Logistic Regression Gradient Ascent

Lecture # 11

Logistic Regression: Definition

Regression is finding the best-fitting line to a dataset. This is mainly an optimization problem. Logistic regression is when we have a bunch of data and with the data we try to build an equation to do classification for us. To do this, we try to find the best-fit set of parameters. Finding the "best fit" is regression and in this way we train our classifier.

Regression can be defined as the relation between a dependent variable, y, and a set of independent variables, x, that describes the expectation value of y given x: E[y|x].

The Regression Problem

We define three types of regression:

Linearity: when a parametric model is linear in all model parameters. A linear regression can be defined as:

$$f(x|\theta) = \sum_{p=1}^{k} \theta_p g_p(x)$$

where functions g_p (x) do not depend on any free model parameters- this is linear regression. Regressions of models that include nonlinear dependence on θ_p such as

$f(x|\theta) = \theta_1 + \theta_2 \sin(\theta_3 x)$ is called nonlinear regression.

Complexity: The complexity of the error covariance matrix increases by increasing the number of independent variables. This is the limiting factor in nonlinear regression.

Error behavior: the uncertainties in the values of independent and dependent variables, and their correlations, are the main factors that determine which regression method to use. The structure of error covariance matrix and deviations from Gaussian error behavior, could make the problem complicated.

Regression for Linear Models

The simplest case for regression is the case of a linear model where an independent variable x and a dependent variable y, are considered defining the linear model

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i$$

Where θ_0 and θ_1 are the regression coefficients that we are trying to estimate and ε_i is the additive noise. We could write the data likelihood as:

$$p(\{y_i\}|\{x_i\}, \theta, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(\frac{-(y_i-(\theta_0+\theta_1 x_i))2}{2\,\sigma i^2}\right)$$

By taking the logarithm of this posterior, we arrive at the classic definition of regression in terms of log-likelihood

$$\ln(L) = \ln\big((\theta|\{x_i, y_i\}, I)\big) \propto \sum_{i=1}^{N}\left(\frac{-\big(y_i - (\theta_0 + \theta_1 x_i)\big)^2}{2\sigma_i^2}\right)$$

Maximizing the log-likelihood as a function of the model parameters, θ, is achieved by minimizing the sum of the square errors

$$\ln(L) \propto \sum_{i=1}^{N} \frac{-|y_i - (\theta_0 + \theta_1 x_i)|}{\sigma_i}$$

For Gaussian uncertainties the minimization of this equation results in

$$\theta_1 = \frac{\sum_{i=1}^{N} x_i y_i - \bar{x}\,\bar{y}}{\sum_{i=1}^{N} (x_i - \bar{x})\text{^}2}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

