Course : COMP6577 – Machine Learning

Effective Period : February 2020

Feature Engineering: Feature Extraction & Selection

Session 15 & 16

People
Innovation
Excellence

Learning Outcome

- LO3: Student be able to experiment classification and clustering algorithm from given dataset

Outline

- Correlation and Covariance between Variables
- Normalization of Data: Standardization & Simple Range Scaling
- Statistical Tools for Variable Selection: Partial Correlation, Multiple Regression and Best-Subsets Regression
- Case Study

Correlation between Variables

- The correlation coefficient *r* is a measure of the strength of relationship between two variables.
- The higher the correlation coefficient, the stronger the relationship. The correlation coefficient for two variables x_1 and x_2 with mean $\bar{x}_1$ and $\bar{x}_2$ can be expressed as

$$r = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{N}(x_{2i} - \bar{x}_2)^2}}$$

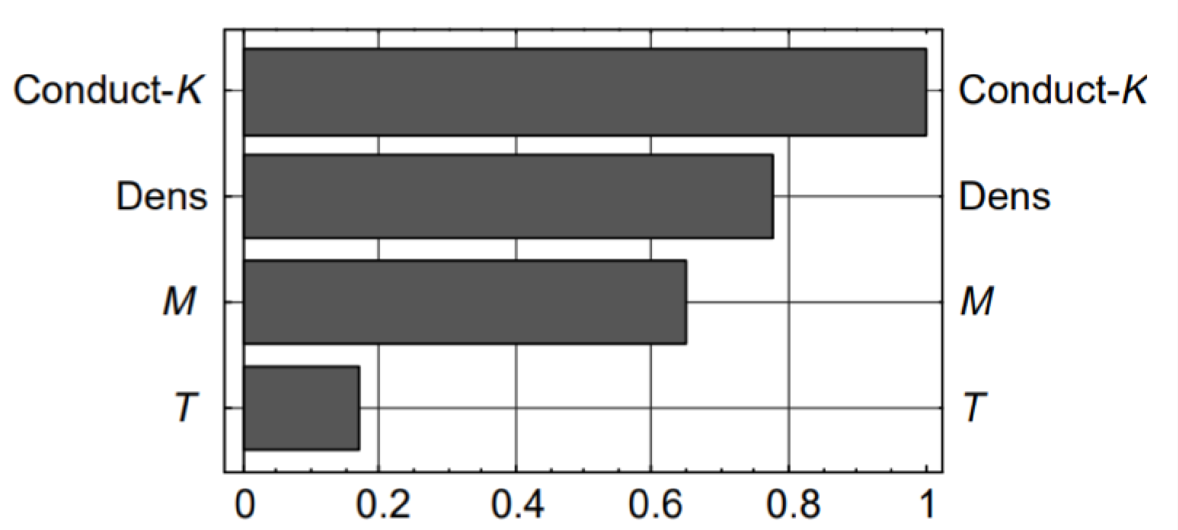
Thermal Conductivity Dataset (Example)

- Given Sample of Five Records from the Dataset for Thermal Conductivity and Related Variables (Table 1)

Species	Temp. (Celcius)	Moisture (percent)	Density (kg/m^3)	Conductivity (W/m K)
Ash white	29	15.6	647	0.1742
Red oak	29	12.4	697	0.1944
Japanese cedar	20	0	294	0.0778
Japanese beech	25	50	800	0.2132
Silver birch	100	0	680	0.25

Correlations between Variables (2)

- The linear correlation coefficients for the variables in the thermal conductivity dataset are illustrated in the bar chart in Figure below, which shows that conductivity is highly correlated with density (0.775), is reasonably highly related to moisture content (0.647), and has little correlation to temperature (0.172).



Correlations between Variables (3)

- The correlation matrix depicting correlation between all four variables is presented in this table:

	T	*M*	*Dens*	*K*
T	1.0	-0.221	-0.077	0.172
M	-0.221	1.0	0.583	0.647
Dens	-0.077	0.583	1.0	0.775
K	0.172	0.647	0.775	1.0

Table 2. Correlation Matrix for the Input and Output Variables

Correlations between Variables (4)

- The correlation matrix is symmetric with the diagonal values representing the correlation of a variable to itself, which is 1.0.
- Off-diagonal values are the correlations between pairs of variables denoted by the labels indicated in the first row and column.
- Moisture content (M) and density (Dens) are correlated at 0.583, but density has very little correlation with temperature (T) (K0.077), and moisture is weakly and negatively related to temperature (K0.221).
- Thus, judging by the correlation, conductivity is mainly influenced by density and moisture content, and these are reasonably highly correlated.
- Temperature has a weaker relationship to all the variables and is especially weak in its correlation to density

Covariance between Variables

- The covariance of two variables is expressed by

$$\mathrm{COV} = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

- Basically, the two expressions within parentheses in the Equation above, each compute the difference between the value of an input variable and its mean. When two variables coincide (i.e., $x_1 = x_2$), the result is the covariance of one variable with respect to itself, which is its variance. When $x_1 \neq x_2$, the result is the covariance between the two variables.

