Machine Learning Prep Guide

Anjali Chauhan

2022-04-27

Contents

1 Prerequisites **5**

2 Introduction **7**

2.1 Linear Regression 7

2.2 What's the difference between Linear Regression and Logistic Regression? 7

2.3 What is overfitting? 8

2.4 What is the bias-variance tradeoff? 8

2.5 What are ridge and lasso regression and what are the differences between them? 9

2.6 What's the difference between L2 and L1 regularization? 9

2.7 What's Regularization? and what's the difference between L1 and L2 regularization? 10

2.8 Can we use L1 regularization for feature selection? 10

2.9 When do we need to perform feature normalization for linear models? When it's okay not to do it? 11

2.10 Logistic Regression 11

2.11 What is logistic regression? Or State an example when you have used logistic regression recently. 11

3 Literature **13**

4 Methods **15**

4.1 math example 15

5 Applications **17**

5.1 Example one 17

5.2 Example two 17

6 Final Words **19**

Chapter 1

Prerequisites

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Chapter 2

Introduction

2.1 Linear Regression

Linear Regression involves finding a 'line of best fit' that represents a dataset using the least squares method. The least squares method involves finding a linear equation that minimizes the sum of squared residuals. A residual is equal to the actual minus predicted value.

To give an example, the red line is a better line of best fit than the green line because it is closer to the points, and thus, the residuals are smaller.

Linear Regression is one of the most fundamental algorithms used to model relationships between a dependent variable and one or more independent variables. In simpler terms, it involves finding the 'line of best fit' that represents two or more variables.

The line of best fit is found by minimizing the squared distances between the points and the line of best fit — this is known as minimizing the sum of squared residuals. A residual is simply equal to the predicted value minus the actual value.

2.2 What's the difference between Linear Regression and Logistic Regression?

Linear Regression is used to predict a continuous variable and is mainly used to solve regression problems. Linear regression finds the best fit line by which the output numerical value can be predicted.

Logistic Regression is used to predict categorical values and is mainly used in classification problems. Logistic regression produces an S curve that classifies,

the output is binary or categories.

2.3 What is overfitting?

Overfitting is an error where the model 'fits' the data too well, resulting in a model with high variance and low bias. As a consequence, an overfit model will inaccurately predict new data points even though it has a high accuracy on the training data.

Overfitting is a modeling error when a function fits the data too closely, resulting in high levels of error when new data is introduced to the model.

There are a number of ways that you can prevent overfitting of a model:

- **Cross-validation**: Cross-validation is a technique used to assess how well a model performs on a new independent dataset. The simplest example of cross-validation is when you split your data into two groups: training data and testing data, where you use the training data to build the model and the testing data to test the model.
- **Regularization**: Overfitting occurs when models have higher degree polynomials. Thus, regularization reduces overfitting by penalizing higher degree polynomials.
- **Reduce the number of features**: You can also reduce overfitting by simply reducing the number of input features. You can do this by manually removing features, or you can use a technique, called Principal Component Analysis, which projects higher dimensional data (eg. 3 dimensions) to a smaller space (eg. 2 dimensions).
- **Ensemble Learning Techniques**: Ensemble techniques take many weak learners and converts them into a strong learner through bagging and boosting. Through bagging and boosting, these techniques tend to overfit less than their alternative counterparts.

Overfitting is an error where the model 'fits' the data too well, resulting in a model with high variance and low bias. As a consequence, an overfit model will inaccurately predict new data points even though it has a high accuracy on the training data.

2.4 What is the bias-variance tradeoff?

The bias of an estimator is the difference between the expected value and true value. A model with a high bias tends to be oversimplified and results in underfitting. Variance represents the model's sensitivity to the data and the noise. A model with high variance results in overfitting.

Therefore, the bias-variance tradeoff is a property of machine learning models in which lower variance results in higher bias and vice versa. Generally, an optimal balance of the two can be found in which error is minimized.

2.5 What are ridge and lasso regression and what are the differences between them?

Both L1 and L2 regularization are methods used to reduce the overfitting of training data. Least Squares minimizes the sum of the squared residuals, which can result in low bias but high variance.

