Unit 3 REGRESSION AND GENERALIZATION

Regression: Assessing performance of Regression – Error measures, Overfitting and Underfitting, Catalysts for Overfitting, VC Dimensions

Linear Models: Least Square method, Univariate Regression, Multivariate Linear Regression, Regularized Regression - Ridge Regression and Lasso
Introduction to Regression Analysis, Chapter 13,

• **Regression analysis** is used to:

  – Predict values of a dependent variable, \( Y \), based on its relationship with values of at least one independent variable, \( X \).

  – Explain the impact of changes in an independent variable on the dependent variable by estimating the **numerical value** of the relationship

**Dependent variable:** the variable we wish to explain

**Independent variable:** the variable used to explain the dependent variable
Simple **Linear** Regression Model

• Only **one** independent variable (thus, simple), X

• Relationship between X and Y is described by a linear function

• Changes in Y are assumed to be caused by changes in X, that is,
  - Change In X  **Causes**  Change in Y
The graph shows the relationship between weight (in kg) and height (in cm). The equation of the line is given as:

\[ y = 0.2811x + 13.9 \]

with an \( R^2 \) value of 0.4218.
Types of Relationships

Linear relationships

Curvilinear relationships
Types of Relationships

Strong relationships

Y

X

Y

X

Weak relationships

Y

X

Y

X
Types of Relationships

No relationship

X

Y

(continued)
The population regression model: This is a conceptual model, a hypothesis, or a postulation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- **Dependent Variable**
- **Population Y intercept**
- **Population Slope Coefficient**
- **Independent Variable**
- **Random Error term**
- **Linear component**
- **Random Error component**
• The model to be estimated from sample data is:

\[ Y_i = b_0 + b_1 X_i + e_i \]

• The actual estimated from the sample

Estimated (or predicted) \( Y \) value for observation \( i \)

- Where

\[ \hat{Y}_i = b_0 + b_1 X_i \]

Residual (random error from the sample)

\[ e_i = Y_i - \hat{Y}_i \]
• The individual random error terms, $e_i$, have a mean of zero, i.e., $\sum_{i=1}^{n} e_i = 0$

• Since the sum of random error is zero, we try to estimate the regression line such that the sum of squared differences are minimized, thus, the name Ordinary Least Squared Method (OLS)

• i.e.,

$$\min \sum e^2_i = \min \sum (Y_i - \hat{Y}_i)^2 =$$

$$\min \sum (Y_i - (b_0 + b_1 X_i))^2$$

• The estimated values of $b_0$ and $b_1$ by OLS are the only possible values of $b_0$ and $b_1$ that minimize the sum of the squared differences between $Y$ and $\hat{Y}$
Simple Linear Regression Model

\[ \hat{Y}_i = b_0 + b_1 X_i \]

- Observed Value of Y for \( X_i \)
- Predicted Value of Y for \( X_i \)
- Intercept = \( b_0 \)
- Slope = \( b_1 \)

\( e_i = Y_i - \hat{Y}_i \) Random Error for this \( X_i \) value

\( \hat{Y}_{i, \text{estimated}} \)
Interpretation of the slope and the intercept

\[ \beta_0 = E(Y \mid X = 0) \quad ; \quad \beta_1 = \Delta \frac{E(Y \mid X)}{\Delta (X)}; \]

- \( b_0 \) is the estimated average value of \( Y \) when the value of \( X \) is \( b_0 \)
- \( b_1 \) is the estimated change in the average value of \( Y \) as a result of a one-unit change in \( X \)
- Units of measurement of \( X \) and \( Y \) are very important for the correct interpretation of the slope and the intercept
- How Good is this prediction?
How Good is the Model’s prediction Power?

- Total variation is made up of two parts:

\[ \text{SST} = \text{SSR} + \text{SSE} \]

Total Sum of Squares  Regression Sum of Squares  Error Sum of Squares

\[ \text{SST} = \sum (Y_i - \bar{Y})^2 \quad \text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2 \quad \text{SSE} = \sum (Y_i - \hat{Y}_i)^2 \]

where:

\( \bar{Y} \) = Average value of the dependent variable
\( Y_i \) = Observed values of the dependent variable
\( \hat{Y}_i \) = Predicted value of \( Y \) for the given \( X_i \) value
• **SST = total sum of squares**
  – Measures total variation of the $y_i$ values around their mean

• **SSR = regression sum of squares (Explained)**
  – Explained portion of total variation attributed to $Y$’s relationship with $X$

• **SSE = error sum of squares (Unexplained)**
  – Variation of $Y$ values attributable to other factors than its relationship with $X$
\[ SST = \sum (Y_i - \bar{Y})^2 \]

\[ SSE = \sum (Y_i - \hat{Y}_i)^2 \]

\[ SSR = \sum (\hat{Y}_i - \bar{Y})^2 \]

\[ Y, \text{ w/o the effect of } X \]
How Good is the Model’s prediction Power?

- The coefficient of determination is the portion of the total variation in the dependent variable, Y, that is explained by variation in the independent variable, X.
- The coefficient of determination is also called r-squared and is denoted as \( r^2 \).

\[
r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}
\]

\( 0 \leq r^2 \leq 1 \)
**Standard Error of Estimate**

- The standard deviation of the variation of observations around the regression line is estimated by
  
  \[ S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{MSE} \]

  - Where SSE = error sum of squares; n = sample size

  - The concept is the same as the standard deviation (average difference) around the mean of a univariate
Comparing Standard Errors

• $S_{yx}$ is a measure of the variation of observed Y values from the regression line.