Covariance between Variables (2)

- The covariance matrix containing the covariance between each pair of the three input variables is presented in this table:

	T	*M*	*Dens*
T	543	-141	-351
M	-141	753	3117
Dens	-351	3117	37 888

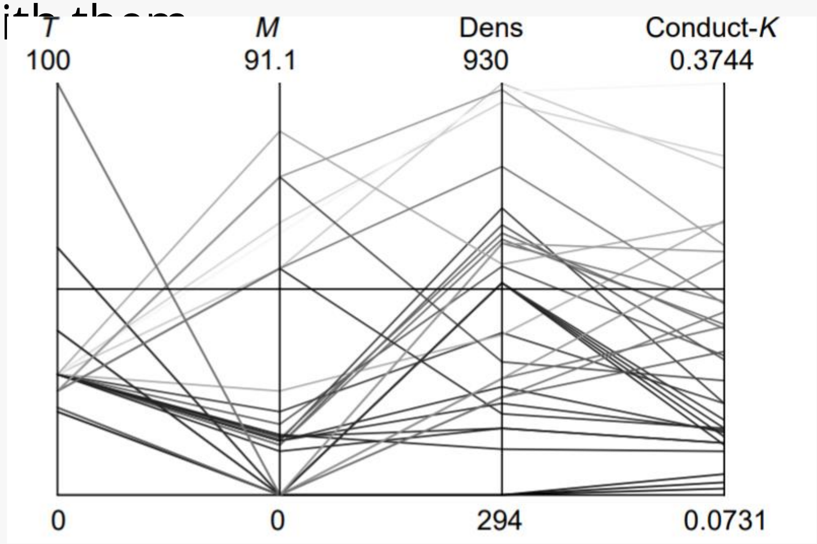
- As with the correlation matrix, the covariance matrix is symmetric.
- Here, the diagonal values represent the variance of each variable, and off-diagonal values represent the covariance between pairs of variables denoted by the labels in the first row and column

Covariance between Variables (3)

- Covariance is a measure of how two variables co-vary in relation to one another. When two variables are not related, their covariance is zero. If two variables move in the same direction, covariance is positive and if they move in opposite directions, it is negative.
- The covariance table data indicates a high positive covariance between moisture content (M) and density (Dens) and smaller negative covariance between these and temperature (T).
- Data including spread, trends, relationships, correlations, and covariances helps clarify the problem to determine appropriate strategies for normalization of data and extract relevant inputs and features for model development.

Normalization of Data

- The figure below indicates that variables in the thermal conductivity dataset have very dissimilar ranges. When variables with large magnitudes are combined with those with small magnitudes, the former can mask the effect of the latter due to the sheer magnitude of the inputs leading to larger weights associated with them.



- **Normalization** puts all inputs variables in a similar range so that true influence of variables can be ascertained.

Standardization

- There are many ways to normalize data. A simple approach is to standardize the data with respect to mean and the standard deviation using a linear transformation. This transforms all variables into a new variable with zero mean and unit standard deviation.
- To do this, each input variable is treated separately, and for each variable x_i in the training set, the mean $\bar{x}_i$ and variance σ_i^2 are calculated using:

$$\bar{x}_i = \frac{1}{N} \sum_{n=1}^{N} x_i^n$$

$$\sigma_i^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x_i^n - \bar{x}_i)^2$$

- where n = 1, ..., *N* is the pattern number.

Standardization (2)

- With the mean and the standard deviation σ_i, each input variable is normalized as:

$$x_{Ti}^{n} = \frac{x_i^n - \bar{x}_i}{\sigma_i}$$

where x_{Ti}^{n} is the normalized (transformed) value of the *n*th observation of the variable x_i.

- The new transformed variable now has zero mean and unit standard deviation. The actual range of the data depends on the original data, but most data fall within ±2σ.

People
Innovation
Excellence

Standardization (3)

- For prediction problems, the target output is also normalized using the same procedure for consistency. With the normalization, the inputs and target variables are of the same order; therefore, final weights will also be of order unity.
- This, furthermore, prevents weights from growing too large and causing training problems in situations where large weights throw the current training into a flat area of the error curve.

