STA3010 Regression Analysis

Feng Yin

The Chinese University of Hong Kong (Shenzhen)

February 22, 2020

Overview

Reference

The material of this lecture is mainly from Section 13.3 of "Introduction to Applied Linear Algebra" written by S. Boyd and L. Vandenberghe; and partly from Chapter 8 of our textbook.

Indicator Variables

We differentiate quantitative variables and qualitative/categorical variables in that:

1. Quantitative variables have well-defined scale of measurement such as temperature, distance, pressure, and income;
2. Qualitative/Categorical variables have no natural scale of measurement, such as employment status (employed or unemployed), shifts (day, evening, or night), and gender (male or female), etc.

For qualitative variable, we sort of have to manually assign a set of levels to account for the effect that the variable may have on the response. This is done through the use of indicator variables.

Example of Indicator Variables

Example 8.2 From Textbook

An electric utility studies the effect of the size of a single-family house and the type of air conditioning used in the house on the total electricity consumption during warm weather months. Let y be the total electricity consumption (in kilowatt-hours) during the period June through September and x_1 be the size of the house (square feet of floor space). There are four types of air conditioning systems: (1) no air conditioning, (2) window units, (3) heat pump, and (4) central air conditioning. The four levels of this factor can be modeled by three indicator variables, x_2 x_3, and x_4, defined as follows:

Type of Air Conditioning	x_2	x_3	x_4
No air conditioning	0	0	0
Window units	1	0	0
Heat pump	0	1	0
Central air conditioning	0	0	1

One-Hot-Encoding for Indicator Variables

- Expanding a categorical feature with m values into $m - 1$ features that encode whether the feature has one of the (non-default) values is sometimes called one-hot encoding, because for any data example, only one of the new feature values is one, and the others are zero.
- There is no need to expand an original feature that is Boolean (i.e., takes on two values). For example, use gender as an input.
- House price prediction example.

Unit Scaling/Standardization/Normalization

Example

Consider $y = 5 + \beta_1 x_1 + \beta_2 x_2$, where y is measured in liters, x_1 is a very large number measured in milliliters, and x_2 is a very small number measured in liters.

Motivation

Often, it is (numerically) helpful to work with scaled inputs and output that produce dimensionless model parameters/regression coefficients, which are known as standardized model parameters/regression coefficients.

Two popular ways are used to generate standardized model parameters, they are:

1. Unit normal scaling
2. Unit length scaling

Unit normal scaling: The idea is to scale the inputs and output as follows:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, ..., n, \;\; j = 1, 2, ..., k \tag{1}$$

$$y_i^* = \frac{y_i - \bar{y}}{s_y}, i = 1, 2, ..., n \tag{2}$$

where the scaling factors are computed by

$$s_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}, \qquad s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \tag{3}$$

It is easy to verify that both the scaled inputs and the scaled outputs have sample mean equal to zero and sample variance equal to 1.
Finally, we have a new regression model:

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + ... + b_k z_{ik} + \varepsilon_i, \; i = 1, 2, ..., n \tag{4}$$

where there is no need of b_0.

Unit length scaling: The idea is to scale the inputs and output as follows:

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{jj}^{1/2}}, i = 1, 2, ..., n, \ j = 1, 2, ..., k \tag{5}$$

$$y_i^0 = \frac{y_i - \bar{y}}{SS_T^{1/2}}, i = 1, 2, ..., n \tag{6}$$

where the scaling factors are computed by

$$s_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2, \qquad SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{7}$$

It is easy to verify that the length $\sqrt{\sum_{i=1}^{n} (w_{ij} - \bar{w}_j)^2} = 1$.
Finally, we have a new regression model:

$$y_i^0 = b_1 w_{i1} + b_2 w_{i2} + ... + b_k w_{ik} + \varepsilon_i, \ i = 1, 2, ..., n \tag{8}$$

where there is no need of b_0.

Handling Large Values

Winsorizing large values:

- It is common to clip the data when its value is much larger than expected and may be regarded as outlier.
- The term "winsorize" is named after the Statistician Charles P. Winsor.

