Department of Computer Science

CS57300: Data Mining

Linear Regression

Prof. Chris Clifton

10 February 2022

Indiana Center for Database Systems ™

Department of Computer Science

Regression

- Task specification: Prediction
  - *But predicting a continuous value*
- Data representation: Homogeneous IID data
  - *"Class" value (response / target / dependent variable) is continuous*
- Knowledge representation
- Learning technique
- Prediction technique

2

Linear Regression

- Task specification: Regression
- Data representation: Homogeneous IID data
- Knowledge representation: Regression coefficients
- Learning technique: Matrix inversion
- Prediction technique: Linear function evaluation

Idea: Linear Model

- Task: Learn a predictive function *f(**x**;θ)=y*
  - Look familiar?
- Knowledge Representation: $\hat{y} = a_0 + \sum_{j=1}^{p} a_j x_j$
  - $\hat{y}$ is the prediction
- What assumption are we making?
  - *What does this remind you of?*

PURDUE UNIVERSITY
Department of Computer Science

Score Function

- How do we evaluate?
  - Error (residual): Difference between prediction and true value
- $y(i) = \hat{y}(i) + e(i)$

$$= a_0 + \sum_{j=1}^{p} a_j x_j(i) + e(i), 1 \le i \le n$$

- Matrix Representation: $y = \mathbf{X}\mathbf{a} + e$
  - *Note that a_0 is incorporated in the vector / Matrix*

5

PURDUE UNIVERSITY
Department of Computer Science

Score Function (2)

- Common technique for scoring: *Mean Squared Error*
  - Sum of squares of e_i

$$\frac{1}{n}\sum_{i=1}^{n} e(i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y(i) - (a_0 + \sum_{j=1}^{p} a_j x_j(i))\right)^2$$

- Learning problem: determine $\mathbf{a} = a_j$
- Work out the math, for MSE: $\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
  - *Note: This essentially works out to MLE*

6

PURDUE UNIVERSITY
Department of Computer Science

Linear Regression

Advantages

- *Optimal* solution given assumptions
- Closed form, easy to calculate
- Often works well in practice

Disadvantages

- Assumption that model is linear may not hold
- Not always easy to calculate
  - $\mathbf{X}^T\mathbf{X}$ must be invertible
  - Even when invertible, may be *unstable*

7

PURDUE UNIVERSITY
Department of Computer Science

Solving linear regression

- Direct matrix inversion
- LU decomposition: $\begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$
- Singular-Value Decomposition: $\mathbf{X} = \mathbf{USV}^T$
  - $\boldsymbol{S}$ is a diagonal matrix

8

PURDUE UNIVERSITY
Department of Computer Science

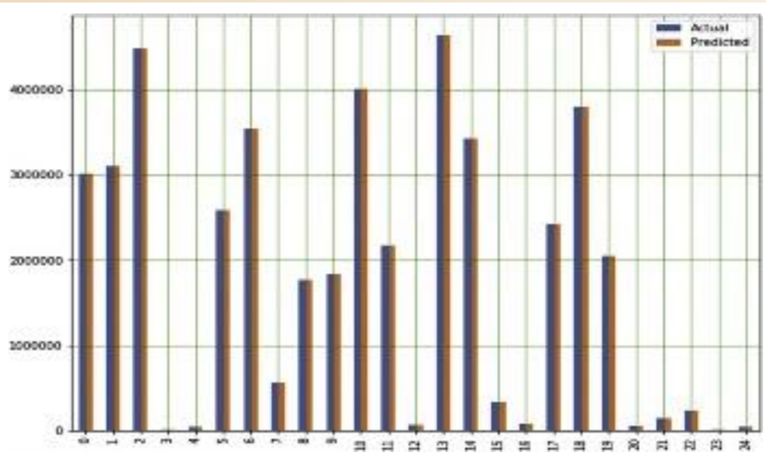
Evaluating Models

- Standard training/test, cross-validation approaches
  - What is MSE on holdout?
- How hard is the problem?
  - $\hat{y}' = \sum_{i=1}^{n} y(i)$
  - MSE difference between $\hat{y}$ and $\hat{y}'$

9

PURDUE UNIVERSITY
Department of Computer Science

Linear Regression: COVID-19 Prediction



Diabetes Metab Syndr. 2020 September-October; 14(5): 1467–1474.
Published online 2020 Aug 1. doi: 10.1016/j.dsx.2020.07.045

10

PURDUE UNIVERSITY
Department of Computer Science

Beyond Linear Regression

- CART – Classification and Regression Tree
  - Decision tree where each node is a regression model
- Artificial Neural Network (ANN) based approaches
  - Represent (arbitrarily) complex functions

11

PURDUE UNIVERSITY
Department of Computer Science

Explainability

- Linear Regression has good explainability properties
  - Low coefficient for a_i ➔ y does not have a linear relationship with x_i
  - High coefficient suggests strong relationship
- But only *linear* relationships
  - $y=x_j^2$ will end up with a low *linear* coefficient

12