LINEAR REGRESSION

CSCI-B365

Introduction to Data Analysis and Mining
Linear Relationships

Consider 2 points of the form \((x, y)\). They are \((2, 3)\) and \((4, 8)\).

Question: how to make a line out of these two points?
Linear Relationships

Consider 2 points of the form \((x, y)\). They are \((2, 3)\) and \((4, 8)\).

Question: how to make a line out of these two points?

\[
y = k \cdot x + n
\]

Rise: \( k = \frac{p}{q} \)
Linear Relationships

Consider 2 points of the form \((x, y)\). They are \((2, 3)\) and \((4, 8)\).

Question: how to make a line out of these two points?

\[
y = k \cdot x + n
\]

**Intercept:**

\[
n = -2 \cdot k + 3
\]
Matlab Code

clear
clc

% define 2 data points a and b
a = [2 3];
b = [4 8];

% calculate rise and intercept
k = (b(2) - a(2))/(b(1) - a(1));
n = -k * a(1) + a(2);

% calculate points of the line
x = 0 : 0.1 : 8;
y = k * x + n;

% plot the line and two data points
plot(x, y, a(1), a(2), '*', ... 
     b(1), b(2), '*')
axis([0 8 -4 18])
Linear Relationships

Linear relationship given two points $a = (a_1, a_2)$ and $b = (b_1, b_2)$ is calculated as:

$$y - a_2 = \frac{b_2 - a_2}{b_1 - a_1} \cdot (x - a_1)$$
Linear Relationships with 3 Data Points

Consider now 3 points of the form \((x, y)\). They are \(a = (2, 3)\), \(c = (3, 5.5)\) and \(b = (4, 8)\).

Let's calculate the rise using points \(b\) and \(c\):

\[
k_{bc} = \frac{5.5 - 8}{3 - 4} = 2.5
\]

Let's calculate the rise using points \(a\) and \(c\):

\[
k_{ac} = \frac{5.5 - 3}{3 - 2} = 2.5
\]

Conclusion:
If three (or more) points are co-linear, it does not matter which two we use to construct a line.
3 Data Points that are not Co-Linear

Consider 3 different points of the form \((x, y)\). They are \(a = (2, 3)\), \(b = (4, 8)\), and \(c = (3.5, 5.5)\).

Clearly, we cannot draw a line that goes through all three data points.

Let's calculate rise using points \(a\) and \(c\):

\[
k_{ac} = \frac{5.5 - 3}{3.5 - 2} = 1.67
\]

Let's calculate rise using points \(b\) and \(c\):

\[
k_{bc} = \frac{5.5 - 8}{3.5 - 4} = 5
\]
Introducing Error

Square error for each point:

\[ e^{2}(a, \text{line}) = 0 \]
\[ e^{2}(b, \text{line}) = 0 \]
\[ e^{2}(c, \text{line}) = 1.25^2 = 1.56 \]

If we had 1\textsuperscript{st} component of vector \( c \) and used the line to predict its 2\textsuperscript{nd} component, \( e \) would be a prediction error.
What if $a$ and $c$ are Used for Line?

Square error for each point:

$e^2(a, \text{line}) = 0$
$e^2(b, \text{line}) = (-1.67)^2 = 2.78$
$e^2(c, \text{line}) = 0$
Given points a, b and c, find a line such that $(e_1^2 + e_2^2 + e_3^2)$ is minimized!
Application: Measuring Acceleration

Idea: apply force (F) to an object and measure its acceleration (a).

Use many different forces ($F_1, F_2, \ldots$) and measure acceleration for each force ($a_1, a_2, \ldots$). This will be our data set collection process.

FIGURE 1

Picture from: https://wiki.brown.edu/confluence/display/PhysicsLabs/LAB+3+-+FORCE+AND+ACCELERATION
Look at Data Points

F = m \cdot a

Steps:

• Collect data points (a, F)
• Do linear regression
• If you have a good instrument and do it right, your rise should be equal to object’s mass and intercept should be zero
Problem Formulation

Data set: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$

$f(x) = w_0 + w_1 \cdot x$

$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$
Problem Formulation

Given:

\[ D = \{(x_i, y_i)\}_{i=1}^{n} \]

Find:

best \( w_0 \) and \( w_1 \) (from the line equation \( f(x) = w_0 + w_1 \cdot x \))

Let's expand \( x \) into a vector \( \mathbf{x} = (1, x) \).
Let's use \( \mathbf{w} = (w_0, w_1) \)

