LOGISTIC REGRESSION

What do you mean by the Logistic Regression?

- It's a classification algorithm that is used where the target variable is of categorical nature. The main objective behind Logistic Regression is to determine the relationship between features and the probability of a particular outcome.
- For Example, when we need to predict whether a student passes or fails in an exam given the number of hours spent studying as a feature, the target variable comprises two values i.e. pass and fail.

What are the different types of Logistic Regression?

Three different types of Logistic Regression are as follows:

1. **Binary Logistic Regression**: In this, the target variable has only two 2 possible outcomes. For Example, 0 and 1, or pass and fail or true and false.
2. **Multinomial Logistic Regression:** In this, the target variable can have three or more possible values without any order. For Example, predicting preference of food i.e. Veg, Non-Veg, Vegan.
3. **Ordinal Logistic Regression:** In this, the target variable can have three or more values with ordering. For Example, Movie rating from 1 to 5.

What are the odds?

Odds are defined as the ratio of the probability of an event occurring to the probability of the event not occurring.

For Example, let's assume that the probability of winning a game is 0.02. Then, the probability of not winning is 1- 0.02 = 0.98.

- The odds of winning the game= (Probability of winning)/(probability of not winning)
- The odds of winning the game= 0.02/0.98
- The odds of winning the game are 1 to 49, and the odds of not winning the game are 49 to 1.

Is the decision boundary Linear or Non-linear in the case of a Logistic Regression model?

- The decision boundary is a line or a plane that separates the target variables into different classes that can be either linear or nonlinear. In the case of a Logistic Regression model, the decision boundary is a straight line.
- Logistic Regression model formula = α+1X1+2X2+….+kXk. This clearly represents a straight line.
- It is suitable in cases where a straight line is able to separate the different classes. However, in cases where a straight line does not suffice then nonlinear algorithms are used to achieve better results.

What is the difference between the outputs of the Logistic model and the Logistic function?

The Logistic model outputs the logits, i.e. log-odds; whereas the Logistic function outputs the probabilities.

- Logistic model = α+1X1+2X2+….+kXk. Therefore, the output of the Logistic model will be logits.
- Logistic function = f(z) = 1/(1+e-(α+1X1+2X2+….+kXk)). Therefore, the output of the Logistic function will be the probabilities.

How do we handle categorical variables in Logistic Regression?

- The inputs given to a Logistic Regression model need to be numeric. The algorithm cannot handle categorical variables directly. So, we need to convert the categorical data into a numerical format that is suitable for the algorithm to process.
- Each level of the categorical variable will be assigned a unique numeric value also known as a dummy variable. These dummy variables are handled by the Logistic Regression model in the same manner as any other numeric value.

What are the assumptions made in Logistic Regression?

1. It assumes that there is minimal or no multicollinearity among the independent variables i.e, predictors are not correlated.
2. There should be a linear relationship between the logit of the outcome and each predictor variable. The logit function is described as logit(p) = log(p/(1-p)), where p is the probability of the target outcome.
3. Sometimes to predict properly, it usually requires a large sample size.
4. The Logistic Regression which has binary classification i.e, two classes assume that the target variable is binary, and ordered Logistic Regression requires the target variable to be ordered. For example, Too Little, About Right, Too Much.
5. It assumes there is no dependency between the observations.

Can we solve the multiclass classification problems using Logistic Regression? If Yes, then How?

Yes, in order to deal with multiclass classification using Logistic Regression, the most famous method is known as the one-vs-all approach.

- In this approach, a number of models are trained, which is equal to the number of classes.
- For Example,
    - the first model classifies the datapoint depending on whether it belongs to class 1 or some other class (not class 1)
    - the second model classifies the datapoint into class 2 or some other class (not class 2) and so-on for all other classes.
    - So, in this manner, each data point can be checked over all the classes.

How can we express the probability of a Logistic Regression model as conditional probability?

- We define probability P(Discrete value of Target variable | X1, X2, X3…., Xk) as the probability of the target variable that takes up a discrete value (either 0 or 1 in the case of binary classification problems) when the values of independent variables are given.
- For Example, the probability an employee will attain (target variable) given his attributes such as his age, salary, etc.