The magnitude of $S_{yx}$ should always be judged relative to the size of the Y values in the sample data:

i.e., $S_{yx} = $36.34K is moderately small relative to house prices in the $200 - $300K range (average 215K)
Assessing performance of Regression – Error measures

- Sum of Squared Error (SSE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Goodness-fit ($R^2$)
- Bias
- Variance
Bias

• The error due to bias is taken as the difference between the expected (or average) prediction of model and the correct value which we are trying to predict.

• **Low Bias**: It means **fewer assumptions** about the form of the target function.

• **High-Bias**: It means **more assumptions** about the form of the target function.

• Examples of **low-bias machine learning algorithms** are: Decision Trees, Support Vector Machines and k-Nearest Neighbours.

• Examples of **high-bias machine learning algorithms** include: Linear Regression, Logistic Regression.
Variance

• Difference between what the model has learned from particular dataset and what the model was expected to learn

• **Low Variance**: It means small changes to the estimate of the target function with changes to the training dataset.

• **High Variance**: It means large changes to the estimate of the target function with changes to the training dataset.

• Examples of **low-variance machine learning algorithms** include: Linear Regression, Logistic Regression.

• Examples of **high-variance machine learning algorithms** include: Decision Trees, Support Vector Machines and k-Nearest Neighbours.
Overfitting and Underfitting

Ein=In sample error:
  Error in Training Data

Eout=Out of sample error or Generalization error or Validation error
  Error in Testing Data
Overfitting

• “Fitting the data more than necessary”
  It makes - Model more complex
  It makes - Model uses additional degrees of freedom
• Validation error (Eout is high, training error Ein low)
• low bias but high variance
• Some methods to overcome overfitting,
  1. Cross Validation
  2. Reduce the model complexity
  3. Regularization
Underfitting

• “underfitting occurs when the model or the algorithm does not fit the data well enough”

  It makes - Model more simple

• Validation and training error high

• low variance but high bias
High bias (underfit)  

\[ \theta_0 + \theta_1 x \]

“Just right”  

\[ \theta_0 + \theta_1 x + \theta_2 x^2 \]

High variance (overfit)  

\[ \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 \]
Comment on UF, OF, BF

• Underfitting – Validation and training error high
• Overfitting – Validation error is high, training error low
• Good fit – Validation error low, slightly higher than the training error
The VC dimension is an estimate for the capability of a binary classifier. It is a formal measure of bias. The VC dimension of a representation system is defined to be the maximum number of data points that can be separated (i.e., grouped) in all possible ways. VC dimension, is related to the trade-off between bias and variance.
VC Dimensions

- Dataset of N points
- Labelled in $2^N$ ways as positive or negative
- $2^N$ learning problems are defined.
- A hypothesis $h \in H$ that separates the positive example from negative example
- $H$ shatters N points
- Aim is to learn N examples with no error
- *The* maximum number of points shattered by $H$ is called VC Dimensions.
• To prove $\text{VC}(H)=3$ you must show:

1. There is a set of cardinality 3 shattered by $H$, and

2. Any set of cardinality greater than 3 is not shattered by $H$
VC Dimension – Shattering
VC Dimension – Shattering
VC Dimension – Shattering
VC Dimension – Shattering
VC Dimension – Shattering
VC Dimension – Shattering
VC Dimension – Shattering
VC Dimension – Shattering
• consider hyperplanes (i.e. lines in 2D).
• It is easy to find a set of three points that can be classified correctly no matter how they are labeled
• VC(H)=3
• For all $2^3=8$, $2^3=8$ possible labeling a hyperplane is available that separates them perfectly.
VC(H) = 4
• VC dimension is pessimistic: in general we do not need to worry about all possible labeling.

• It is important to remember that one can choose the arrangement of points in the space, but then the hypothesis must be consistent with all possible labeling of those fixed points.
Theory of Generalization

The ability to categorize correctly new examples that differ from those used for training is known as generalization

1. Bias and Variance Dilemma
2. Training and Testing Curves
3. Case Study of Polynomial Curve Fitting
Bias and Variance Dilemma

• The goal is to achieve low bias and low variance. -> good prediction performance.
• Parametric or linear machine learning -> high bias but a low variance.
• Non-parametric or non-linear machine learning algorithms -> low bias but a high variance.
• no escaping the relationship between bias and variance Increasing the bias will decrease the variance.
• Increasing the variance will decrease the bias.
• the algorithms you choose and the way you choose to configure them are finding different balances in this trade-off for your problem
• In reality, we cannot calculate the real bias and variance error terms because we do not know the actual underlying target function.
Training and Testing Curves

- Training error increases for a particular size of training data set. Also it remains stable for large sizes.
- But testing error will be maximum for small set size which will remains stable after some testing.
Case Study of Polynomial Curve Fitting

• The polynomial models can be used in those situations where the relationship between study and explanatory variables is curvilinear.

\[ y(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \]
• The constant \((M = 0)\) and first order \((M = 1)\) polynomials give rather poor fits to the data and consequently rather poor representations of the function \(\sin(2\pi x)\).

• The third order \((M = 3)\) polynomial seems to give the best fit to the function \(\sin(2\pi x)\)

• For higher order polynomial \((M = 9)\), an excellent fitting to the training data is achieved. In fact, the polynomial passes exactly through each data point and \(E(w_\_)) = 0\).

• However, the fitted curve oscillates wildly and gives a very poor representation of the function \(\sin(2\pi x)\). This behavior is known as overfitting