Regression Methods

In this lecture. . .

- Linear and logistic regression

- Suitable cost functions

- Finding parameters by gradient descent

Introduction

Regression methods are supervised-learning techniques. They try to explain a dependent variable in terms of independent variables. The independent variables are numerical and we fit straight lines, polynomials or other functions to predict dependent variables.

The method can also be used for classification, when the dependent variables are usually zeroes and ones.

What Is It Used For?

Regression methods are used for

- Finding numerical relationships between dependent and independent variables based on data
- Classifying data based on a set of numerical features
- Example: Find a relationship between the number of tomatoes on a plant, the ambient temperature and how much they are watered
- Example: Determine the probability of getting cancer given lifestyle choices (quantity of alcohol consumed, cigarettes smoked, etc.)

You will no doubt know about regression from fitting straight lines through a set of data points. For example, you have values and square footage for many individual houses, is there a simple linear relationship between these two quantities? Any relationship is unlikely to be perfectly linear so what is the *best* straight line to put through the points? Then you can move on to higher dimensions. Is there a linear relationship between value and both square footage and number of garage bays?

As an introduction to this supervised-learning technique we shall start with looking for linear relationships in multiple dimensions, and then move on to logistic regression which is good for classification problems.

Linear Regression In Many Dimensions

Linear regression in multiple dimensions is not much harder than in one dimension.

We have M independent, explanatory, variables, the features, $x_1, \ldots, x_m, \ldots, x_M$ and so we write all xs as vectors.

For each of N data points we will have the independent variable $\mathbf{x}^{(n)}$ (square footage of a house, number of garage bays, etc.) and the dependent variable $y^{(n)}$ (property value).

We will fit the linear function $h_\theta(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}$ to the ys, where $\boldsymbol{\theta}$ is the vector of the as-yet-unknown parameters.

One subtle point is that because we usually want to have a parameter θ_0 that doesn't multiply any of the independent variables we write $\mathbf{x}$ as $(1, x_1, \ldots, x_M)^T$, where T means transpose, so it has dimension $M + 1$.

The cost function is usually the quadratic function:

$$J(\boldsymbol{\theta}) = \frac{1}{2N} \sum_{n=1}^{N} \left(h_\theta(\mathbf{x}^{(n)}) - y^{(n)} \right)^2 .$$

We can differentiate this cost function with respect to each of the θs, set the result equal to zero, and solve for the θs.

Easy.

However, although there is still technically an analytic expression for the vector $\boldsymbol{\theta}$ it involves matrix inversion so in practice you might as well use gradient descent.

Finding the parameters numerically using gradient descent

Although numerical methods are not needed when you have linear regression in a single dimension you will need to use some numerical method for anything more complicated.

E.g. batch gradient descent and stochastic gradient descent.

To use these methods you will need

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^{(n)} \left(h_{\theta}(\mathbf{x}^{(n)}) - y^{(n)} \right).$$

For example, the batch gradient descent algorithm is as follows.

The Batch Gradient Descent Algorithm

Step 1: Iterate

New $\boldsymbol{\theta}$ = Old $\boldsymbol{\theta}$ $-\beta\,\partial J/\partial\boldsymbol{\theta}$

I.e. New $\boldsymbol{\theta}$ = Old $\boldsymbol{\theta}$ $-\beta/N\sum_{n=1}^{N}\mathbf{x}^{(n)}\left(h_\theta(\mathbf{x}^{(n)})-y^{(n)}\right)$