Why can't we use Linear Regression in place of Logistic Regression for Binary classification?

Linear Regressions cannot be used in the case of binary classification due to the following reasons:

1. **Distribution of error terms:** The distribution of data in the case of Linear and Logistic Regression is different. It assumes that error terms are normally distributed. But this assumption does not hold true in the case of binary classification.
2. **Model output:** In Linear Regression, the output is continuous (or numeric) while in the case of binary classification, an output of a continuous value does not make sense. For binary classification problems, Linear Regression may predict values that can go beyond the range between 0 and 1. In order to get the output in the form of probabilities, we can map these values to two different classes, then its range should be restricted to 0 and 1. As the Logistic Regression model can output probabilities with Logistic or sigmoid function, it is preferred over linear Regression.
3. **The variance of Residual errors:** Linear Regression assumes that the variance of random errors is constant. This assumption is also not held in the case of Logistic Regression.

What is a logistic function? What is the range of values of a logistic function?

f(z) = 1/(1+e -z) The values of a logistic function will range from 0 to 1. The values of Z will vary from -infinity to +infinity.

Why is logistic regression very popular?

Logistic regression is famous because it can convert the values of logits (logodds), which can range from -infinity to +infinity to a range between 0 and 1. As logistic functions output the probability of occurrence of an event, it can be applied to many real-life scenarios. It is for this reason that the logistic regression model is very popular.

What is the formula for the logistic regression function?

f(z) = 1/(1+e-(α+1X1+2X2+….+kXk))

What are the outputs of the logistic model and the logistic function?

The logistic model outputs the logits, i.e. log odds; and the logistic function outputs the probabilities.

1. **Logistic model =** α+1X1+2X2+….+kXk. The output of the same will be logits.
2. **Logistic function =** f(z) = 1/(1+e-(α+1X1+2X2+….+kXk)). The output, in this case, will be the probabilities.

How to interpret the results of a logistic regression model? Or, what are the meanings of alpha and beta in a logistic regression model?

- Alpha is the baseline in a logistic regression model. It is the log odds for an instance when all the attributes (X1, X2,…………Xk) are zero. In practical scenarios, the probability of all the attributes being zero is very low. In another interpretation, Alpha is the log odds for an instance when none of the attributes is taken into consideration.
- Beta is the value by which the log odds change by a unit change in a particular attribute by keeping all other attributes fixed or unchanged (control variables).

What is odds ratio?

Odds ratio is the ratio of odds between two groups. For example, let's assume that we are trying to ascertain the effectiveness of a medicine. We administered this medicine to the 'intervention' group and a placebo to the 'control' group.

Odds ratio (OR) = (odds of the intervention group)/(odds of the control group)

Interpretation

- If odds ratio = 1, then there is no difference between the intervention group and the control group
- If odds ratio is greater than 1, then the control group is better than the intervention group
- If odds ratio is less than 1, then the intervention group is better than the control group.

Is the decision boundary linear or nonlinear in the case of a logistic regression model?

- The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.
- Logistic regression model formula = α+1X1+2X2+….+kXk. This clearly represents a straight line. Logistic regression is only suitable in such cases where a straight line is able to separate the different classes. If a straight line is not able to do it, then nonlinear algorithms should be used to achieve better results.

What is the Maximum Likelihood Estimator (MLE)?

The MLE chooses those sets of unknown parameters (estimator) that maximise the likelihood function. The method to find the MLE is to use calculus and setting the derivative of the logistic function with respect to an unknown parameter to zero, and solving it will give the MLE. For a binomial model, this will be easy, but for a logistic model, the calculations are complex. Computer programs are used for deriving MLE for logistic models.

Here's another approach to answering the question.

MLE is a statistical approach to estimating the parameters of a mathematical model. MLE and ordinary square estimation give the same results for linear regression if the dependent variable is assumed to be normally distributed. MLE does not assume anything about independent variables.

What are the different methods of MLE and when is each method preferred?

In case of logistics regression, there are two approaches of MLE. They are conditional and unconditional methods. Conditional and unconditional methods are algorithms that use different likelihood functions. The unconditional formula employs joint probability of positives (for example, churn) and negatives (for example, non-churn). The conditional formula is the ratio of the probability of observed data to the probability of all possible configurations.