L2 Regularization, also called ridge regression, minimizes the sum of the squared residuals plus lambda times the slope squared. This additional term is called the Ridge Regression Penalty. This increases the bias of the model, making the fit worse on the training data, but also decreases the variance.

If you take the ridge regression penalty and replace it with the absolute value of the slope, then you get Lasso regression or L1 regularization.

L2 is less robust but has a stable solution and always one solution. L1 is more robust but has an unstable solution and can possibly have multiple solutions.

2.6 What's the difference between L2 and L1 regularization?

- **Penalty terms**: L1 regularization uses the sum of the absolute values of the weights, while L2 regularization uses the sum of the weights squared.
- **Feature selection**: L1 performs feature selection by reducing the coefficients of some predictors to 0, while L2 does not. Computational efficiency: L2 has an analytical solution, while L1 does not.
- **Multicollinearity**: L2 addresses multicollinearity by constraining the coefficient norm.
- L1 effectively removes features that are unimportant, and doing this too aggressively can lead to underfitting. L2 weighs each feature instead of removing them entirely, which can lead to better accuracy. Briefly, L1 removes features while L2 doesn't, L2 regulates their weights instead.

2.7 What's Regularization? and what's the difference between L1 and L2 regularization?

Regularization in machine learning is the process of regularizing the parameters that constrain, regularizes, or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, avoiding the risk of Overfitting. Regularization basically adds the penalty as model complexity increases which can help avoid overfitting.

2.7.1 Ridge Regression

Ridge regression, also known as L2 Regularization, is a regression technique that introduces a small amount of bias to reduce overfitting. It does this by minimizing the sum of squared residuals plus a penalty, where the penalty is equal to lambda times the slope squared. Lambda refers to the severity of the penalty.

Without a penalty, the line of best fit has a steeper slope, which means that it is more sensitive to small changes in X. By introducing a penalty, the line of best fit becomes less sensitive to small changes in X. This is the idea behind ridge regression.

2.7.2 Lasso Regression

Lasso Regression, also known as L1 Regularization, is similar to Ridge regression. The only difference is that the penalty is calculated with the absolute value of the slope instead.

2.8 Can we use L1 regularization for feature selection?

Yes, because the nature of L1 regularization will lead to sparse coefficients of features. Feature selection can be done by keeping only features with non-zero coefficients.

2.9 When do we need to perform feature normalization for linear models? When it's okay not to do it?

Feature normalization is necessary for L1 and L2 regularizations. The idea of both methods is to penalize all the features relatively equally. This can't be done effectively if every feature is scaled differently.

Linear regression without regularization techniques can be used without feature normalization. Also, regularization can help to make the analytical solution more stable, — it adds the regularization matrix to the feature matrix before inverting it.

2.10 Logistic Regression

Logistic Regression is a classification technique that also finds a 'line of best fit'. However, unlike linear regression where the line of best fit is found using least squares, logistic regression finds the line (logistic curve) of best fit using maximum likelihood. This is done because the y value can only be one or zero.

Logistic regression is similar to linear regression but is used to model the probability of a discrete number of outcomes, typically two. For example, you might want to predict whether a person is alive or dead given their age.

At a glance, logistic regression sounds much more complicated than linear regression, but really only has one extra step.

First, you calculate a score using an equation similar to the equation for the line of best fit for linear regression.

The extra step is feeding the score that you previously calculated in the sigmoid function below so that you get a probability in return. This probability can then be converted to a binary output, either 1 or 0.

To find the weights of the initial equation to calculate the score, methods like gradient descent or maximum likelihood are used. Since it's beyond the scope of this article, I won't go into much more detail, but now you know how it works!

2.11 What is logistic regression? Or State an example when you have used logistic regression recently.

Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example,

if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

Chapter 3

Literature

Here is a review of existing methods.

Chapter 4

Methods

We describe our methods in this chapter.

Math can be added in body using usual syntax like this

4.1 math example

p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt{(\frac{p(1-p)}{n})} \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$

You can also use math in footnotes like this[1].

We will approximate standard error to 0.027[2]

[1] where we mention $p = \frac{a}{b}$

[2] p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt{(\frac{p(1-p)}{n})} \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.