PRE-PROCESSING DATA

LECTURE 8
SECTION 2
JUNE 15TH

INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

UNIVERSITY of
RWANDA

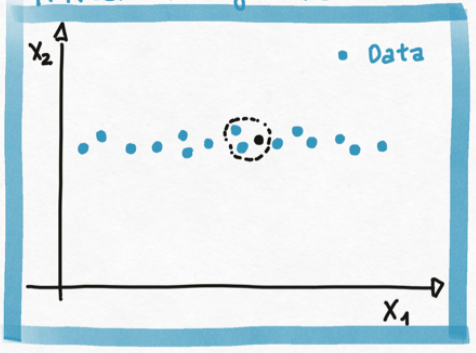
TRANSFORMING THE DATA

RESCALING THE DATA:

In model training and interpretation, we need to look at the units and scale of each covariate.

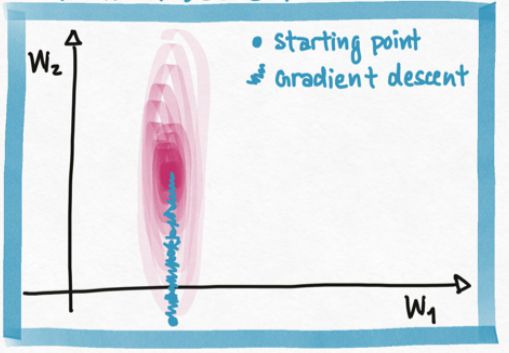
If the covariates have very different scales, this difference can skew training, interpretation and evaluation.

K-Nearest Neighbors



Nearest neighbor determined completely by x_1, x_2 doesn't even affect the computation.

Gradient Descent



If one dimension is very large then the gradient update for the coefficient for this dimension will be very small. This will cause slow convergence.

Model Interpretation

Interpret:

$y = 2{,}000x_1 + 0.1x_2 + 3$

y: price (C$)

x_1: distance to city center (mega m) } very small #

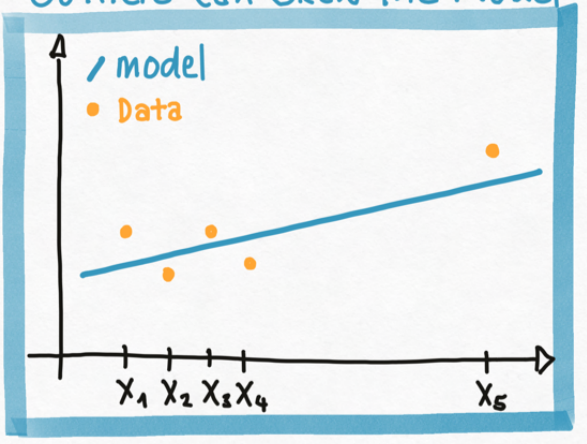
x_2: size (cm^2) } very large #

Knowing the units of the covariates can we still conclude that x_1 is more important than x_2?

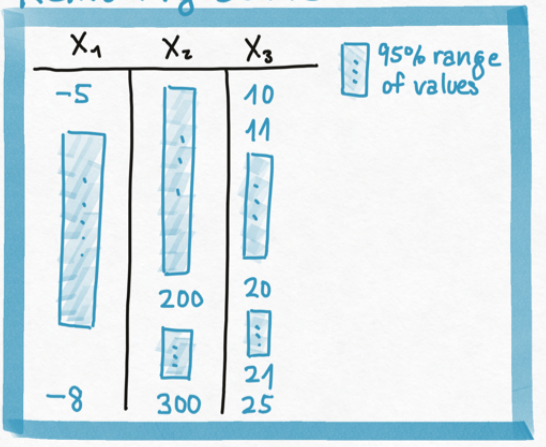
OUTLIERS IN THE DATA:

Often times in real data there are errors in measurements, either due to sensor errors or manual entry errors. These values can often be too large or too small when compared to real data entries, e.g. age = -10. We call these values outliers.

Outliers Can Skew the Model



Removing Outliers



Removing rows with an outlier value for X_i one covariate X_i at a time is not always a good idea, since an outlier may be a particular combination of values, e.g. salary = $100,000 and savings = -$2.00.

HANDLING MISSINGNESS

MISSING VALUES IN THE DATA:

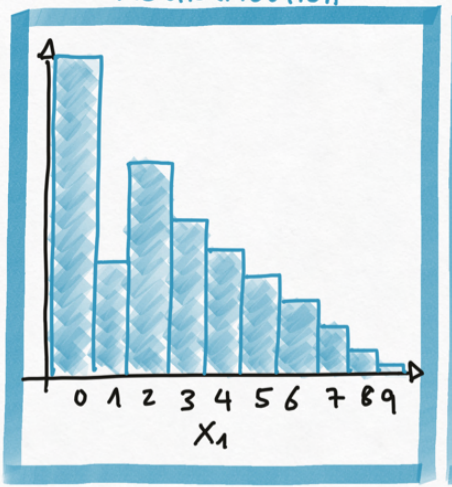
Often times there are missing values in the data. In order to feed the data into machine learning models f_w, we need to fill in these values. This is called imputation.

The easiest thing is to impute the missing values using zero, or the average value of the covariate.

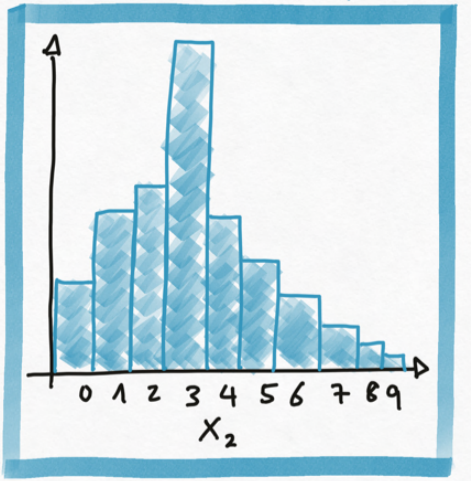
Imputing missing values

X_1	X_2	
0	2.5	
3	6.8	
10	4.1	
9	?	← replace with 0 or average (X_2)
?	0	
⋮	⋮	

Imputing with zero skews the distribution



Imputing with average skews the distribution



Imputing with constants can skew the distribution of the data, creating false signal.

MODEL BASED IMPUTATION:

We can impute missing values with non-constant values, i.e. using functions. These methods are called model-based imputation.

Regression Based Imputation

X_1	X_2
0	2.5
3	6.8
10	4.1
9	?
2	0
⋮	⋮

train model $f_w(X_1) = X_2$ on rows with no missing values

use trained model to predict: $f_w(9) = X_2$

kNN Imputation

X_1	X_2
0	2.5
3	6.8
10	4.1
9	?
2	0
⋮	⋮
8	3

Find the k-nearest neighbors of the row with missing values, based on the non-missing values. Then use the average of the neighbors to impute.

$X_2 = \frac{4.1+3}{2}$

We can re-use all our models for regression and classification to impute missing values: instead of predicting y using x_1, x_2, we use x_1 to predict missing values in x_2.