The unconditional method is preferred if the number of parameters is lower compared to the number of instances. If the number of parameters is high compared to the number of instances, then conditional MLE is to be preferred. Statisticians suggest that conditional MLE is to be used when in doubt. Conditional MLE will always provide unbiased results.

What are the advantages and disadvantages of conditional and unconditional methods of MLE?

Conditional methods do not estimate unwanted parameters. Unconditional methods estimate the values of unwanted parameters also. Unconditional formulas can directly be developed with joint probabilities. This cannot be done with conditional probability. If the number of parameters is high relative to the number of instances, then the unconditional method will give biased results. Conditional results will be unbiased in such cases.

What is the output of a standard MLE program?

The output of a standard MLE program is as follows:

Maximized likelihood value: This is the numerical value obtained by replacing the unknown parameter values in the likelihood function with the MLE parameter estimator.

Estimated variance-covariance matrix: The diagonal of this matrix consists of estimated variances of the ML estimates. The off diagonal consists of the covariances of the pairs of the ML estimates.

Why can't we use Mean Square Error (MSE) as a cost function for logistic regression?

In logistic regression, we use the sigmoid function and perform a non-linear transformation to obtain the probabilities. Squaring this non-linear transformation will lead to non-convexity with local minimums. Finding the global minimum in such cases using gradient descent is not possible. Due to this reason, MSE is not suitable for logistic regression. Cross-entropy or log loss is used as a cost function for logistic regression. In the cost function for logistic regression, the confident wrong predictions are penalised heavily. The confident right predictions are rewarded less. By optimising this cost function, convergence is achieved.

How to choose a cutoff point in case of a logistic regression model?

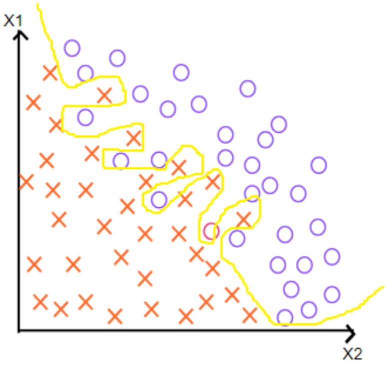
The cutoff point depends on the business objective. Depending on the goals of your business, the cutoff point needs to be selected. For example, let's consider loan defaults. If the business objective is to reduce the loss, then the specificity needs to be high. If the aim is to increase profits, then it is an entirely different matter. It may not be the case that profits will increase by avoiding giving loans to all predicted default cases. But it may be the case that the business has to disburse loans to default cases that are slightly less risky to increase the profits. In such a case, a different cutoff point, which maximises profit, will be required. In most of the instances, businesses will operate around many constraints. The cutoff point that satisfies the business objective will not be the same with and without limitations. The cutoff point needs to be selected considering all these points. As a thumb rule, choose a cutoff value that is equivalent to the proportion of positives in a dataset.

How can we avoid over-fitting in regression models?

Your answer: Regularization technique can be used to avoid over-fitting in regression models. The basic idea is to penalize the complex models, i.e, adding a complexity term that would give a bigger loss for complex models.

Let's say that we have a decision boundary given by the equation:

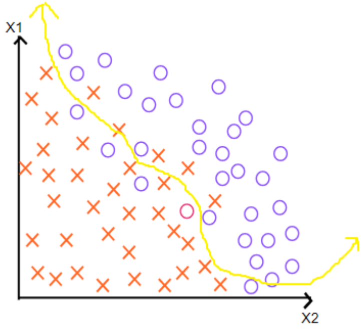
$$W_0 + W_1X_1 + W_2X_2^2 + W_3X_1^3 + W_4X_2^4 = 0$$



As we can clearly see that the model is over-fitting. so we want to get rid of w3 and w4. Then the equation becomes:

$$W_0 + W_1X_1 + W_2X_2^2 = 0$$

Now with this equation boundary may look like following:



So we ended with a quadratic function which is essentially good. In this specific example we saw effect of penalizing the 2 parameters.

