Regression

Topics covered so far

1. Regression: Linear
    a. General Statistical Framework
    b. Maximum Likelihood and Bayesian Estimators
    c. Linear Regression
    d. Performance Assessment - Estimating parameter means and confidence intervals for prediction

1. Regression: Model Evaluation
    a. Prediction vs Modeling
    b. Assumptions behind Regression
    c. Bias-variance tradeoff
    d. Overfitting and Regularization
    e. Cross-validation
    f. Bootstrapping

Discussion questions

1. What is the difference between machine learning and the statistical point of view?
2. What is linear regression and how does it work?
3. What is multiple linear regression? Give an example where it is used.
4. How do you measure the performance of a linear regression model?

Statistics vs Machine Learning

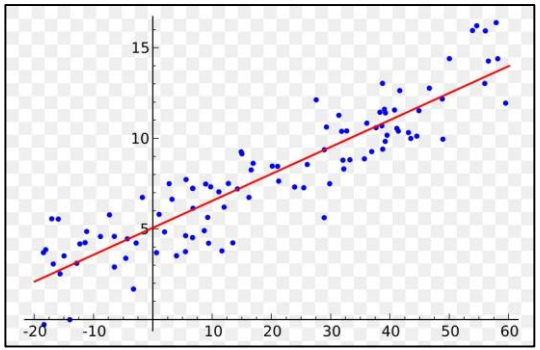
The difference between machine learning and statistics is their purpose. Machine learning models are designed to make the most accurate predictions possible. Statistical models are designed for inference about the relationships between variables.
The following table provides the major differences between statistics and the machine learning point of view:

Statistics	Machine Learning
Emphasis on deep theorems on complex models	Emphasis on the underlying algorithm
Focus on hypothesis testing and interpretability	Focus on predicting accuracy of the model
Inference on parameter estimation, errors and predictions	Inference on prediction
Deep understanding of simple models	Theory does not always explain success

Linear Regression

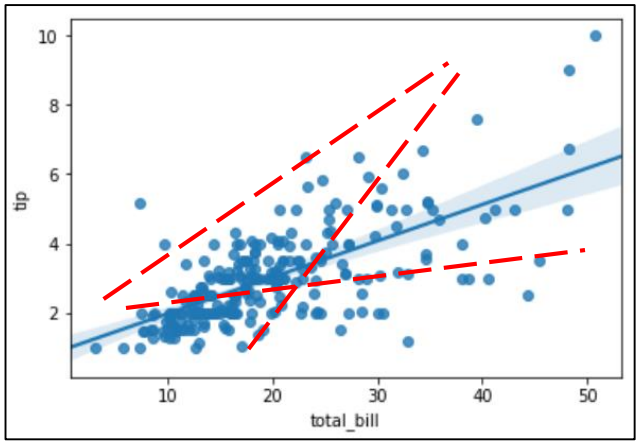
- Linear regression is a way to identify a relationship between the independent variable(s) and the dependent variable
- We can use these relationships to predict values for one variable for given value(s) of other variable(s)
- It assumes the relationship between variables can be modeled through linear equation or an equation of line.
- The variable, which is used in prediction is termed as independent/explanatory/regressor where the predicted variable is termed as dependent/target/response variable.
- In case of linear regression with a single explanatory variable, the linear combination can be expressed as :

response = intercept + constant * explanatory variable



Best fit line in the linear regression model

- Learning from the data, the model generates a line that fits the data.
- Our aim is to find a regression line that best fits the data
- By best fit, it means that the line will be such that the cumulative distance of all the points from the line is minimized
- Mathematically, the line that minimizes the sum of squared error of residuals is called Regression Line or the Best Fit Line.



In the example here, you can see a scatter plot between the *tip* amount and the *total_bill* amount

We can see that there is positive correlation between these two - as the bill amount increases, the tip increases

The line in blue that you see is the 'best fit' line - those in red are some examples of all other lines that are not the 'best fit'

What is Multiple linear regression?

- This is just the extension of the concept of simple linear regression with one variable
- In the real world, any phenomenon or outcome could be driven by many different independent variables
- Therefore there is a need to have a mathematical model that can capture this relationship

  - **Ex:** Predicting the price of a house, we need to consider various attributes such as area, number of rooms, number of kitchens etc. Such a regression problem is an example of multiple linear regression.
  - The equation for multiple linear regression can be represented by :

target = intercept + constant 1*feature 1 + constant 2*feature 2 + constant 3*feature 3 + ...

