UCLA CS97 Homework Assignment 1 Part 1

Yizhou Sun (yzsun@cs.ucla.edu)

June 19, 2024

0 Instruction and Preparation

Due date: Wednesday, 6/26 at 10:00pm PT **Instructions:** Be sure to clearly label where each problem and sub-problem begins. All problems must be submitted in order.

Goal: This homework aims to help you go through main concepts covered so far with toy examples.

Notations. A list of notations that might be new to you are provided below. (Skip this part if you already know them)

1. Summation Notation: $\sum$. When summing multiple items together, this notation can help us shorten the formula. For example, $x_1 + x_2 + x_3 + x_4 + x_5 + x_6$ can be written as $\sum_{i=1}^{6} x_i$, where i is the index for ith item.
2. Product Notation: $\prod$. Similar to summation notation, product notation is to help us to shorten the formula when *multiplying* multiple items together. For example, $x_1 \times x_2 \times x_3 \times x_4 \times x_5 \times x_6$ can be written as $\prod_{i=1}^{6} x_i$.

1 (20pt) Know your data

area (100 sq.ft.)	bedroom	zipcode	price ($100K)
15	2	11111	5
20	3	22222	7
18	3	11111	6
16	3	33333	5.5

Table 1.1: House Price Training Dataset

1.1 Mean

The **mean** of a set of n observations of a variable is denoted $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

(4pt) Exercise : Compute mean for *price* in Table 1.1.

1.2 Variance and Standard Deviation

The (sample) **variance** of a set of n observations of a variable is denoted as s^2 and is defined as:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x}^2)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2}$$

(4pt) Exercise : Compute variance for *price* in Table 1.1.

The **standard deviation** (std) is the square root of variance, denoted as s.

(2pt) Exercise : Compute standard deviation for *price* in Table 1.1.

1.3 Normalization

Normalization is to transform a numerical variable by re-centering and scaling. One of the most popular normalization is z-score normalization, which is also called standardization. For each observation x_i of variable X, it will be transformed by subtracting from mean and dividing by standard deviation:

$$x_i^{(new)} = \frac{x_i - \bar{x}}{s} \tag{3}$$

This process has been seen in our Homework 1 without disclosing the details.

(4pt) Exercise: Normalize *area* in Table 1.1 via z-score normalization.

1.4 One-hot encoding

In order to handle categorical features, we need to create dummy variables to convert discrete values into numerical ones. A standard way of doing so is to create a binary dummy variable for each possible value for that variable (e.g., in sklearn they adopt this way).

(4pt) Exercise: Write down the one-hot encoding for every data point for variable *zipcode* in Table 1.1. Hint: Fill in the table below.

area (100 sq.ft.)	bedroom	is_11111	is_22222	is_33333	price ($100K)
15	2				5
20	3				7
18	3				6
16	3				5.5

Table 1.2: House Price Training Dataset with One-hot Encoding

1.5 The relationship between two variables

Correlation between two variables is important to decide the relationship between two variables. Right now, you are not required to know the definition of correlation. Roughly, given two variables, X and Y, if Y increases when X increases, they are positively correlated. If Y decreases when X increases, they are negatively correlated. If no obvious trend between them, they are not correlated. Scatter plot gives us a good understanding of such correlation.

(2pt) Exercise: Are *area* and *price* correlated in Table 1.1? If yes, are they positively correlated or negatively correlated?