AIML - CS 419

Lecture 2: Regression and Loss Functions

18/01/2023

Lecturer: Abir De *Scribe: Groups 3 & 4*

In the previous lecture we introduced canonical machine learning problems and understood what is exactly meant by data and noise in data. Let's continue by appreciating why understanding the domain of data is important and a brief overview of loss functions.

1 Introduction

Understanding the domain of data helps us to select our model appropriately, engineer the features (i.e. remove outliers, remove noise, etc.) and identify underfitting and overfitting. We might also be able to apply appropriate data transformations using the domain knowledge and convert it into a form such that even a simpler model would suffice. Noise can have a negative impact on the performance of machine learning models by making it more challenging to find the underlying patterns in the data. **Curve Fitting** is a process to construct a mathematical function with the best fit to a given series of data points. **Loss Function** estimates how well does the algorithm perform on training, validation and test sets.

2 Loss Functions

We create loss functions because we require a measurable method to express how large the error is. We don't use loss functions like $\sum_i^n (y_i - f(x_i))$ or $\sum_i^n (y_i - f(x_i))^3$ as they do not account for positive and negative errors instead they nullify each other. By choosing the modulus or square function we penalize both positive and negative errors. We also try to keep the loss function differentiable and continuous so that it can be minimized effectively using appropriate algorithms.

2.1 Few types of Loss Functions

- **Mean Squared Error** : A popular loss function used in machine learning for regression is mean squared error (MSE). The average of the squared discrepancies between the expected and actual values is what is measured.

$$MSE = \frac{\sum_1^n (y_i - f(x_i))^2}{n} \tag{1}$$

  It's important to note that the MSE can be sensitive to outliers and it can be affected by the scale of the data. This is defined for continuous values and thus can't be used for classification problems.

- **Binary Cross Entropy Loss** : This is commonly used in classification tasks where the labels are binary. It uses the logarithm function to convey the dissimilarity between the predicted and true distribution.

$$BCELoss = -(y * log(p) + (1 - y) * log(1 - p)) \tag{2}$$

where y is the true binary label (0 or 1) and p is the probability of the positive class. Thus when the actual label is 1 and the probability that label is true (predicted using f) is also 1, the loss is 0 as expected.

A similar loss function called the Sparse Categorical Entropy Loss Function is typically used for multi-class classification problems.

$$Loss = \sum_{i}^{n}(-y * log(p)) \tag{3}$$

where y is the true integer class label, p is the predicted probability of the true class.

- **Mean absolute error** : It is defined as

$$\frac{\sum_1^n |y_i - f(x_i)|}{n} \tag{4}$$

It is very similar to MSE apart from the fact that it is less sensitive to outliers as it does not penalize the square of the difference but the only the modulus.

2.2 How to choose a Loss Function?

Choosing a loss function is very crucial and problem - specific. This example illustrates how we can design a loss function such that minimizing it will help us to meet our requirements.
Example : Let x_i be the features used to predict the target variable y_i after carrying out suitable feature engineering. Our aim is to find a function f such that for a given ϵ

$$|y_i - f(x_i)| < \epsilon \quad \forall i \tag{5}$$

It may be possible that if ϵ is too small for such a function to exist. Assuming ϵ and other hyperparameters of the function f are chosen appropriately, we can optimize the parameters of f using the following the loss function :

$$max(0, |y_i - f(x_i)| - \epsilon) \tag{6}$$

This loss function penalizes the model when $|y_i - f(x_i)| > \epsilon$ and the loss is 0 when $|y_i - f(x_i)| < \epsilon$. Thus, when we minimize the above loss function for all training examples, we indeed get the optimal parameters. But let's say we have a function *f*, then the least value of ϵ such that (1) holds

true ∀i is the maximum of the differences between the training examples and the predicted values i.e.

$$\max_i |y_i - f(x_i)|$$

Thus by minimizing over all functions the best bound of ϵ can be given by :

$$\min_f \max_i |y_i - f(x_i)| \tag{7}$$

3 Data

In machine learning (ML), data refers to the input information about the problem in quantized form that is used to train and test ML models. This data can come in many forms, such as numerical values, images, or text, and is often organized into datasets that are used to train and test ML algorithms. Quality data is fundamental to any Machine Learning project. To gain actionable insights, the appropriate data must be sourced and cleansed. There are two key stages of understanding data: a Data Assessment and Data Exploration.

3.1 Data Assessment

If we want to make predictions using machine learning we need to have a data which also includes one or more parameters that we need to predict. The data should also contain the relavent parameters that can be used to predict the desired value. Adding unrelated features to a machine learning model, makes the algorithm look for connections that aren't there. We also need to make sure that our data holds the information in relation to the value we need to predict using ML. This can result in decreased performance. And also, to successfully build the machine learning model there must be enough data points available with sufficient variance.

3.2 Data Exploration.

Data exploration in machine learning refers to the process of analyzing and visualizing the data to better understand its characteristics and properties. This includes looking for missing values, outliers, duplicate data, examining the relationship between different features in the data, such as correlation or mutual information.

An ideal data set would be complete, with valid values for every observation. However, in reality, we will come across many "NULL" or "NaN" values. We could remove the missing values entirely, but this could lead to loss of some information and may also introduce bias. An alternative to removing data is imputing values; replacing missing values with an appropriate substitute. For continuous variables, the mean, median, or mode are often used.

An outlier is a data point that is significantly different from other observations. Outliers could indicate bad data: data that was incorrectly collected. Alternatively, these values could be interesting and useful for your machine learning model. Some machine learning algorithms, such as

linear regression, can be sensitive to outliers. Hence we might need to make suitable adjustments to our model and data to deal with the outliers.

