where the average number of years in school is 13 and life expectancy is 53 years, and a third country where the average number of years in school is 20 and life expectancy is 80 years.

You can drag the end points of the solid line to position the line in such a way that it follows the trend of the data points. Note that you will not be able to get the line fit perfectly with the data points, and this is fine: some of the data points will lie above the line, and some below it. The most important part is that the line describes the overall trend.

After you have positioned the line you can use it to predict the life expectancy.

Given the data, what can you tell about the life expectancy of people who have 15 years of education? Important: Notice that even if you can obtain a specific prediction, down to a fraction of a year, by adjusting the line, you may not necessarily be able to give a confident prediction. Take the limited amount of data into account when giving your answer.

Select the correct answer

It is exactly 64 years

It is certainly between 60 and 70 years

It is certainly 70 years or less

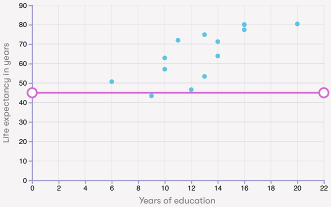
It is probably less than 90

Submit

Unanswered

Exercise 18: Life expectancy and education (part 2 of 2)

In the previous exercise, we only had data from three countries. The full data set consists of data from 14 different countries, presented here in a graph:



Based on this data, would your prediction about the life expectancy of people with 15 years of education change? If so, why?

Which of the following options would best match your estimate of the life expectancy for people with 15 years of education? Choose the most specific option that you think is justified by fitting the straight line model to the above data.

Select the correct answer

Probably between 45 and 50 years

Probably between 50 and 90 years

Probably between 69 and 71 years

Probably between 15 and 150 years

Submit

It should be pointed out that studies like those used in the above exercises cannot identify causal relationships. In other words, from this data alone, it is impossible to say whether studying actually increases life expectancy through a betterinformed and healthier life-style or other mechanisms, or whether the apparent association between life expectancy and education is due to underlying factors that affects both. It is likely that, for example, in countries where people tend to be highly educated, nutrition, healthcare, and safety are also better, which increases life expectancy. With this kind of simple analysis, we can only identify associations, which can nevertheless be useful for prediction.

Machine learning applications of linear regression

Linear regression is truly the workhorse of many AI and data science applications. It has its limits but they are often compensated by its simplicity, interpretability and efficiency. Linear regression has been successfully used in the following problems to give a few examples:

- prediction of click rates in online advertising
- prediction of retail demand for products
- prediction of box-office revenue of Hollywood movies
- prediction of software cost
- prediction of insurance cost
- prediction of crime rates
- prediction of real estate prices

Could we use regression to predict labels?

As we discussed above, linear regression and the nearest neighbor method produce different kinds of predictions. Linear regression outputs numerical outputs while the nearest neighbor method produces labels from a fixed set of alternatives (“classes”).

Where linear regression excels compared to nearest neighbors is interpretability. What do we mean by this? You could say that in a way, the nearest neighbor method and any single prediction that it produces are easy to interpret: it’s just the nearest training data element! This is true, but when it comes to the interpretability of the learned model, there is a clear difference. Interpreting the trained model in nearest neighbors in a similar fashion as the weights in linear regression is impossible: the learned model is basically the whole data, and it is usually way too big and complex to provide us with much insight. So what if we’d like to have a method that produces the same kind of outputs as the nearest neighbor, labels, but is interpretable like linear regression?

Logistic regression to the rescue

Well there is good news for you: we can turn the linear regression method’s outputs into predictions about labels. The technique for doing this is called logistic regression. We will not go into the technicalities, suffice to say that in the simplest case, we take the output from linear regression, which is a number, and predict one label A if the output is greater than zero, and another label B if the output is less than or equal to zero. Actually, instead of just predicting one class or another, logistic regression can also give us a measure of uncertainty of the prediction. So if we are predicting whether a customer will buy a new smartphone this year, we can get a prediction that customer A will buy a phone with probability 90%, but for another, less predictable customer, we can get a prediction that they will *not* buy a phone with 55% probability (or in other words, that they will buy one with 45% probability).

It is also possible to use the same trick to obtain predictions over more than two possible labels, so instead of always predicting either yes or no (buy a new phone or not, fake news or real news, and so forth), we can use logistic regression to identify, for example, handwritten digits, in which case there are ten possible labels.

An example of logistic regression

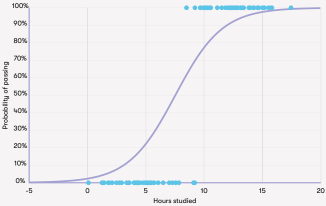
Let’s suppose that we collect data of students taking an introductory course in cookery. In addition to the basic information such as the student ID, name, and so on, we also ask the students to report how many hours they studied for the exam (however you study for a cookery exam, probably cooking?) – and hope that they are more or less honest in their reports. After the exam, we will know whether each student passed the course or not. Some data points are presented below:

Student ID	**Hours studied**	**Pass/fail**
24	15	Pass
41	9.5	Pass
58	2	Fail
101	5	Fail
103	6.5	Fail
215	6	Pass

Based on the table, what kind of conclusion could you draw between the hours studied and passing the exam? We could think that if we have data from hundreds of students, maybe we could see the amount needed to study in order to pass the course. We can present this data in a chart as you can see below.