Standardization (4)

- Look at the thermal conductivity dataset (in previous slides). The means for the four variables—temperature, moisture content, density, and thermal conductivity—represented in vector form are

$$\bar{\mathbf{x}} = -\{34.2, 1.20, 583, 0.1811\}.$$

- Similarly, the standard deviations for the same variables represented in vector form are

$$\sigma_{\mathbf{x}} = \{23.29, 27.45, 194.6, 0.0827\}.$$

- The rescaled data x_{Ti} for those in Table Thermal Dataset using normalized eqx_{Ti}^{n}ion is shown in the next table.

Standardization (5)

- Standardized Values for Thermal Conductivity and Related Variables

Species	Temp	Moisture	Density	Conductivity
White ash	-0.222	-0.204	0.327	-0.083
Red oak	-0.222	-0.321	0.584	0.1611
Japanese cedar	-0.608	-0.772	-1.49	-1.25
Japanese beech	-0.393	1.05	1.11	0.388
Silver birch	2.82	-0.772	0.496	0.833

- Now variables are unit free and have a similar range that varies between ±3 with 0 mean and a standard deviation of 1. The correlations established earlier are not altered by this standardization.

Simple Range Scaling

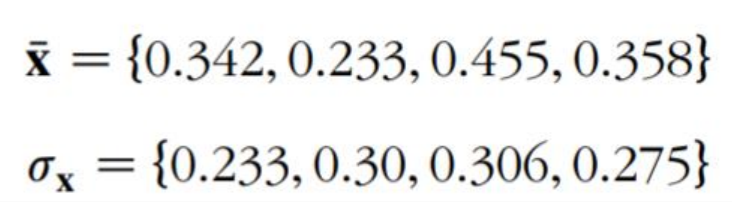
- Simple Range Scaling is a simpler approach to fix the minimum and maximum values for the normalized variables to 0 and 1 or 1 and -1, respectively. In this case, the mean and the standard deviation of the normalized inputs vary from one input variable to another, but the observations stay in the same range.
- A simple linear transformation in the range from 0 to 1 is

$$x_{Ti} = \frac{x_i - x_{i\min}}{x_{i\max} - x_{i\min}}$$

where x_{imin} and x_{imax} are the minimum and the maximum values of the variable x_i

Simple Range Scaling (2)

- A similar transformation can be made for any desired range, e.g., -1 or 1, or any other. For the example thermal conductivity problem, each of the four variables were transformed using the equation before, and the resulting mean and standard deviation for the variables are



$$\bar{\mathbf{x}} = \{0.342, 0.233, 0.455, 0.358\}$$

$$\sigma_{\mathbf{x}} = \{0.233, 0.30, 0.306, 0.275\}$$

- The above linear transformations are done for each individual variable separately without any consideration given to the correlations among data. The whole set of input variables can be considered together and linear transformations that take into account the correlations among inputs can be done.