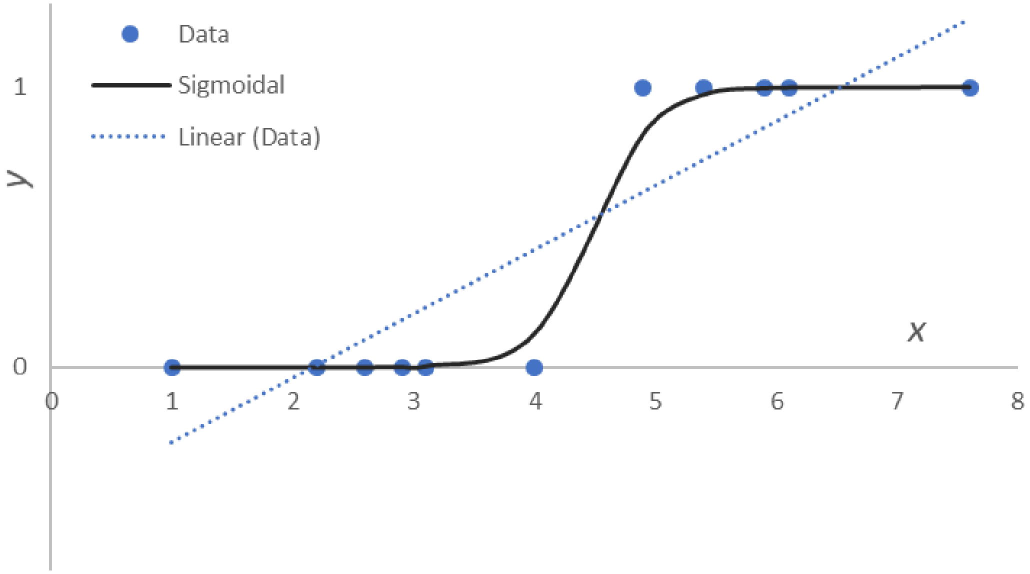
Step 2: Update

Update all θ_k simultaneously. Repeat until convergence

Linear regression is such a common technique that I'm going to skip examples and go straight to something different, regression used for classification.

Logistic Regression

Suppose you want to classify an email as to whether or not it is spam. The independent variables, the xs, might be features such as number of !s or number of spelling mistakes in each email. And your ys will all be either 0, not spam, or 1, spam. Linear regression is not going to do a good job of fitting in this case. Would you want to fit a straight line through this?



One way to see why this is nonsense is to add another dot to this figure, a dot way off to the right, with $y = 1$. It is clearly not conflicting with the data that's already there, if it had $y = 0$ then that might be an issue. But even though it seems to be correctly classified it would cause a linear fit to rotate, to level out and this would affect predictions everywhere.

Also sometimes the vertical axis represents a probability, the probability of being in one class or another. In that case any numbers that you get that are below zero or above 1, which you will get with a linear fit, will also be nonsense.

For classification problems you need something better than linear. We tend to use logistic regression. Instead of using a linear function we use the sigmoidal function

$$h_\theta(\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}}.$$

This function is shown in the previous figure.

We fit this function to the data and then given a new data point we would decide whether it was spam or not according to a **threshold** for h_θ. We might have a threshold of 0.5, so that anything above that level would go into our spam box. If we are worried that genuine emails are going into our spam box then the threshold might be 0.8, say.

The cost function

An important difference when we want to do logistic regression instead of linear regression is in the choice of cost function.

There are some obvious properties for a cost function: It should be positive except when we have a perfect fit, when it should be zero; And it should have a single minimum. Our previous OLS cost function satisfies the first of these but when h_θ is the logistic function it does not necessarily have a single minimum.

However, the following cost function does have these properties.

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)}\ln\left(h_\theta(\mathbf{x}^{(n)})\right) + (1-y^{(n)})\ln\left(1-h_\theta(\mathbf{x}^{(n)})\right)\right). \tag{1}$$

So why does this cost function work?

Remember that y can only take values 0 or 1. When it is one the function of h_θ is smoothly, monotonically decreasing to a value of zero when h_θ is also one. And when $y = 0$ the cost function is also zero when $h_\theta = 0$.

We no longer have an analytic solution for the minimum of the cost function so we have to solve numerically.

But rather conveniently we find that with the new definition for h_θ and the new cost function we still have

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^{(n)} \left(h_\theta(\mathbf{x}^{(n)}) - y^{(n)} \right).$$

I.e. the same as for linear regression and the quadratic cost function.

This means that our gradient descent algorithm remains, surprisingly, unchanged.

Example: Political speeches

I shall take speeches/writings by N politicians. And I shall label each politician as either '0' for left wing or '1' for right wing. These are the $y^{(n)}$ for $n = 1, \ldots, N$.

I shall then look at the *types* of words used, and whether they are the words positive words, negative words, irregular verbs, etc. for a total of M features.

So the n^{th} politician is represented by $\mathbf{x}^{(n)}$, a vector of length $M + 1$.

The first entry is 1. The second entry would be the fraction of positive words the n^{th} politician uses, the third entry the fraction of negative words, and so on.

But how do I know whether a word is positive, negative, etc.? For this I need a special type of dictionary used for such textual analysis.

One such dictionary is the Loughran–McDonald dictionary. This is a list of words classified according various categories.

These categories are: Negative; Positive; Uncertainty; Litigious; Constraining; Superfluous; Interesting; Modal; Irregular Verb; Harvard IV; Syllables. This list is often used for financial reporting (hence the 'litigious' category!). Most of the category meanings are clear. 'Modal' concerns degree from words such as 'always' (strong modal, 1) through 'can' (moderate, 2) to 'might' (weak, 3). Harvard IV is a Psychosociological Dictionary.