- The model aims to find the constants and intercept such that this line is the best fit

Regression Model Evaluation Metrics

R-squared	Adjusted R-squared	Mean Absolute Error	Root Mean Square Error
• Measure of the % of variance in the target variable explained by the model • Generally the first metric to look at for linear regression model performance • Higher the better	• Conceptually, very similar to R-squared but penalizes for addition of too many variables • Generally used when you have too many variables as adding more variables always increases R^2 but not Adjusted R^2 • Higher the better	• Simplest metric to check prediction accuracy • Same unit as dependent variable • Not sensitive to outliers i.e. errors doesn't increase too much if there are outliers • Difficult to optimize from mathematical point of view (pure maths logic) • Lower the better	• Another metric to measure the accuracy of prediction • Same unit as dependent variable • Sensitive to outliers - errors will be magnified due to square function • But has other mathematical advantages that will be covered later • Lower the better

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - yhat_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}$$

Discussion questions

1. What are the underlying assumptions in the linear regression model?
2. What is Bias Variance trade-off?
3. What is Regularization? What are its different types?
4. Why do we use cross-validation? How does it work?
5. What is the concept of bootstrapping and why do we need it?

Assumptions of Linear Regression

Assumption	How to test	How to fix
There should be a linear relationship between dependent and independent variables	Pairplot / Correlation of each independent variables with dependent variable	Transform variables that appear non-linear (log, square root etc.)
No multicollinearity in independent variables	Heatmaps of correlations or VIF (Variance inflation factor)	Remove correlated variables or merge them
No Heteroskedasticity - residuals should have constant variance	Plot residuals vs. fitted values and check the plot	Non-linear transformation of dependent variable or add other important variables
Residuals must be normally distributed	Plot residuals or use Q-Q plot	Non-linear transformation of independent or dependent variable

Bias-Variance: Underfitting and Overfitting

Bias: Bias is the difference between the prediction of our model and the correct value which we are trying to predict. The model with high bias gives less attention to the training data and overgeneralizes the model which leads to a high error on training and test data.

Variance: Variance is the value that tells us the spread of our data. A model with high variance pays a lot of attention to training data and does not generalize on the test data. Therefore, such models perform very well on training data but have a high error on test data.

In supervised learning, **underfitting** happens when a model is not able to capture the underlying pattern of the data. These models usually have high bias and low variance whereas, **overfitting** happens when our model captures the noise along with the underlying pattern in data. These models usually have low bias and high variance.

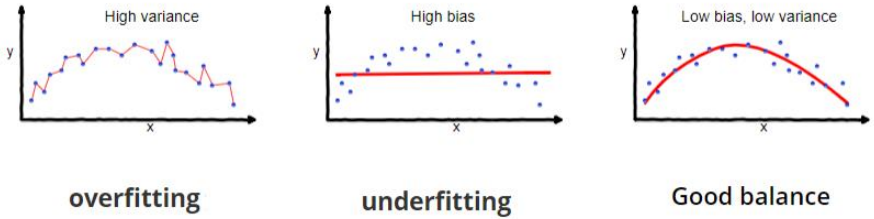


Image Source

Bias-Variance Tradeoff

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

An optimal balance of bias and variance would never overfit or underfit the model.

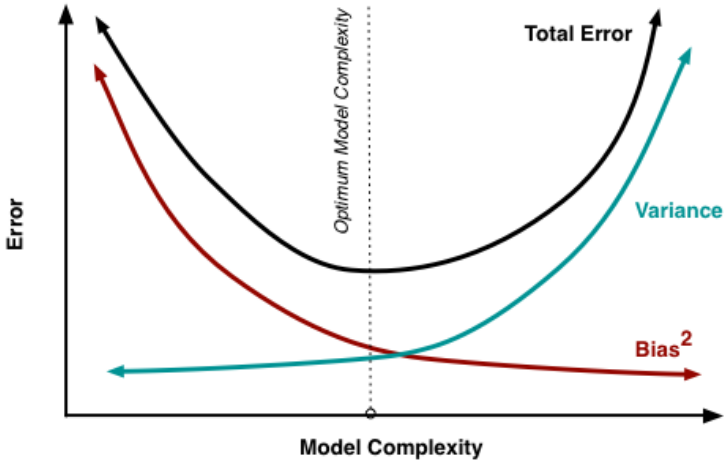


Image Source

Regularization and its types

- Regularization is the process which regularizes or shrinks the coefficients towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
- Regularization, significantly reduces the variance of the model, without substantial increase in its bias.
- There are two types of regularization:
    - **Lasso Regression:** In this technique we add $\alpha * \sum|\beta|$ as the shrinkage quantity. It only penalizes the high coefficients. It has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. This technique is also called L1 regularization.
    - **Ridge Regression:** In this technique, we modify the residual sum of squares by adding the shrinkage quantity $\alpha * \sum\beta^2$ and use α as the tuning parameter that decides how much we want to penalize the flexibility of our model. This technique is also called L2 regularization.