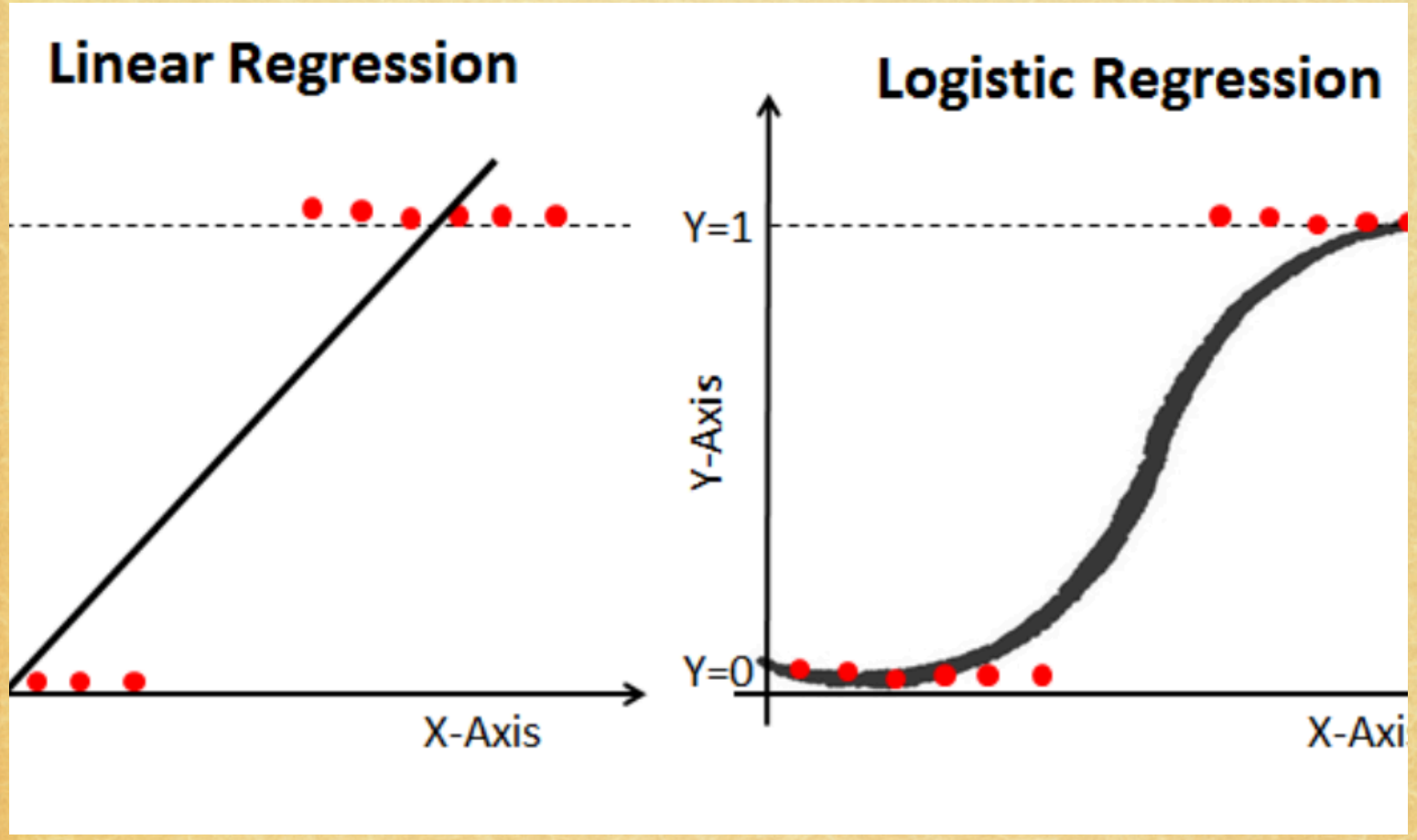
Where $\bar{x}$ and $\bar{y}$ are the mean values of x and y respectively.

Logistic Regression

Linear regression is the process of finding a function to fit the x's that vary linearly with y with the objective of being able to use the function as a model for prediction. The key assumption here is that both the predictor and target variables are continuous. In other words, when x increases, y also increases along the slope of the line.

Now, what happens if the target variable is not continuous? Suppose the target variable is the response to advertisement campaigns- if more than a threshold number of customers buy, for example, then the response is considered to be 1; if not, the response is 0.

