Assignment 3

Score: 55 / 56

What is the fundamental difference between linear regression and logistic regression?

The goal of linear regression is to simulate the link between characteristics, or independent variables, and outcomes, or continuous dependent variables. It makes a numerical value prediction and assumes a linear connection.

For binary classification issues, on the other hand, logistic regression is employed. To confine predicted values between 0 and 1, the logistic function, also known as the sigmoid function, is generally used to describe the chance that a given input belongs to a specific class.

Feedback

Great answer! The explanation clearly distinguishes between linear and logistic regression with appropriate detail.

Explain the concept of support vectors in Support Vector Machines (SVM).

The data points that are closest to the line dividing one class from another are known as support vectors. Because they specify the ideal hyperplane that maximises the margin between classes, they are essential. Only these points have an impact on where the decision border is.

Feedback

Great answer! Well explained with a good understanding of support vectors in SVM.

How is the decision boundary determined in logistic regression?

Based on the anticipated probabilities, the decision border in logistic regression divides several classes. Generally, an input is categorised into one class if the projected probability for that input is greater than a threshold (usually 0.5), and into the other class if it is not. This threshold determines the decision limit.

Feedback

Great answer! Well explained with the concept of threshold in logistic regression.

What are the key assumptions of linear regression?

Linear relationship: The assumption is that the independent and dependent variables have a linear relationship.

Relative independence: implies that mistakes, or residuals, are unrelated to one another.

Constant variance of residuals: This is based on the assumption of homoscedasticity, which states that the residual variance is constant at all levels of the independent variables.

Normal distribution of residuals: Assumes that the residuals have a normal distribution.

Feedback

The response correctly identifies some key assumptions of linear regression but lacks depth and detail. It needs to cover all the essential assumptions and provide more explanation for each.

Explain the difference between SVC and SVR in Support Vector Machines (SVM).

When doing classification tasks, SVC (Support Vector Classifier) is employed to use a hyperplane to divide data points into distinct classes.

In regression problems, the SVR (Support Vector Regressor) is used to construct a hyperplane that maximises the margin while optimally fitting the data points to predict a continuous numerical value.

Feedback

Great explanation distinguishing between SVC and SVR in SVM!

What are the advantages and disadvantages of using decision trees for classification?

Benefits:

requires minimum data preparation (such as normalisation), handles both numerical and categorical data, is simple to read and visualise, and implicitly selects features.

Cons:

can be sensitive to slight changes in the data;
can be prone to overfitting, especially with deep trees and noisy data;

may not adequately capture complicated relationships in the absence of ensemble techniques like random forests.

Feedback

Great answer! You provided a comprehensive list of benefits and drawbacks of using decision trees for classification.

Discuss the concept of regularization in the context of linear regression.

Regularisation lowers the complexity of the model and adds a penalty term to the loss function to avoid overfitting.

To promote sparsity and feature selection, L1 regularisation (Lasso) adds the absolute values of the coefficients as the penalty.

By penalising big coefficients and lessening their influence on the model, L2 regularisation (Ridge) adds the squared values of the coefficients as the penalty.

Feedback

Great answer! You provided a clear explanation of regularization in linear regression, covering both L1 (Lasso) and L2 (Ridge) regularization techniques. Well done!

How does the kernel trick work in SVM? Provide an example.

SVMs may intuitively translate data into a higher-dimensional space without explicitly computing the modified feature vectors, thanks to the kernel technique.

SVMs are effective for nonlinear classification tasks because they can locate nonlinear decision boundaries in the original feature space by using kernel functions such as polynomials, radial basis functions (RBF), or sigmoids.

Feedback

Great answer! You've demonstrated a clear understanding of how the kernel trick works in SVM by explaining how SVMs handle nonlinear classification tasks using kernel functions effectively.

Explain the term 'entropy' and its role in decision tree algorithms.

A dataset's impurity, or randomness, is measured by entropy. Entropy is used in decision tree techniques such as ID3 and C4.5 to measure the degree of uncertainty in a dataset both before and during its division based on distinct properties.

The objective is to identify the characteristic that, when combined with information gain, maximises entropy reduction and produces the most informative splits.

Feedback

Great Answer! Well explained and covers all aspects of entropy in decision tree algorithms.

What is overfitting in the context of support vector machines? How can it be prevented?

Preventing overfitting in support vector machines

When an SVM model learns the training set too well, it becomes overfit and starts to identify noise and outliers rather than the underlying patterns.

It can be avoided by:

making appropriate parameter adjustments, such as changing the regularisation parameter.
using methods like cross-validation to assess model performance.
lowering the number of support vectors or utilising smaller kernels to simplify the model's complexity.

Feedback

Great Answer! Keep it up! The response demonstrates a clear understanding of overfitting in SVMs and provides effective prevention strategies.