A dataset is unbalanced if each class does not have a similar number of examples. This is common with classification problems such as fraud detection; the majority of transactions are normal, whilst a small proportion are fraudulent. Our data is unbalanced, the model may not be able to identify the patterns that are associated with the minority categories.

4 Regression

- Regression is a method for understanding the relationship between independent variables or features and a dependent variable or outcome. Outcomes can then be predicted once the relationship between independent and dependent variables has been estimated. Regression is a field of study in statistics which forms a key part of forecast models in machine learning. It's used as an approach to predict continuous outcomes in predictive modelling, so has utility in forecasting and predicting outcomes from data. Machine learning regression generally involves plotting a line of best fit through the data points. The distance between each point and the line is minimised to achieve the best fit line.

- **Definition:** In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features').

- **Linear regression:** Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

- Following figure shows the regression:

- **Regression example:** Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients. For example, researchers might administer various dosages of a certain drug to patients and observe how their blood pressure responds. They might fit a simple linear regression model using dosage as the predictor variable and blood pressure as the response variable.

5 Polynomial Fitting

The regression procedure known as "Polynomial Regression" describes the connection between a dependent variable (y) and an independent variable (x) as an nth degree polynomial. If we apply

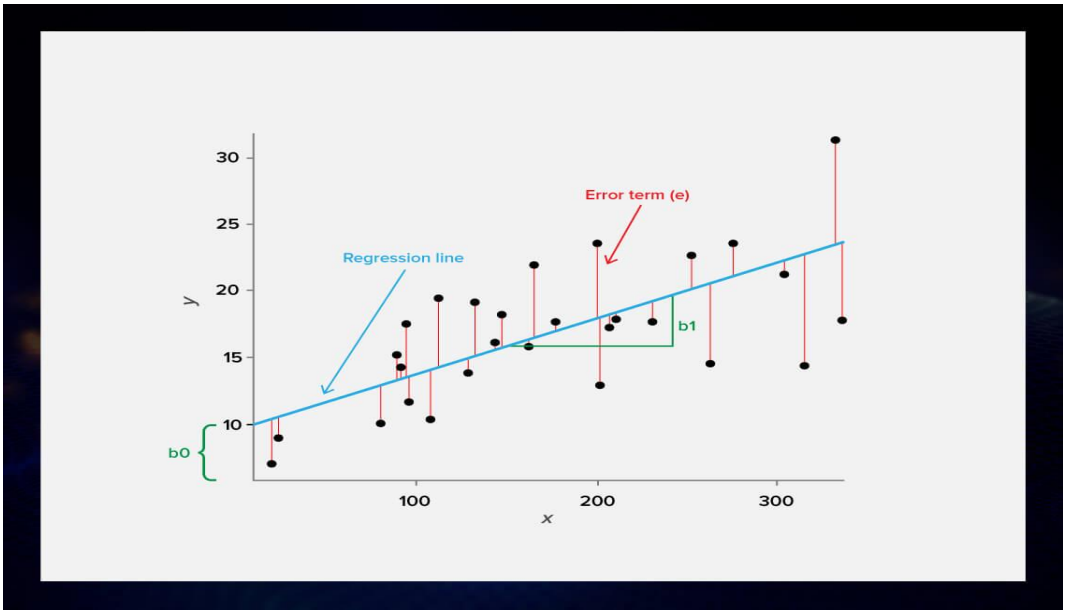

Figure 1: Figure showing linear regression.

a linear model to a linear data set, it gives us a nice result. However, if we apply the same model, unchanged, to a non-linear data set, the results would be drastically different. The result will be an increase in loss function, a high mistake rate, and a decline in accuracy.

- Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points, possibly subject to constraints.
- Thus we need a criteria to compare two curves on a data-set.
- To do so, we describe an error function E(f, D) which takes a curve f and data-set D as input and returns a real number.
- The error function must be such that it can capture how much worse the curve is at approximating the data-set.

We try to fit the data using a polynomial function of the form

$$y(x, w) = w_o + w_1 x + w_2 x^2 + ... + w_m x^m = \Sigma_{j=0}^{M} w_j x^j \tag{8}$$

The values of the coefficients will be determined by fitting the polynomial to the training data. This can be accomplished by minimising an error function that assesses the mismatch between the training set data points and the function y(x,w), for every given value of w.
Error Function: the sum of the squares of the errors between the predictions for each data point x_n.

$$E(w) = \frac{1}{2}(\Sigma_{n=1}^{N}(y(x_n, w) - t_n)^2) \tag{9}$$

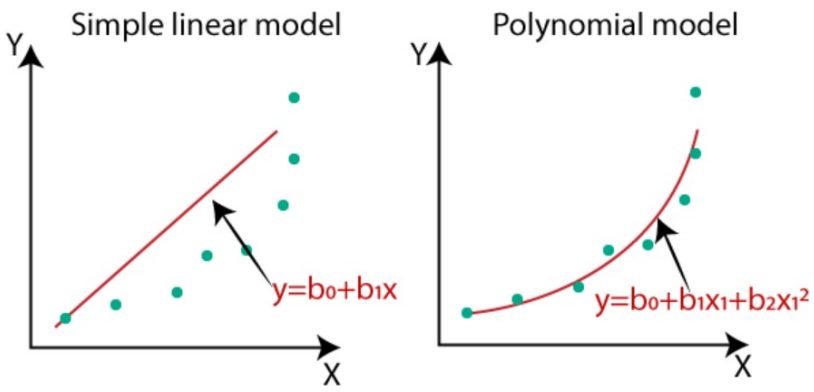

Figure 2: linear vs polynomial model

The minimization of the error function has a unique solution, represented by w^*, since the error function is a quadratic function of the coefficients, its derivatives with respect to the coefficients will be linear in the components of w.