Now, if we consider a dataset with following parameters :

Features : X1, X2, X3, ……………, X500

Weights : w0, w1, w2, ………………, w500

In such type of problems we actually don't know that which of the weights we should penalize in order to get smoother curve. So we actually penalize all the weights.

Then the optimization equation changes to ,

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} |w_i|$$

in case of Lasso regularization, and

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} w_i^2$$

in case of Ridge regularization,

where λ =regularization parameter (Hyper parameter).

Regularization parameter will control trade-off between two different objectives. The first objective is we want to fit the training data by adding polynomial features and second objective is we want to keep the weights small, which makes the hypothesis simpler.

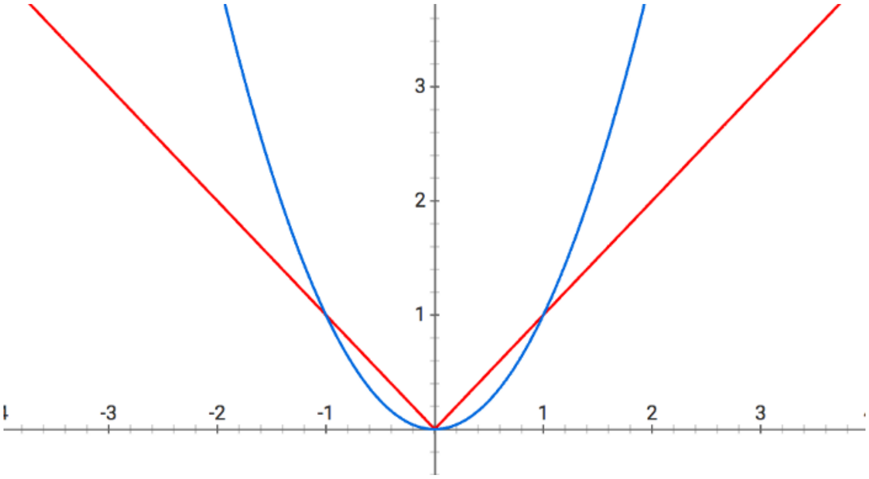
What is the key difference between ridge and lasso regularization?

Ridge regression adds “squared magnitude” of coefficient as penalty term to the loss function whereas Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds “absolute value of magnitude” of coefficient as penalty term to the loss function.

Lasso regression tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero. The key difference between these techniques is that as Lasso shrinks the less important feature’s coefficient to zero thus, it removes some feature altogether. So, this works well for feature selection in case we have a huge number of features.

In the context of L1-regularization(lasso), that the coefficients are pulled towards zero proportionally to their absolute values — they lie on the red curve.

In the context of L2-regularization(ridge), the coefficients are pulled towards zero proportionally to their squares — the blue curve.



How do you implement multinomial logistic regression?

The multinomial logistic classifier can be implemented using a generalization of the sigmoid, called the softmax function. The softmax represents each class with a value in the range (0,1), with all the values summing to 1. Alternatively, you could use the one-vs-all or one-vs-one approach using multiple simple binary classifiers.

Suppose that you are trying to predict whether a consumer will recommend a particular brand of chocolate or not. Let us say your hypothesis function outputs h(x)=0.55 where h(x) is the probability that y=1 (or that a consumer recommends the chocolate) given any input x. Does this mean that the consumer will recommend the chocolate?

The answer to this question is 'cannot be determined.' And this will remain the case unless you are provided additional data on the decision boundary. Let us say that you set the decision boundary such that $y=1$ is $h(x)\geq 0.5$ and 0; otherwise, then the answer for this question would be a resounding YES. However, if you set the decision boundary (although this is not very common practice) such that $y=1$ is $h(x)\geq 0.6$ and 0, otherwise the answer will be a NO.

If you observe that the cost function decreases rapidly before increasing or stagnating at a specific high value, what could you infer?

A trend pattern of the cost curve exhibiting a rapid decrease before then increasing or stagnating at a specific high value indicates that the learning rate is too high. The gradient descent is bouncing around the global minimum but missing it owing to the larger than necessary step size.

How many binary classifiers would you need to implement one-vs-all for three classes? How does it work?

