Cross-Validation, Regularization

Rachel Marcone & Mauro Delorenzi

September 2023

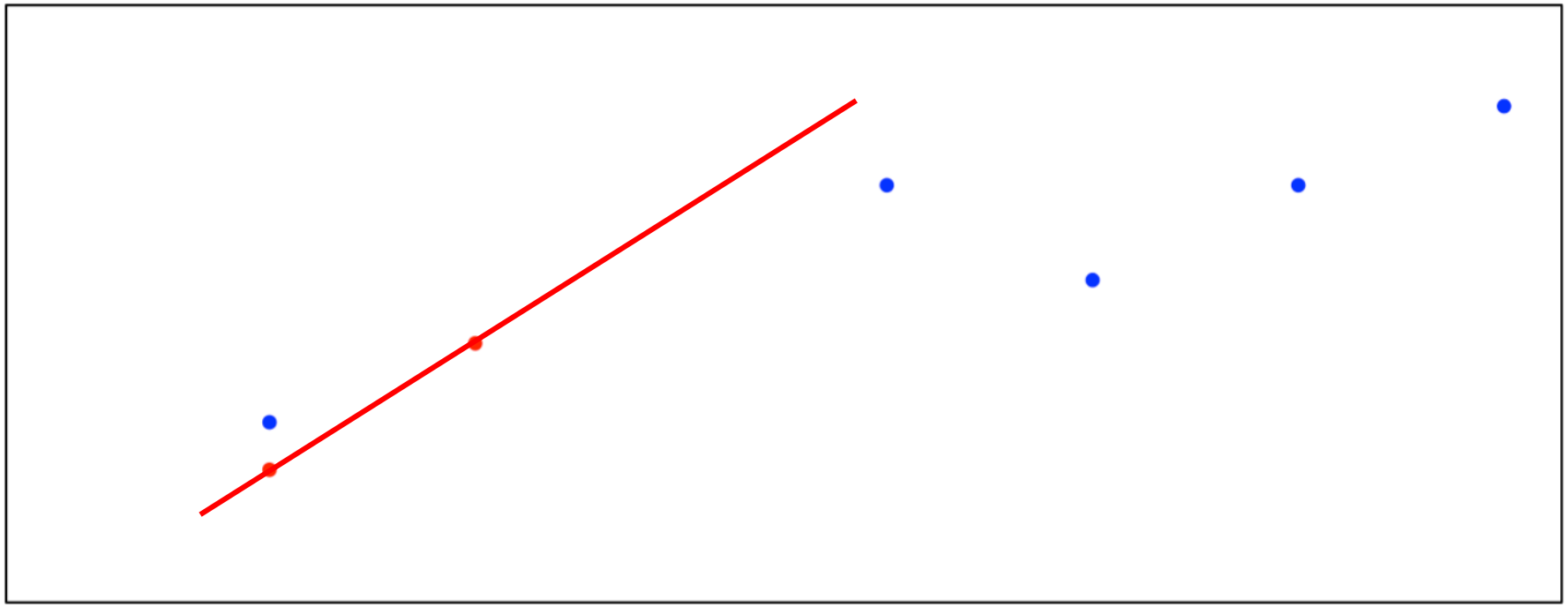
Check out Josh Starmer's channel

- https://www.youtube.com/watch?v=Q81RR3yKn30 (Ridge Regression)
- https://www.youtube.com/watch?v=NGf0voTMlcs (Lasso Regression)
- https://www.youtube.com/watch?v=1dKRdX9bfIo&t=208s (Elastic net Regression)
- https://www.youtube.com/watch?v=fSytzGwwBVw (cross validation)

- https://www.youtube.com/watch?v=iJE2fZcNPlA&t=149s (lecture comparing the 3)

Regularization

- The goal is to better predict knowing you might not have a good look at the full picture (enough points to predict the real situation).



Regularization

- Regularization are regression that optimize function with a penalty on the coefficients that are used for prediction of the outcome.
- This reduces the variance in the data and reduce overfitting and getting a better balance for the parameters.
- Some regularization will "remove" coefficients that are unnecessary for the regression
- Regularization methods seek to both minimize the sum of squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model

3 popular methods

- Lasso Regression (least absolute shrinkage and selection operator): Ordinary least squares is modified to also minimize the absolute sum of the coefficient (called also L1-regularization)
- Ridge Regression: Ordinary least squares is modified to also minimize the squared sum of the coefficient (called also L2-regularization)
- Elastic Net Regression : A combination of both of the above.

- Useful when : you have many parameters that you want to use to predict the model, when you have colinearity, or when you want to be more sure of the prediction you do.

Formulas

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$

Cross validation

- … or how to have a number on how accurate the model is.
- … or how to know how the model varies according to the points that are included
- … or how to select a model
- … or how to exclude variables
- … or how to show overfitting
- … or how to assess how good the model is

Cross validation explained

- Separates the dataset into a training and a test set
- Train the model (i.e. find the coefficients) on the training set.
- Test the model (i.e. find the outcome) for the test set and see how accurate the model is

K-Fold cross validation