Example: An input x_5 has already been normalized, and then winsorized with a threshold equal to 3:

$$\tilde{x}_5 = \begin{cases} x_5 & |x_5| \le 3 \\ 3 & x_5 > 3 \\ -3 & x_5 < -3. \end{cases}$$

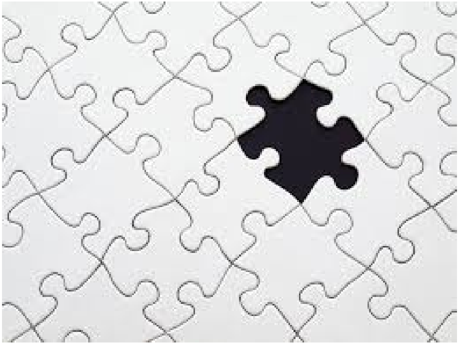
Log-Transform of large positive values:

- When feature values are positive and vary over a wide range, it is common to use their logarithms in practice.
- If the feature value may be equal to zero, where the logarithm is undefined, we often use the log-transformation $\tilde{x}_k = \log(x_k + 1)$ instead.
- The above strategies can be applied to the output values in the same way.

Example: Ambient PM 2.5 concentration in Beijing as an input x for a regression task.

Handling Missing Values

Dealing with missing data is a nightmare of almost every data scientist.



	district	dealDate	totalPrice(in Mio.)	tranCycle.day.	area	orientation	decoration	elevator	loft	years	metro
2	BaijiaLake	2017/6/28	325	18	138.32	6	NA	1	28	2006	1
3	BaijiaLake	2017/5/4	235	240	100.82	5	2	1	31	2009	1
4	BaijiaLake	2017/3/15	228.5	31	80.84	7	3	1	31	2008	1
5	BaijiaLake	2017/2/19	248	123	127.91	7	NA	1	17	2008	1
6	BaijiaLake	2016/11/27	278	99	137.76	7	3	1	28	2006	1
7	BaijiaLake	2016/9/29	290	19	127.91	7	1	1	17	2008	1
8	BaijiaLake	2016/9/26	205	37	88.56	2	3	1	28	2006	1
9	BaijiaLake	2016/9/24	238	82	105.32	0	3	1	25	2009	1

If the value of one or more entries are missing in a data point. We could do the following:

- Discard this data item when you have abundant of data.
- Replace the missing value with the sample mean of the corresponding input/feature.
- Construct a regression/imputation model to predict its value. [Additional Readings: "Multiple Imputation by Chained Equations: What is it and how does it work?", M. J. Azur *et.al.*, Int. J. Methods Psychiatr Res. 2011 March 1; 20(1): 4049. doi:10.1002/mpr.329.]
- Treat it as a latent variable and perform probabilistic inference.

Generating New Features

- Products and Interactions: New features can be developed from pairs of original features, for example, their product. Or more systematically using polynomial expansion.
- Random Features: The new features are given by a nonlinear function of a random linear combination of the original features. To add m new features of this type, we first generate a random $m \times p$ matrix $\boldsymbol{R}$. Generate new features as $|\boldsymbol{Rx}|$, which can be very effective.

Be aware: adding too many new features may lead to over-fit! Consider feature selection or take into account regularization.

Feature Selection

Delete redundant features in order to

- Reduce overfitting
- Improve modeling accuracy
- Save training time

Feature Selection Methods: PCC

For feature X_i and output Y, the Pearson correlation coefficient (PCC) is computed as

$$P(i) = \frac{Cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}} \tag{9}$$

Pearson correlation coefficient is only able to detect linear relationship.

Feature Selection Methods: MI

For feature X_i and output Y, assuming we know the joint pdf $p(x_i, y)$ and the marginals $p(x_i)$ and $p(y)$, the mutual information (MI) is computed as

$$M(i) = \int_{x_i} \int_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} dx_i dy. \tag{10}$$

For continuous variables, we need to perform discretization of the variables before computing the MI. Discretization granularity matters.

Feature Selection Methods: MC

For feature X_i and output Y, the maximum correlation (MC) [Hirschfeld, 1935][Gebelein, 1941][Rènyi, 1959]

$$\rho(X_i, Y) = \max_{f,g} \mathbb{E}[f(X_i)g(Y)]$$

- $f : \mathcal{X} \mapsto \mathbb{R}, \ g : \mathcal{Y} \mapsto \mathbb{R}$
- $\mathbb{E}[f(X_i)] = \mathbb{E}[g(Y)] = 0, \ \mathbb{E}[f^2(X_i)] = \mathbb{E}[g^2(Y)] = 1$

Maximum correlation is able to able to reveal almost all sorts of linear and nonlinear relationship.

Summary with Keywords

- Indicator variable
- Normalization
- Missing values
- Large values
- Create New Features
- Feature Selection