You would need three binary classifiers to implement one-vs-all for three classes since the number of binary classifiers is precisely equal to the number of classes with this approach. If you have three classes given by $y=1$, $y=2$, and $y=3$, then the three classifiers in the one-vs-all approach would consist of $h(1)(x)$, which classifies the test cases as 1 or not 1, $h(2)(x)$ which classifies the test cases as 2 or not 2 and so on. You can then take the results together to arrive at the correct classification. For example, with three categories, Cats, Dogs, and Rabbits, to implement the one-vs-all approach, we need to make the following comparisons:

1. Binary Classification Problem 1: Cats vs. Dogs, Rabbits (or not Cats)
2. Binary Classification Problem 2: Dogs vs. Cats, Rabbits (or not Dogs)
3. Binary Classification Problem 3: Rabbits vs. Cats, Dogs (or not Rabbits)

What is the importance of regularisation?

Regularisation is a technique that can help alleviate the problem of overfitting a model. It is beneficial when a large number of parameters are present, which help predict the target function. In these circumstances, it is difficult to select which features to keep manually.

Regularisation essentially involves adding coefficient terms to the cost function so that the terms are penalized and are small in magnitude. This helps, in turn, to preserve the overall trends in the data while not letting the model become too complex. These penalties, in effect, restrict the influence a predictor variable can have over the target by compressing the coefficients, thereby preventing overfitting.

Will the decision boundary be linear or non-linear in logistic regression models? Explain with an example.

The decision boundary is essentially a line or a plane that demarcates the boundary between the classes to which linear regression classifies the dependent variables. The shape of the decision boundary will depend entirely on the logistic regression model.

For logistic regression model given by hypothesis function $h(x)=g(Tx)$where g is the sigmoid function, if the hypothesis function is $h(x)=g(1+2x2+3x3)$then the decision boundary is linear. Alternatively, if $h(x)=g(1+2x22+3x32)$then the decision boundary is non-linear.

Provide a mathematical intuition for Logistic Regression?

Logistic regression can be seen as a transformation from linear regression to logistic regression using the logistic function, also known as the sigmoid function or $S(x)$:

$$S(x) = \frac{1}{1 + e^{-x}}$$

Given the linear model:

$y = b_0 + b_1 \cdot x$

If we apply the sigmoid function to the above equation it results:

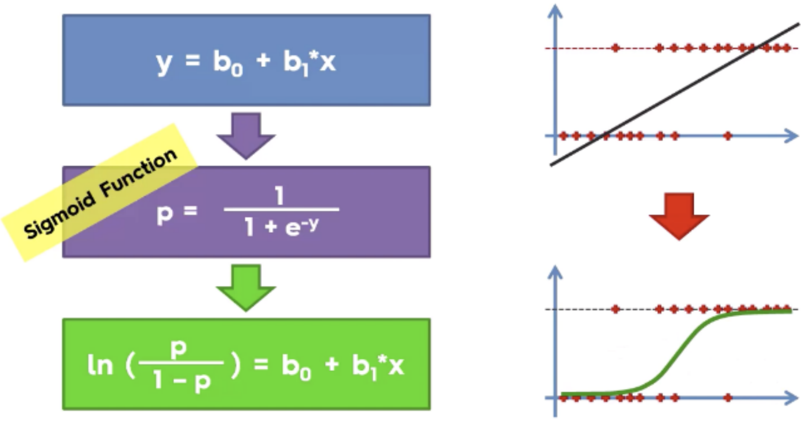
$S(y) = \frac{1}{1+e^{-y}} = p$

where `p` is the probability and it takes values between 0 and 1. If we now apply the logit function to `p`, it results:

$logit(p) = log(\frac{p}{1-p}) = b_0 + b_1 \cdot x$

The equation above represents the logistic regression. It fits a logistic curve to set of data where the dependent variable can only take the values 0 and 1.

The previous transformation can be illustrated in the following figure:



logistic function (also called the 'inverse logit').

What's the difference between Softmax and Sigmoid functions?