Cross-validation and its types

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

- It provides some kind of assurance that your model has got most of the pattern from the data set correct and it is not picking up some noise.
- We will be discussing two types of cross validation techniques -
    1. K-Fold Cross-validation
    2. Leave-One-Out Cross-validation (LOOCV)

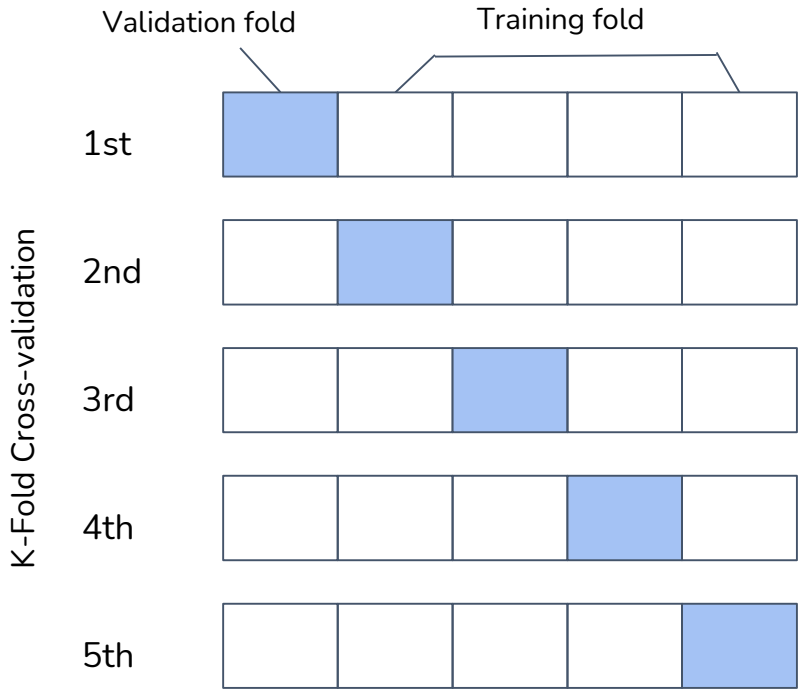
K-fold Cross-validation

This algorithm has a single parameter called K that refers to the number of groups that a given data sample is to be split into.

This algorithm has following procedure:

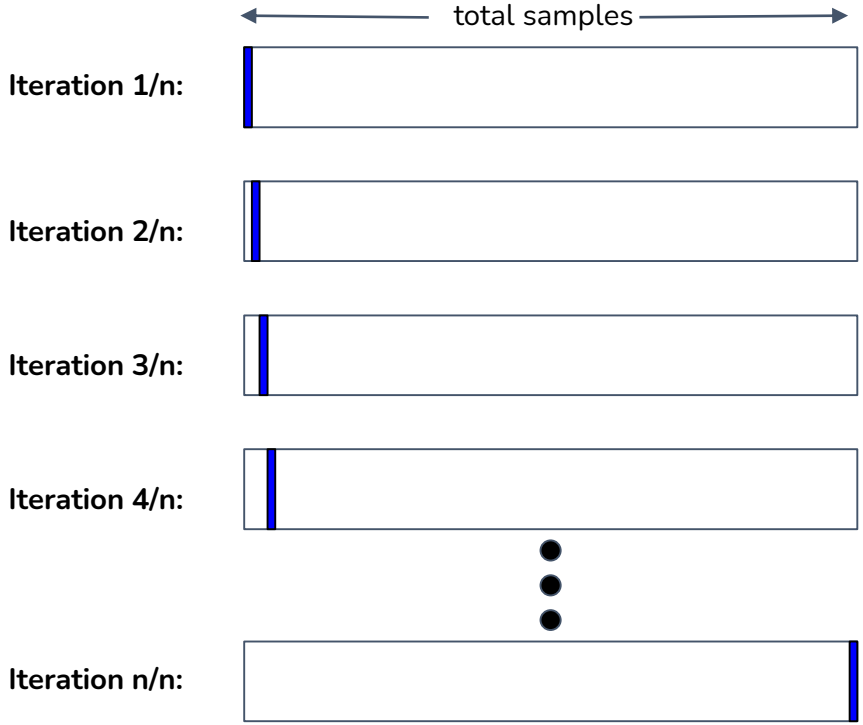
1. Shuffle the dataset randomly.
2. Split the whole dataset into K groups
3. For each unique group, take one as a hold out set and remaining as training set.
4. Repeat the step 3, for all groups
5. Summarize the skill of the model using the sample of model evaluation scores of all groups

Here, K = 5



Leave-One-Out Cross-validation (LOOCV)

- LOOCV is a special case of K-fold cross validation where K equals n, n being the number of data points in the sample.
- This approach leaves 1 data point out of the training data, i.e. if there are n data points in the original sample then, n-1 samples are used to train the model and p points are used as the validation set.
- This is repeated for all combinations in which the original sample can be separated this way, and then the error is averaged for all trials, to give overall effectiveness.
- The number of possible combinations is equal to the number of data points in the original sample or n.



Bootstrapping

Bootstrapping (also called Bootstrap sampling) is a resampling method that involves drawing of sample data repeatedly with replacement to estimate a population parameter.

It involves the following steps:

1. Choose a number of bootstrap samples to perform
2. Choose a sample size n
3. For each bootstrap sample
    1. Draw a sample with replacement with the chosen size
    2. Calculate the statistic on the sample
4. Calculate the mean of the calculated sample statistics.

Bootstrap sampling can be used to estimate the parameter of a population (i.e. mean, standard error etc.)

Case study

Great Learning

Happy Learning !