Linear vs. Logistic Regression



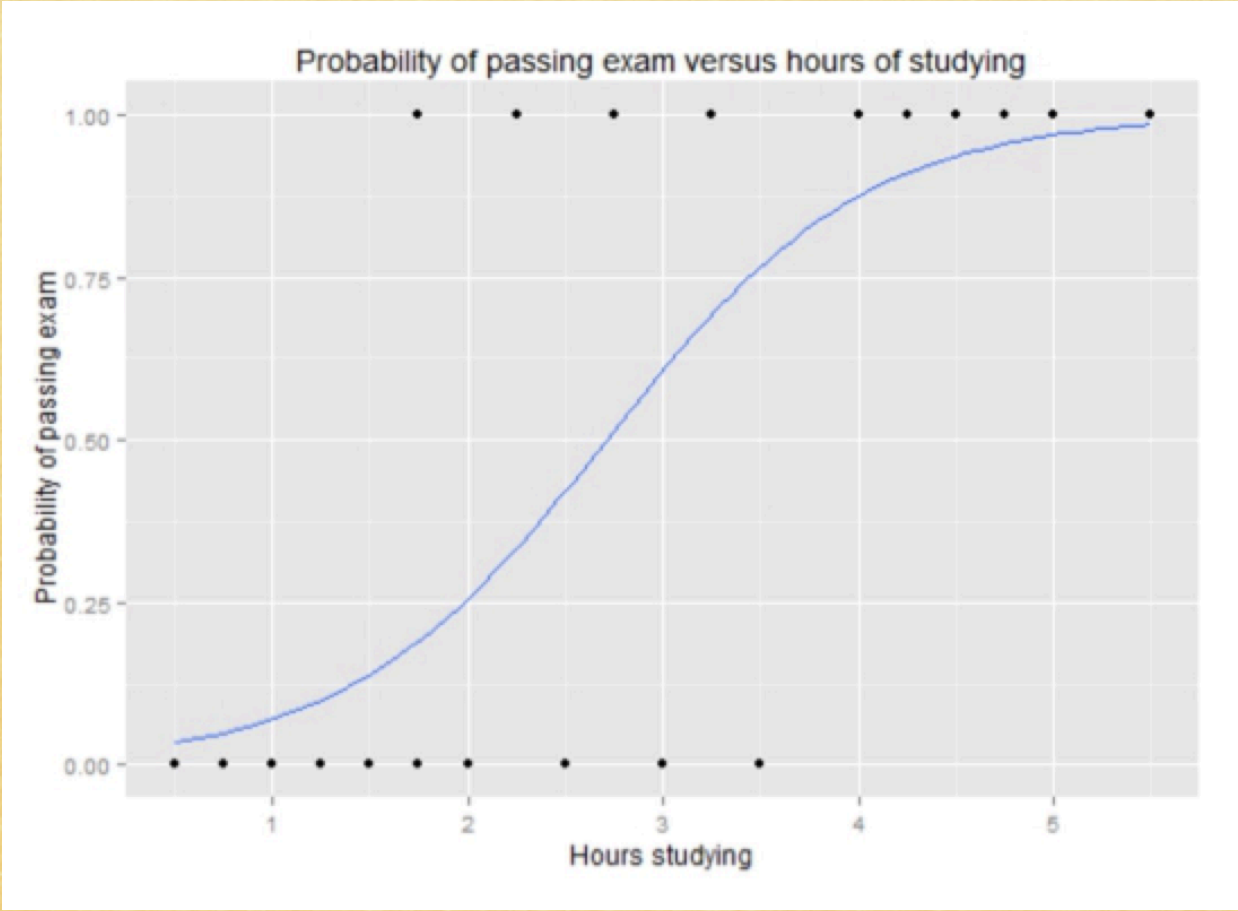
Classification with Logistic Regression

Regression is basically finding the best-fitting line to a dataset. This is mainly an optimization problem. Logistic regression is when we have a bunch of data and with the data we try to build an equation to do classification for us. To do this, we try to find the best-fit set of parameters. Finding the “best fit” is regression and in this way we train our classifier.

In the case the response Y is discrete, the straight line is no longer a fit. There is no gradual transition as the Y value abruptly jumps from one binary outcome to another. The straight line is therefore a poor fit for these data. A better fit would be an S-shaped curve. If the equation to this sigmoid curve is known, then it can be used effectively as the straight line in the case of linear regression.

Logistic regression is the process of obtaining an appropriate non-linear curve to fit the data when the target variable is discrete. How is the sigmoid curve obtained? How does it relate to the predictors?

Probability of passing an exam as a function of hours studied



How Logistic Regression finds the Sigmoid Curve?

In case of linear regression, we only need two parameters, slope (b_1) and zero-point (b_0). This completely specifies the way the independent and dependent variables are related to one another. It is more complicated however, to find the parameters that fit an S-shaped curve.

Often when the use of logistic regression is needed, the "y" is a yes/no type of response. This is interpreted as the probability of an event happening (y=1) or not happening (y=0). This can be constructed as follows:

- If y is an event (response, pass/fail etc)
- And p is the probability of the event happening (y=1)
- Then (1 - p) is the probability of the event not happening (y=0)
- And p/(1-p) is the odds of the event happening

The logarithm of the odds, log (p/(1-p)) is linear in the predictor x, with log (p/(1-p)) – log of the odds- called *the logit* function.