The results of the analysis of the speeches is shown below. The fractions of positive, negative, etc. words is very small because the vast majority of words in the dictionary I used are not positive, negative, etc.

	Benn	**Church.**	**Corbyn**	**JFK**	**M&E**	**May**	**Thatch.**	**Trump**
Negative	0.501%	1.048%	1.216%	0.562%	2.067%	0.395%	1.626%	1.261%
Positive	0.213%	0.380%	0.517%	0.243%	0.760%	0.137%	1.018%	1.048%
Uncertainty	0.395%	0.441%	0.274%	0.137%	0.289%	0.091%	0.425%	0.289%
Litigious	0.152%	0.061%	0.274%	0.213%	0.441%	0.091%	0.304%	0.228%
Constraining	0.000%	0.030%	0.137%	0.122%	0.304%	0.122%	0.046%	0.122%
Superfluous	0.000%	0.000%	0.000%	0.000%	0.030%	0.000%	0.000%	0.000%
Interesting	0.030%	0.061%	0.000%	0.000%	0.152%	0.030%	0.106%	0.061%
Modal	1.307%	1.975%	1.945%	0.881%	1.276%	0.380%	2.097%	2.112%
Irregular vb	0.228%	0.745%	0.608%	0.334%	0.988%	0.289%	0.942%	0.699%
Harvard IV	1.352%	3.206%	4.513%	1.869%	7.765%	1.596%	4.270%	4.255%
Syllables	1.52	1.44	1.57	1.41	1.68	1.68	1.50	1.49

Because I was only using eight politicians for the training I did not use all 11 of these categories. If I did so then I would probably get a meaningless perfect fit. (Twelve unknowns and eight equations.) So I limited the features to just positive, negative, irregular verbs and syllables.

I found that the θs for positive words, irregular verbs and number of syllables were all positive, with the θ for negative words being negative.

Finally I thought I would classify my own writing. So I took a sample from that other famous political text *Paul Wilmott On Quantitative Finance, second edition*. And I found that I was as right wing as Margaret Thatcher.

If you were to do this for real you would use the above methodology but improve on its implementation in many ways:

- More training data, meaning many, many more speeches/writings from a larger selection of politicians

- Use a better dictionary, one more suited to classifying politicians

Other Regression Methods

You can go a lot further with regression methods than the techniques I've covered here. Just briefly. . .

The usual suspects

There are many (basis) functions in common use. You will no doubt have used some yourself. For example, Fourier series, Legendre polynomials, Hermite polynomials, radial basis functions, wavelets, etc. All might have uses in regression. Some are particularly user friendly being orthogonal (the integral of their products over some domain, perhaps with a weighting function, is zero).

Polynomial regression

Rather obviously fit a polynomial in the independent variables. However, the higher the degree of the polynomial the greater the danger of overfitting.

You can use OLS again to find the parameters by treating each non-linear term as if it were another independent variable. So a polynomial fit in one dimension becomes a linear fit in higher dimensions. Instead of

$$\theta_0 + \theta_1 x + \theta_2 x^2$$

think of

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2.$$

Because of the obvious correlation between x and x^2 it can be tricky to interpret exactly what the coefficients in the polynomial mean.

Ridge regression

One sometimes adds a regularization term to OLS:

$$J(\boldsymbol{\theta}) = \frac{1}{2N}\left(\sum_{n=1}^{N}\left(h_\theta(x^{(n)}) - y^{(n)}\right)^2 + \lambda|\bar{\boldsymbol{\theta}}|^2\right).$$

Here I am using $\bar{\boldsymbol{\theta}}$ to mean the vector with the first entry being zero (i.e. there is no θ_0). And this is the L^2 or Euclidean norm. This extra penalty term has the effect of reducing the size of the (other) coefficients.

Why would we want to do this?

Generally it is used when there is quite a strong relationship between several of the factors. Fitting to both height and age might be a problem because height and age are strongly related. An optimization without the regularization term might struggle because there won't, in the extreme case of perfect correlation, be a unique solution. Regularization avoids this and would balance out the coefficients of the related features.

Lasso regression

Or rather LASSO for Least Absolute Shrinkage and Selection Operator. This is similar to ridge regularization but the penalty term is now the L^1 norm, the sum of the absolute values rather than the sum of squares.

Not only does Lasso regression shrink the coefficients it also has a tendency to set some coefficients to zero, thus simplifying models.

Summary

Please take away the following important ideas

- There are different types of regression besides linear

- For classification you will usually use sigmoidal functions

- Paul Wilmott is as right wing as Margaret Thatcher