Unanswered

Exercise 19: Logistic regression



Each dot on the figure corresponds to one student. On the bottom of the figure we have the scale for how many hours the student studied for the exam, and the students who passed the exam are shown as dots at the top of the chart, and the ones who failed are shown at the bottom. We´ ll use the scale on the left to indicate the predicted probability of passing, which we´ ll get from the logistic regression model as we explain just below. Based on this figure, you can see roughly that students who spent longer studying had better chances of passing the course. Especially the extreme cases are intuitive: with less than an hour's work, it is very hard to pass the course, but with a lot of work, most will be successful. But what about those that spend time studying somewhere inbetween the extremes? If you study for 6 hours, what are your chances of passing?

We can quantify the probability of passing using logistic regression. The curve in the figure can be interpreted as the probability of passing: for example, after studying for five hours, the probability of passing is a little over 20%. We will not go into the details on how to obtain the curve, but it will be similar to how we learn the weights in linear regression.

If you wanted to have an 80% chance of passing a university exam, based on the above figure, how many hours should you approximately study for?

Select the correct answer

○ 6-7 hours
○ 7-8 hours
○ 8-9 hours
○ 10-11 hours

Submit

Logistic regression is also used in a great variety of real-world AI applications such as predicting financial risks, in medical studies, and so on. However, like linear regression, it is also constrained by the linearity property and we need many other methods in our toolbox. We will return to the linearity issue later when we discuss neural networks.

The limits of machine learning

To summarize, machine learning is a very powerful tool for building AI applications. In addition to the nearest neighbor method, linear regression, and logistic regression, there are literally hundreds, if not thousands, of different machine learning techniques, but they all boil down to the same thing: trying to extract patterns and dependencies from data and using them either to gain understanding of a phenomenon or to predict future outcomes.

Machine learning can be a very hard problem and we can’t usually achieve a perfect method that would always produce the correct label. However, in most cases, a good but not perfect prediction is still better than none. Sometimes we may be able to produce better predictions by ourselves but we may still prefer to use machine learning because the machine will make its predictions faster and it will also keep churning out predictions without getting tired. Good examples are recommendation systems that need to predict what music, what videos, or what ads are more likely to be of interest to you.

The factors that affect how good a result we can achieve include:

- The hardness of the task: in handwritten digit recognition, if the digits are written very sloppily, even a human can’t always guess correctly what the writer intended
- The machine learning method: some methods are far better for a particular task than others
- The amount of training data: from only a few examples, it is impossible to obtain a good classifier
- The quality of the data

Note

Data quality matters

In the beginning of this chapter, we emphasized the importance of having enough data and the risks of overfitting. Another equally important factor is the **quality** of the data. In order to build a model that generalizes well to data outside of the training data, the training data needs to contain enough information that is relevant to the problem at hand. For example, if you create an image classifier that tells you what the image given to the algorithm is about, and you have trained it only on pictures of dogs and cats, it will assign everything it sees as either a dog or a cat. This would make sense if the algorithm is used in an environment where it will only see cats and dogs, but not if it is expected to see boats, cars, and flowers as well.

We'll return to potential problems caused by ”biased” data.

It is also important to emphasize that different machine learning methods are suitable for different tasks. Thus, there is no single best method for all problems (“one algorithm to rule them all…”). Fortunately, one can try out a large number of different methods and see which one of them works best in the problem at hand.

This leads us to a point that is very important but often overlooked in practice: what it means to work better. In the digit recognition task, a good method would of course produce the correct label most of the time. We can measure this by the classification error: the fraction of cases where our classifier outputs the wrong class. In predicting apartment prices, the quality measure is typically something like the difference between the predicted price and the final price for which the apartment is sold. In many real-life applications, it is also worse to err in one direction than in another: setting the price too high may delay the process by months, but setting the price too low will mean less money for the seller. And to take yet another example, failing to detect a pedestrian in front of a car is a far worse error than falsely detecting one when there is none.

As mentioned above, we can’t usually achieve zero error, but perhaps we will be happy with error less than 1 in 100 (or 1%). This too depends on the application: you wouldn’t be happy to have only 99% safe cars on the streets, but being able to predict whether you’ll like a new song with that accuracy may be more than enough for a pleasant listening experience. Keeping the actual goal in mind at all times helps us make sure that we create actual added value.

After completing Chapter 4 you should be able to:

- Explain why machine learning techniques are used
- Distinguish between unsupervised and supervised machine learning scenarios
- Explain the principles of three supervised classification methods: the nearest neighbor method, linear regression, and logistic regression

Please join the Elements of AI community at [Spectrum]() to discuss and ask questions about this chapter.

You reached the end of Chapter 4!

Correct answers

○ 0 %

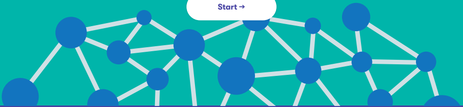
Exercises completed

0 /25

Next Chapter

Neural networks

Start →



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Reaktor