The logit can be specified as a linear function of the predictor, x, similar to the linear regression model shown before.

Logit = log p/(1-p) = b_0 x + b_1 (1)

And for a more general case involving multiple independent variables, it is

Logit = $b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$ (2)

For each row of the predictor, the logit can now be computed. From the logit, it is now easy to compute the probability of the response y (happening or not happening) as

$P = e^{logit} / (1 + e^{logit})$ (3)

The logistic regression then calculates the probability of y happening (i.e. y=1)- equation 3, given specific values of x from equation 2.

Logistic regression can be defined as a mathematical modeling approach in which a best-fitting model is selected to describe the relationship between several independent variables and a dependent binomial response variable.

From the data given, the x's are known and using equations 2 and 3, we can compute p for any value of x. In order to do this however, the coefficients in equation 2 (b values) need to be determined. Using a training sample, one could compute the relation

$$p^{\gamma} . (1-p)^{(1-\gamma)}$$

Where γ is the original outcome variable (which can be 0 or 1). And p is the probability estimated by the logit equation. For a specific training sample, if the actual outcome was y=0 and the model estimate of p was high (say 0.9)- that is, the model was wrong- this quantity reduces to 0.1. This quantity is a simplified form of the likelihood function, is maximized for good estimates and minimized for poor estimates. If one calculates a summation of the likelihood function across all the training data, then a high value indicates a good model or vice versa.

If X is the independent variable and Y a dependent variable, how can we measure the probability of Y being 1 (or 0)- (P(Y=1 | X) as a function of X?

The linear regression models these probabilities as: $p(X)=\beta 0 + \beta 1X$

The logistic regression equation is derived from the same equation except that the dependent variable must only have categorical values. Logistic Regression does not calculate the outcome as 0 or 1, instead, it calculates the probability (ranges between 0 and 1) of a variable falling in class 0 or class 1. Thus, we can conclude that the dependent variable must be positive and it should lie between 0 and 1 i.e. it must be less than 1. In order to meet the above-mentioned conditions, we must do the following:

- Take the exponent of the equation, since the exponential of any value is a positive number
- Secondly, a number divided by itself + 1 will always be less than 1

$$P(X) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta x)} + 1}$$

Derivation of *logit*

From Edureka web page.

$$P(X) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta x)} + 1}$$

$$\Longrightarrow \quad p(e^{(\beta_0+\beta x)}+1) = e^{(\beta_0+\beta_1 x)}$$

$$\Longrightarrow \quad p.\, e^{(\beta_0+\beta x)} + p = e^{(\beta_0+\beta_1 x)}$$

$$\Longrightarrow \quad p = e^{(\beta_0+\beta_1 x)} - p.\, e^{(\beta_0+\beta x)}$$

$$\Longrightarrow \quad p = e^{(\beta_0+\beta_1 x)}\,(1-p)$$

$$\Longrightarrow \quad \frac{p}{(1-p)} = e^{(\beta_0+\beta_1 x)}$$

$$\Longrightarrow \quad \ln\left[\frac{p}{(1-p)}\right] = (\beta_0 + \beta_1 x)$$

Consider the sigmoid function

$\sigma(z) = 1 / (1 + e^{-z})$

At z=0, the value of the sigmoid is 0.5. For increasing values of z it approaches unity and for decreasing values it approaches 0.

For the logistic regression classifier, we will take our features and multiply each one by a weight and then add them up. This result will be put into the sigmoid, giving a number between 0 and 1. Anything above 0.5 will be classified a 1 and anything below 0.5 will be classified as a 0. The question now is what is the best weight or regression coefficients to use? How do we find them?

Often gradient decent or other non-linear optimization methods is used to search for coefficients, b, with the objective of maximizing the likelihood of correct estimation $p^{\gamma} . (1-p)^{(1-\gamma)}$ summed over all training samples.

