Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Numeric values

Now, on to the (typically) more manageable data type:

Concept 1

Typically, if your feature is **already a numeric value**, then we usually want to **keep it as a data value**.

Example: Heart rate, stock price, distance, reaction time, etc.

However, this may not be true if there is some difference between different ranges of numbers:

- Being below or above the age of 18 (or 21) for legal reasons
- Temperature above or below boiling
- Different age ranges of children might need different range sizes: the difference between ages 1-2 is very different from ages 7-8.

Concept 2

Sometimes, if there are distinct **breakpoints**/boundaries between different values of a numerical feature, we might use **discrete** features to represent those.

Standardizing Values

We still aren't done, if our data is numeric. We likely want to **scale** our features, so that they all tend to be in similar ranges.

Why is that? If some features are much **larger** than others, then they will have a much larger impact on the answer.

For example, suppose we have $x_1 = 4000$, $x_2 = 7$:

$$h(x) = \theta^T x = 4000\theta_1 + 7\theta_2 \tag{1}$$

The first term is going to have a way bigger impact on $h(x)$. If we change x_1 by 10%, that's going to be bigger than if we changed x_2 by 100%!

$4000 * 10\% = 400$
$7 * 100\% = 7$

Concept 3

If one **feature** is much **larger** than **another** feature, it will tend to have a much **larger** effect on the result.

This is often a bad thing: just because one feature is **larger**, doesn't mean it's more **important**!

Example: Income might be in the range of tens of thousands (10,000-100,000), while age is a two-digit number(20-100). Income will be weighed more heavily.

How do we solve that problem? We need to do two things:

- **Shift** the data so that our range is not too high/low. Our goal is to have it centered on 0.
  - We want it centered on 0 so we can distinguish between the above-average and below-average data points.
  - We do this by subtracting the **mean**, or the **average** of all of our data points.

Plus, it's easier to get all of our data to 0, rather than picking some arbitrary value.

$$\phi_1(x) = x - \overline{x} \tag{2}$$

- Scale the **range** of possible values, so they all vary by roughly the same amount.
- : So, if one variable tends to vary by a **larger** amount, it doesn't have a bigger impact on the result.

$$\phi(x_i) = \frac{x_i - \overline{x}_i}{\sigma_i} \tag{3}$$

Where σ is the **standard deviation**.

Note that each feature has its own σ_i: we have to compute this equation for each feature.

If you are interested, we define **standard deviation** below.

Definition 4

To make sure that all of our data is **on the same size scale**, we **normalize**/**standardize** our dataset using the operation

$$\phi(x_i) = \frac{x_i - \overline{x}_i}{\sigma_i}$$

For every variable x_i in a data point x.

- $\overline{x}_i$ is the **mean** of x_i
- σ_i is the **standard deviation** of x_i

This results in a dataset which has

- A mean $\overline{x}_i$ of **0**
- A standard deviation σ_i of **1**

So, all of our features have the same **average**, and **vary** by the same amount.

This prevents some features getting prioritized because they're on different size scales.

Example: Suppose we have 1-D data $x = [1, 2, 3, 4, 5, 6]$

The mean is

$$\overline{x} = \frac{1+2+3+4+5+6}{6} = 3.5 \tag{4}$$

And the standard deviation is

$$\sigma = \sqrt{\frac{2.5^2 + 1.5^2 + .5^2 + .5^2 + 1.5^2 + 2.5^2}{6}} = \sqrt{\frac{35}{12}} \approx 1.7078 \tag{5}$$

Variance and Standard Deviation (Optional)

This section* describes the origin of σ above. Feel free to skip if you're familiar.

In order to scale our data, we need a measure of how much our data **varies**. So, if our data varies by more, we can scale it down, and vice versa.

We can measure this using the **variance**.

Definition 5

We can measure how spread out/varying our data with **variance**

$$\sigma^2 = \sum_i \frac{(x^{(i)} - \overline{x})^2}{n} \tag{6}$$

In other words, the **average squared distance** from the **mean**.

Why do we square the terms? Same reason we square our loss:

- We want only positive values, for distance.
- We don't want to use absolute value, for smoothness.

We also get nicer statistical properties we won't discuss here.

However, this is too large: we want something similar to "average distance from the mean". This is the average **squared** distance.

So, we take a square root!

Definition 6

A more common way to measure how our data varies is using **standard deviation** σ

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_i \frac{(x - \overline{x})^2}{n}}$$

This term is **not** the average distance from the mean, but can be used for **scaling** our data in the same way.

This term allows us to scale our data appropriately. If our data varies by a larger amount, σ will be larger. So, $\frac{1}{\sigma}$ will cancel that variance out.