- **Softmax function**:
  - Is used for *multi-class* classification in logistic regression models, when we have only *one right answer* or *mutually exclusive* outputs.
  - Its probabilities sum will be `1`.
  - Is used in *different layers* of neural networks.
  - It is defined as:

$$\text{softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \forall \text{j} = 1...\text{K}$$

- **Sigmoid function**:
  - Is used for *multi-label* classification in logistic regression models, when we have *more than one right answer* or *non-exclusive* outputs.
  - Its probabilities sum does not need to be `1`.
  - Is used as an *activation function* while building neural networks.
  - It is defined as:

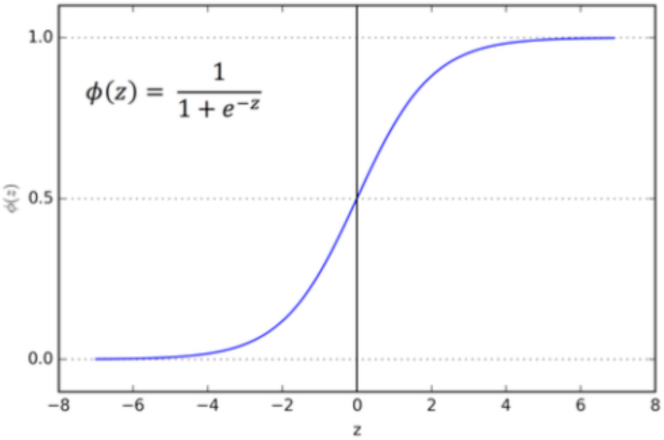
$$\sigma(z_j) = \frac{e^{z_j}}{1 + e^{z_j}}$$

Why is Logistic Regression considered a Linear Model?

A model is considered linear if the transformation of features that is used to calculate the prediction is a linear combination of the features. Although Logistic Regression uses Sigmoid function which is a nonlinear function, the model is a generalized linear model because the outcome always depends on the sum of the inputs and parameters.

i.e the logit of the estimated probability response is a linear function of the predictors parameters.

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i}$$



Why can't we use Mean Square Error (MSE) as a cost function for Logistic Regression?

In Logistic Regression, we use the sigmoid function to perform a non-linear transformation to obtain the probabilities. If we square this nonlinear transformation, then it will lead to the problem of non-convexity with local minimums and by using gradient descent in such cases, it is not possible to find the global minimum. As a result, MSE is not suitable for Logistic Regression.

So, in the Logistic Regression algorithm, we used Cross-entropy or log loss as a cost function. The property of the cost function for Logistic Regression is that:

The confident wrong predictions are penalized heavily

The confident right predictions are rewarded less

By optimizing this cost function, convergence is achieved.

Cost(hΘ(x), Y(actual)) = - log (hΘ(x)) if y=1
-log (1- hΘ(x)) if y=0

What is the use of regularisation? Explain L1 and L2 regularisations.

Regularisation is a technique that is used to tackle the problem of overfitting of the model. When a very complex model is implemented on the training data, it overfits. At times, the simple model might not be able to generalise the data and the complex model overfits. To address this problem, regularisation is used.

Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalised and are small in magnitude. This essentially helps in capturing the trends in the data and at the same time prevents overfitting by not letting the model become too complex.

L1 or LASSO regularisation: Here, the absolute values of the coefficients are added to the cost function. This can be seen in the following equation; the highlighted part corresponds to the L1 or LASSO regularisation. This regularisation technique gives sparse results, which lead to feature selection as well.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

L2 or Ridge regularisation: Here, the squares of the coefficients are added to the cost function. This can be seen in the following equation, where the highlighted part corresponds to the L2 or Ridge regularisation.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

How to choose the value of the regularisation parameter (λ)?

Selecting the regularisation parameter is a tricky business. If the value of λ is too high, it will lead to extremely small values of the regression coefficient β, which will lead to the model underfitting (high bias – low variance). On the other hand, if the value of λ is 0 (very small), the model will tend to overfit the training data (low bias – high variance).

There is no proper way to select the value of λ. What you can do is have a sub-sample of data and run the algorithm multiple times on different sets. Here, the person has to decide how much variance can be tolerated. Once the user is satisfied with the variance, that value of λ can be chosen for the full dataset.

One thing to be noted is that the value of λ selected here was optimal for that subset, not for the entire training data.