Optimization

The input to the sigmoid function will be z, given by the relation

$z = w_0 x_0 + w_1 x_1 + \ldots + w_n x_n$

In vector notation this can be written as

$z = W^T X$

The vector X is our input data and we want to find the best coefficient w, so that this classifier will be as successful as possible. In order to do that, we need to use ideas from optimization theory. We use optimization with gradient ascent. We will then see how we can use this method to find the best parameters to model our dataset.

Gradient Ascent

Gradient Ascent

Gradient ascent is an optimization algorithm. This is based on the idea that if we want to find the maximum point on a function, then the best way to move is in the direction of the gradient. We denote the gradient with symbol ∇ and the gradient of a function f(x,y) by the equation

$$\nabla f(x,y) = \begin{matrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{matrix}$$

This gradient means that we are moving in the x direction by the amount ∂f(x,y)∂x and in the y direction by the amount ∂f(x,y)∂y. Obviously the function f(x,y) must be defined and differentiable around the point it is calculated.

The gradient ascent shown in the figure takes a step in the direction given by the gradient. The gradient operator always points in the direction of greatest increase. Its magnitude is the step-size, α. In vector notation this is defined as

$$w := w + \alpha \nabla_w f(w)$$

This step is repeated until we reach a stopping condition- either a specific number of steps or the algorithm is within a certain tolerance margin.

Gradient decent is similar to gradient ascent with the positive sign changed to negative

$$w := w - \alpha \nabla_w f(w)$$

with the gradient descent we are trying to minimize some function rather than maximize it

Gradient Descent

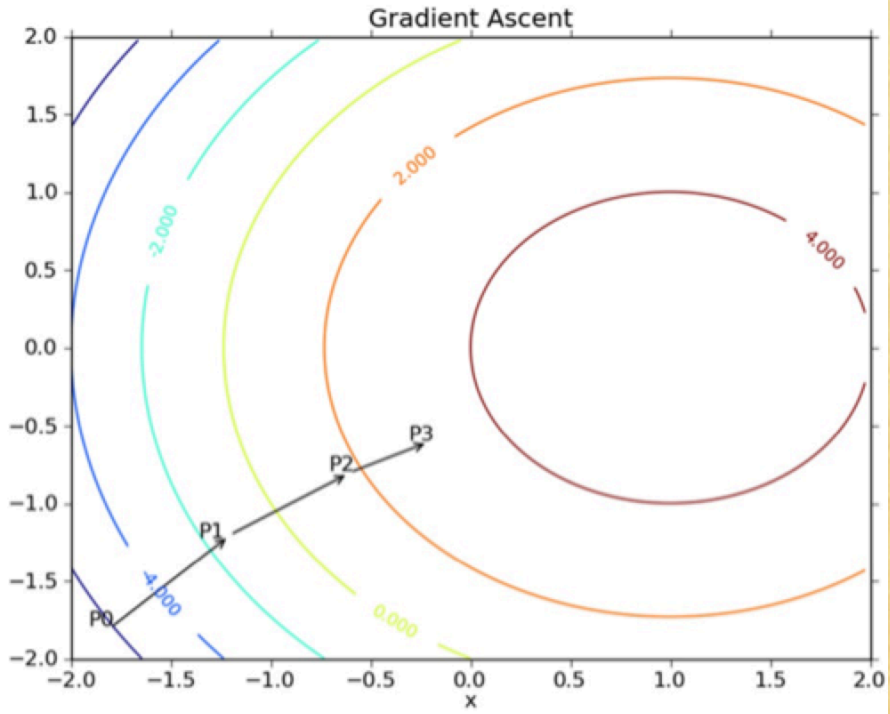


Figure 5.2 The gradient ascent algorithm moves in the direction of the gradient evaluated at each point. Starting with point P0, the gradient is evaluated and the function moves to the next point, P1. The gradient is then reevaluated at P1, and the function moves to P2. This cycle repeats until a stopping condition is met. The gradient operator always ensures that we're moving in the best possible direction.