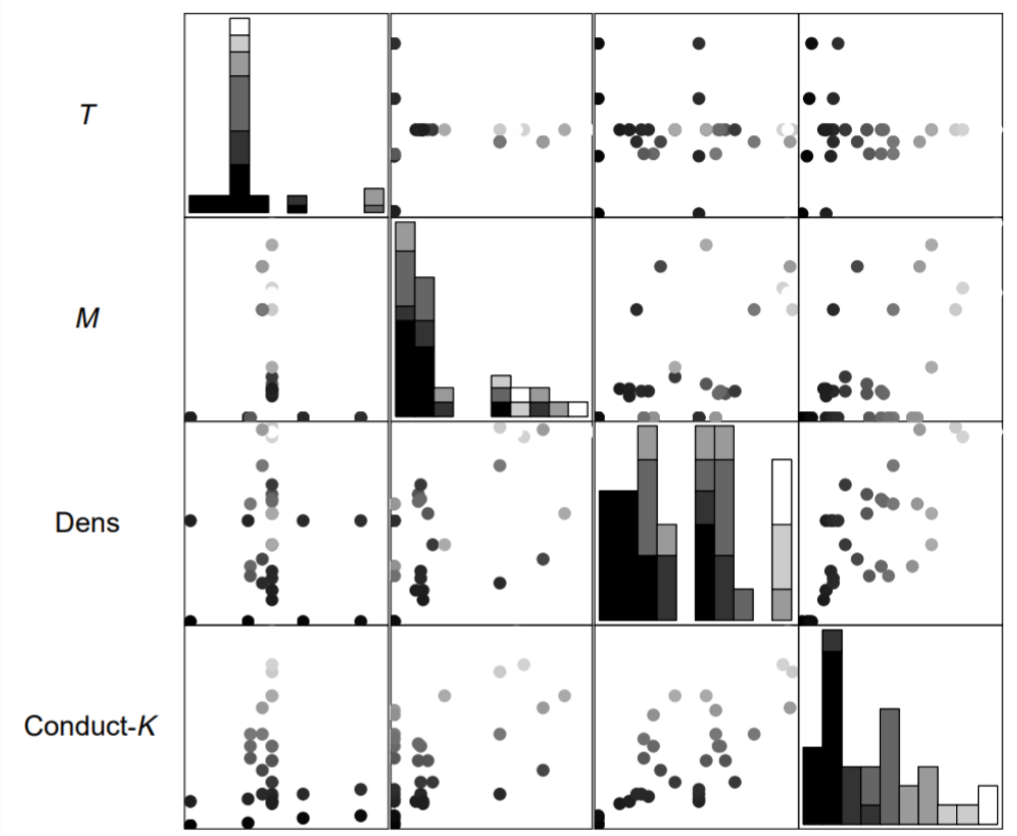
Selecting Relevant Inputs

- In linear regression, for example, too many predictor variables can adversely affect the predicted outcome. Adding a redundant variable to the least squares equation almost always increases the variance of the predicted outcome.
- Thus, too many variables can make a model very sensitive to noise or small changes in a highly correlated dataset and consequently make it less robust. Therefore, selecting a suitable subset of variables from the original set can be crucial.
- Some methods that can be used for this purpose such as Statistical Tools for Variable Selection.

People
Innovation
Excellence

Partial Correlation

- A scatter plot reveals relationships between variables in a dataset, as shown in figure below.



Partial Correlation (2)

- Points lying on a line indicate a linear relationship, a curved set of points denotes a nonlinear relationship, and absence of a pattern indicates that the two variables are uncorrelated. Linear correlation coefficients indicate the strength of the linear relationship between two variables.
- However, this technique alone is not enough for multivariate data because other variables in the set can affect the correlation of two variables, thereby altering the correlation structure.
- In such situations, **partial correlation** can be used to measure the linear association between the two variables while adjusting the effects of other variables by holding them constant.

Partial Correlation (3)

- The partial correlation is calculated from the matrix of simple correlation coefficients, an example of which is presented in the problem of thermal conductivity in relation to density, moisture content, and temperature.
- Suppose the correlation between two variables x_i and y_j is R_{ij}. The partial correlation, r_{ij}, for the two variables is given by

$$r_{ij} = \frac{-C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

where C_{ij} is the inverse of the simple correlation coefficient R_{ij} (i.e., $C_{ij} = 1/R_{ij}$).

Partial Correlation (4)

- Returning to the problem on wood thermal conductivity, the inverse of the correlation matrix in Table 2 (Correlation Matrix for the Input and Output Variables) gives the values shown in the diagonal and the top right triangle of this table utilizing symmetry. The simple linear correlation coefficients are repeated in the bottom left triangle of this table again taking advantage of the symmetry of the correlation matrix

	T	*M*	*Dens*
T	1.055	0.281	-0.082
M	-0.221	1.59	-0.906
Dens	-0.077	0.583	1.522

Inverse of the Correlation Coefficients C_{ij} (Diagonal and Top Right Triangle) and Simple Linear Correlation Coefficients R_{ij} (Bottom Left Triangle)

Partial Correlation (5)

- From the table before, partial correlation can be calculated.

	T	***M***	***Dens***
T	-1.0	-0.216	0.064
M	-0.221	-1.0	0.582
Dens	-0.077	0.583	-1.0

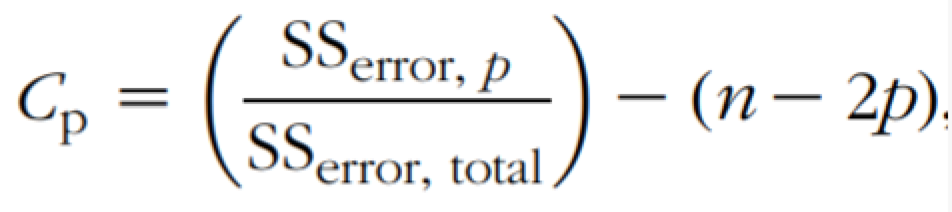
- The partial correlation matrix in the table has a similar structure to the original correlation matrix, which indicates that in this three-variable case, the simple correlations are not influenced significantly by the other variables.
- The reason for this is that only moisture and density are significantly related and temperature is weakly related to both variables.
- . For datasets consisting of many variables, the influence of other variables on the correlation between two variables can be significant.

