CPSC 4300/6300: Applied Data Science

2024

Data preparation for modeling

Dr. Aaron J. Masino
Associate Professor, School of Computing

College of
ENGINEERING, COMPUTING
AND APPLIED SCIENCES

Outline

- Data standardization
- Categorical variable transformation (dummy variables)

Data preparation

Data, especially secondary use retrospective data, is rarely "AI" ready. We will often need to perform:

- Feature selection / engineering
- Address missing values
- Determine segmentation strategy for training, validation, and testing

Covered in other lectures

- Scaling / standardization of continuous variables
- Conversion to dummy values for categorical variables

Data standardization / scaling of continuous features

- Empirical evidence suggest that most machine learning models perform better when continuous data is either standardized or scaled
- Standard practice in modern ML projects
- Standardization – transform data to have zero mean and unit standard deviation. Given samples of a continuous feature, *x*, the standardized values, *z*, are

$$z = \frac{x - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation of *x* over the samples

- Scaling – transform the data to fall within a range *[m, M]* usually *[-1, 1]*

$$x^* = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

$$z = x^*(M - m) + m$$

Mapping categorical features to dummy variables

Categorical features are typically transformed into a set of binary values (a.k.a. dummy variables, one-hot encoding)

One dummy variable is created for each possible value of the categorical feature

For a given sample, the dummy variable is set to 1 if the original value of the categorical variable was equal to the corresponding dummy variable and zero otherwise

The directly observed categorical variable values are not used in models. Instead, the dummy variables are used.

Example: species variable, *s*, with possible values: [dog, cat, fish] → s_dog in [0, 1], s_cat in [0, 1], s_fish in [0, 1]

The species variable, *s*, is converted into three (3) new variables, *s_dog, s_cat, s_fish* that are binary

Sample 1: s = *cat* → s_dog = 0, s_cat = 1, s_fish = 0

For a given sample, the value of *s* is observed. In this example *s=cat*. The sample is assigned the values 0, 1, 0 for the variables *s_dog, s_cat,* and *s_fish*, respectively.