Examples of Sigmoid Function

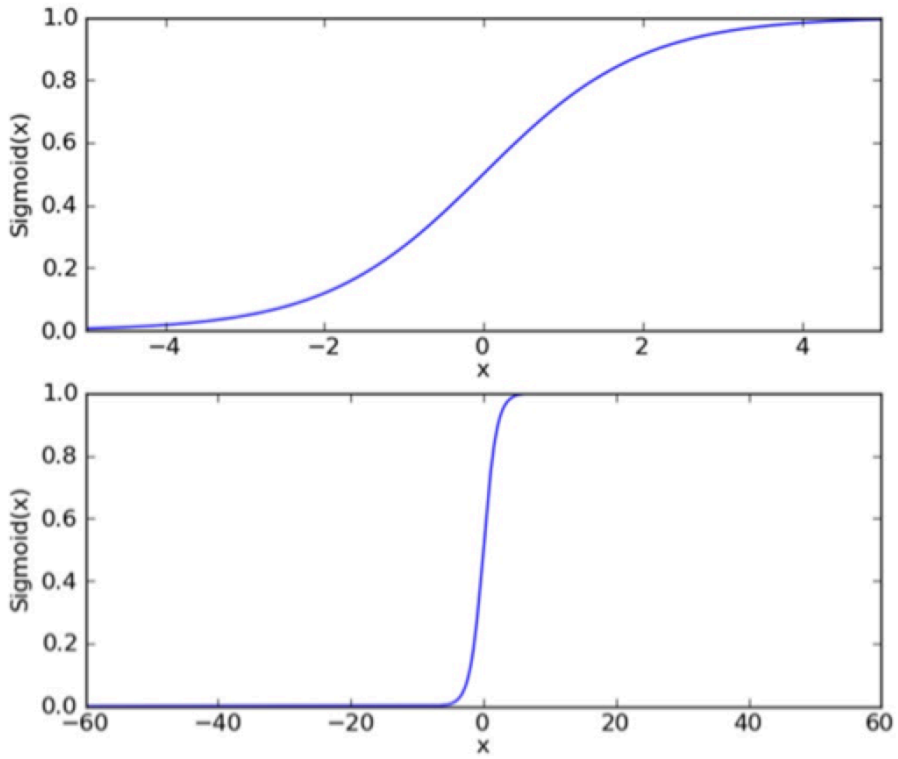


Figure 5.1 A plot of the sigmoid function on two scales; the top plot shows the sigmoid from -5 to 5, and it exhibits a smooth transition. The bottom plot shows a much larger scale where the sigmoid appears similar to a step function at x=0.

Example for Logistic Regression

Probability of passing an exam versus hours of study

Suppose we wish to answer the following question: A group of 20 students spend between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability that the student will pass the exam?