Multiple Regression and Best-Subsets Regression

- Another approach to input selection is multiple regression analysis where a model that linearly fits the output to the input variables is developed through least squares regression.
- The R^2 or the multiple coefficient of determination represents the portion of the variability of the output explained by the predictor variables.
- A value of R^2 near 1 indicates a perfect model, and the variables capture all the variance of the outcome.
- A value near zero indicates a poor model, and the input variables are irrelevant to the outcome.

Multiple Regression and Best-Subsets Regression (2)

- Inputs can be selected based on this approach, but the variance of the predicted output can increase with the inclusion of additional predictor variables. This can cause difficulty in selecting a subset when the number of variables in candidate subsets varies. In such situations, criteria that penalize model complexity are more useful in subset selection.
- Criteria such as Mallow’s C$_p$ statistic have been widely used to evaluate model complexity. This statistic suggests as the criterion the standardized total squared error computed as



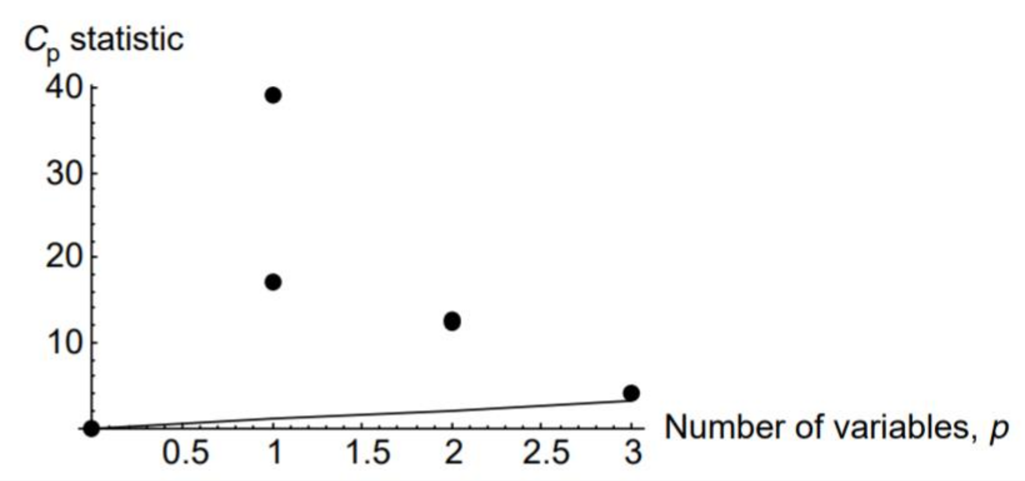
$$C_\mathrm{p} = \left(\frac{\mathrm{SS}_{\text{error},\, p}}{\mathrm{SS}_{\text{error, total}}}\right) - (n - 2p),$$

Multiple Regression and Best-Subsets Regression (3)

- $SS_{error,\ p}$ is the residual error for a multiple linear regression subset model with *p* inputs, and $SS_{error,\ total}$ is the residual error for the model with all *n* inputs.
- The correct model has C*p* value equal or smaller than *p* and a wrong model has a C*p* value larger than *p* due to a bias in the parameter estimation.
- Minimizing C*p* over all possible regression can give the best subset model.
- Good models typically have a (*p*, C*p*) coordinate close to a 45^o on a C*p* versus *p* plot.

Multiple Regression and Best-Subsets Regression (4)

- Regarding the thermal conductivity problem, if models are run with all possible subsets of inputs, the results for the most relevant subsets (others had higher C*p* values) illustrated in Figure below show that the best model has all the variables in the model.



- This case is denoted by the dot lying near the 45º line (note that the scales of the two axes are different).

Case Study

Given data of Singapore Airbnb which can be downloaded in this link

https://www.kaggle.com/jojoker/singapore-airbnb

1. Discuss to find the correlation and covariance of the data (the variables used can be categorical vs categorical, categorical vs numeric, or numeric vs numeric variables)
2. Normalized the data using Standardization or Range Scaling method.
3. Do feature selection and extraction.

End of Session 15 & 16

BINUS UNIVERSITY

References

- Sandhya Samarasinghe. (2006). *Neural Network for Applied Sciences and Engineering*. Auerbach Publications. ISBN: 978-0-8493-3375-0. Chapter 6.
- https://www.kaggle.com/jojoker/singapore-airbnb

People
Innovation
Excellence