Recap

15 May 2023 07:05

feature selection

- filter
  - var threshold
  - chisquare
  - ANOVA
  - mutual info
- Wrapper
  - → exhustive
  - → Backward elimination
  - → forward selection
  - → Rfe
- $\boxed{\text{Embedded}}$

cheatsheet

Embedded Methods

15 May 2023 07:05

Embedded methods are feature selection techniques which perform feature selection as part of the model construction process. They are called embedded methods because feature selection is embedded within the construction of the machine learning model. These methods aim to solve the limitations of filter and wrapper methods by including the interactions of the features while also being more computationally efficient.

f1 f2

filter → feature interaction

Wrapper → slow

embedded → feature sel → model building →

model → data → feature importance

predict

feature impo

machine learning

embedded

coef_

- linear reg
- logistic
- Ridge
- LASSO
- ElasticNET

linear model

Regularization

feature_importance_

- decision tree
- random forest
- gradient boosting

non-linear

attribute

model._

Linear Regression

15 May 2023 07:05

cgpa | iq | lpa

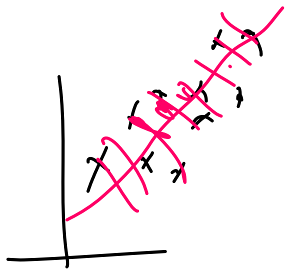
1

$$lpa = \beta_0 + \boxed{\beta_1}\, cgpa + \boxed{\beta_2}\, iq$$

importance Impo

coef_ → feature importance

$\boxed{lr.coef\_}$ → feature



1. Linearity: The relationship between the independent and dependent variables is linear. This also means the change in the dependent variable for a unit change in the independent variable(s) is constant.

2. Independence: The observations are independent of each other. This implies that the residuals (the differences between the observed and predicted values) are independent.

3. Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.

4. Normality: The residuals are normally distributed.

5. No Multicollinearity: The independent variables are not highly correlated with each other. This assumption is really important when you want to interpret the regression coefficients.

Regularized Models

15 May 2023 07:55

Regularized linear models are linear models that include a penalty term in the loss function during training. The penalty term discourages the learning of a too complex model, which can help prevent overfitting.

Ridge

LASSO

ElasticNET

feature selection

$\beta_1 \; \beta_2 \rightarrow 0$

fit_tra

standar → $\dfrac{X_i - \bar{X}}{\sigma}$

age

72

61

scaler.fit(age) → trans

$\bar{X}$ σ

scaler.fit_trans

fit fit_transform

train

transform → train

→ test

Tree Based Models

15 May 2023 07:56

└→ tree → decision tree → 8 cols

embedded →

└→ $\boxed{\text{fimport}}$ → mean

age → 0.01 → < 0.1

bmi → → 0.5 > 0.1 ✓

Recursive Feature Elimination

15 May 2023 07:56

Wrapper → hybrid

{ embedded wrapper }

f1 f2 f3 f4 y

model → coef / featureimportance

best Rfe

f1 f2 f3 y → model → feature

x

f1 f3 y → model → feature

x

f3 → best feature

Advantages and Disadvantages

15 May 2023 14:45

Advantages:

1. Performance: They are generally more accurate than filter methods since they take the interactions between features into account.

2. Efficiency: They are more computationally efficient than wrapper methods since they fit the model only once.

3. Less Prone to Overfitting: They introduce some form of regularization, which helps to avoid overfitting. For example, Lasso and Ridge regression add a penalty to the loss function, shrinking some coefficients to zero.

→ Lasso
→ ridge

Disadvantages:

1. Model Specific: Since they are tied to a specific machine learning model, the selected features are not necessarily optimal for other models.

2. Complexity: They can be more complex and harder to interpret than filter methods. For example, understanding why Lasso shrinks some coefficients to zero and not others can be non-trivial.

3. Tuning Required: They often have hyperparameters that need to be tuned, like the regularization strength in Lasso and Ridge regression. 0.01 0.1

4. Stability: Depending on the model and the data, small changes in the data can result in different sets of selected features. This is especially true for models that can fit complex decision boundaries, like decision trees.

Cheatsheet

15 May 2023 17:38

1. Filter Methods:

- Variance Threshold: Removes all features whose variance doesn't meet a certain threshold. Use this when you have many features and you want to remove those that are constants or near constants.

- Correlation Coefficient: Finds the correlation between each pair of features. Highly correlated features can be removed since they contain similar information. Use this when you suspect that some features are highly correlated.

- Chi-Square Test: This statistical test is used to determine if there's a significant association between two variables. It's commonly used for categorical variables. Use this when you have categorical features and you want to find their dependency with the target variable.

- Mutual Information: Measures the dependency between two variables. It's a more general form of the correlation coefficient and can capture non-linear dependencies. Use this when you want to measure both linear and non-linear dependencies between features and the target variable.

- ANOVA (Analysis of Variance): ANOVA is a statistical test that stands for "Analysis of Variance". ANOVA tests the impact of one or more factors by comparing the means of different samples. Use this when you have one or more categorical independent variables and a continuous dependent variable.

2. Wrapper Methods:

- Recursive Feature Elimination (RFE): Recursively removes features, builds a model using the remaining attributes, and calculates model accuracy. It uses model accuracy to identify which attributes contribute the most. Use this when you want to leverage the model to identify the best features.

- Sequential Feature Selection (SFS): Adds or removes one feature at the time based on the classifier performance until a feature subset of the desired size k is reached. Use this when computational cost is not an issue and you want to find the optimal feature subset.

- Exhaustive Feature Selection: This is a brute-force evaluation of each feature subset. This method, as the name suggests, tries out all possible combinations of variables and returns the best subset. Use this when the number of features is small, as it can be computationally expensive.

3. Embedded Methods:

- Lasso Regression: Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization. Use this when you want to create a simple and interpretable model.

- Ridge Regression: Ridge regression is a method used to analyze multiple regression data that suffer from multicollinearity. Unlike Lasso, it doesn't lead to feature selection but rather minimizes the complexity of the model.

- Elastic Net: This method is a combination of Lasso and Ridge. It incorporates penalties from both methods and is particularly useful when there are multiple correlated features.

- Random Forest Importance: Random forests provide a straightforward method for feature selection, namely mean decrease impurity (MDI). Use this when you want to leverage the power of random forests for feature selection.