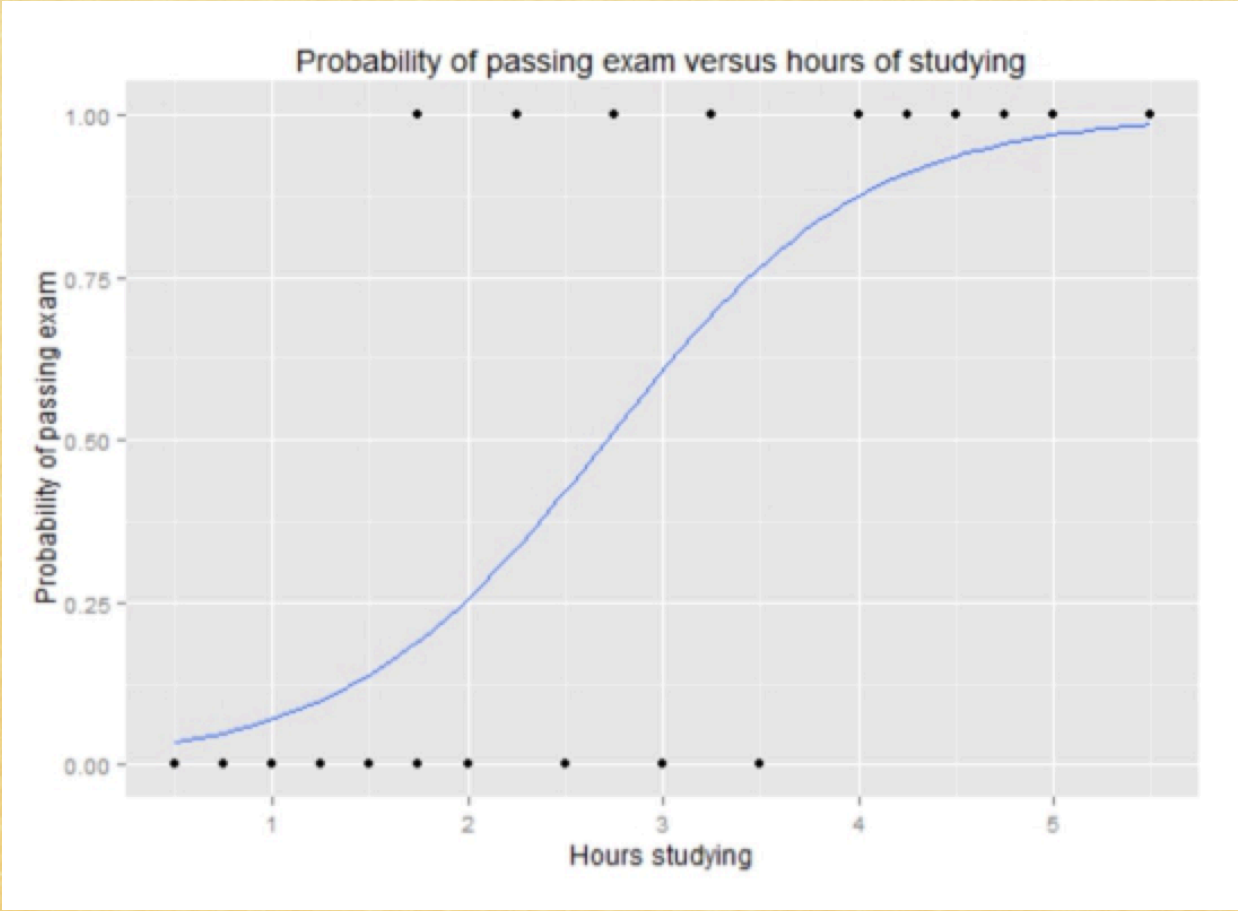
The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0". If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then regression analysis could be used. The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0). The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.

Hours	0.50	0.75	1.	1.25	1.5	1.75	2.	2.25	2.50	2.75	3.	3.25	3.50	4.	4.25	4.50	4.75	5.
Pass	0	0	0	0	0	0	0	1	0	1	1	1	0	1	1	1	1	1

The logistic regression analysis gives the following output.

	Coefficient	Std.Error	z-value	P-value (Wald)
Intercept	–4.0777	1.7610	–2.316	0.0206
Hours	1.5046	0.6287	2.393	0.0167

Probability of passing an exam as a function of hours studied



The output indicates that hours studying is significantly associated with the probability of passing the exam as indicated by P values. Here we have the coefficients for the intercept -4.0777 and Hours = 1.5046. These coefficients are entered into the logistic regression equation to estimate the odds (or probabilities) of passing the exam

Log-odds of passing exam = 1.5046 . Hours – 4.0777 = 1.5046 . (Hours – 2.71)

Odds of passing exam = exp (1.5046 . (Hours – 2.71))

$$probability\ of\ passing\ exam = \frac{1}{1 + \exp\left(-(1.5046\,.Hours - 4.0777)\right)}$$

One additional hour of study will increase log-odds of passing the exam by 1.5046 or odds of passing exam by exp(1.5046) = 4.5.

For a student who studied 2 hours, entering the value Hourse=2 in the equation gives the estimated probability of passing the exam of 0.26.

Sources used for this lecture

Machine Learning In Action

By Peter Harrington

Statistics, Data Mining, and Machine Learining in Astronomy

Z. Ivezic, A. Connolly, J. VanderPlas & A. Gray