Now,

\[
\begin{bmatrix}
  w_0 & w_1 \\
\end{bmatrix}
\begin{bmatrix}
  1 \\
  x \\
\end{bmatrix}
= w_0 + w_1 x
\]

\[ \rightarrow \]

\[ f(x) = \mathbf{w}^T \mathbf{x} = w_0 + w_1 x \]
Problem Formulation

Find coefficients \( w = (w_0, w_1) \) such that the sum of squared errors is minimized.

\[
SSE = \sum_{i=1}^{n} e_i^2
\]

More formally,

\[
w^* = \arg \min_w \sum_{i=1}^{n} e_i^2
\]

Or,

\[
w^* = \arg \min_w \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2
\]

\( w^* \) is an optimal solution
Solving the Problem

Sum of squared errors:

\[ SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2 \]

Find first derivatives and make them equal to zero:

\[
\frac{\partial SSE}{\partial w_0} = -2 \sum_{i=1}^{n} y_i + 2nw_0 + 2w_1 \sum_{i=1}^{n} x_i = 0
\]

\[
\frac{\partial SSE}{\partial w_1} = -2 \sum_{i=1}^{n} x_i y_i + 2w_0 \sum_{i=1}^{n} x_i + 2w_1 \sum_{i=1}^{n} x_i^2 = 0
\]

Solve the system of two equations with two unknowns:

\[ \mathbf{w}^* = (w_0^*, w_1^*) \]
Solving the Problem

Expand $X$ by adding a column of ones:

\[
X = \begin{bmatrix}
1 & x_1 \\
1 & x_2 \\
1 & x_3 \\
\vdots & \vdots \\
1 & x_n \\
\end{bmatrix} \quad y = \begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_n \\
\end{bmatrix}
\]

More elegant way to express the solution (without derivation):

\[
w^* = (X^T X)^{-1} X^T y
\]
Example, 2D input

\[
X = \begin{bmatrix}
3 & 5 \\
7 & 4 \\
2 & 7 \\
5 & 4 \\
9 & 4 \\
\end{bmatrix}
\quad
y = \begin{bmatrix}
2.3 \\
7.2 \\
2.5 \\
5.2 \\
8.4 \\
\end{bmatrix}
\quad
X = \begin{bmatrix}
1 & 3 & 5 \\
1 & 7 & 4 \\
1 & 2 & 7 \\
1 & 5 & 4 \\
1 & 9 & 4 \\
\end{bmatrix}
\quad
y = \begin{bmatrix}
2.3 \\
7.2 \\
2.5 \\
5.2 \\
8.4 \\
\end{bmatrix}
\]

\[w^* = (X^T X)^{-1} X^T y\]

\[w^* = (0.79, 0.99, 0.15)\]
General formulation

Expand $X$ by adding a column of ones:

$$X = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1k} \\
1 & x_{21} & x_{22} & \cdots & x_{2k} \\
1 & x_{31} & x_{32} & \cdots & x_{3k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix}$$

$y = \begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_n
\end{bmatrix}$

Solution:

$$w^* = (X^T X)^{-1} X^T y$$

Making a prediction on a previously unseen vector $\mathbf{x}$:

$$f(\mathbf{x}) = w^{*T} \mathbf{x}$$
Back to our Example

\[
\begin{align*}
\mathbf{e}_1 &= (2, 3) \\
\mathbf{b} &= (4, 8) \\
\mathbf{c} &= (3.5, 5.5) \\
\mathbf{X} &= \begin{bmatrix} 1 & 2 \\ 1 & 3.5 \\ 1 & 4 \end{bmatrix} \\
y &= \begin{bmatrix} 3 \\ 5.5 \\ 8 \end{bmatrix}
\end{align*}
\]
Matlab Code

clear
clc

% define 2 data points a and b
a = [2 3];
b = [4 8];
c = [3.5 5.5];

% define X and y
X = [1 2; 1 4; 1 3.5];
y = [3; 8; 5.5];

% calculate optimal coefficients
w = inv((X' * X)) * X' * y;

% calculate points of the line
x = 0 : 0.1 : 8;
y = w(2) * x + w(1);

% plot the line and two data points
plot(x, y, a(1), a(2), '*' , ... 
    b(1), b(2), '*' , c(1), c(2), '*')
axis([0 8 -4 18])

\[ w^* = \begin{bmatrix} -1.8077 \\ 2.3077 \end{bmatrix} \]

\[ y = w^{*T} \cdot x = 2.3 \cdot x - 1.8 \]
Calculate Errors

\[ f(x) = -1.81 + 2.31 \cdot x \]
\[ \mathbf{b} = (4, 8) \]
\[ \mathbf{c} = (3.5, 5.5) \]

Squared errors:
\[ e_1^2 = (-0.19)^2 = 0.04 \]
\[ e_3^2 = 0.58^2 = 0.33 \]
\[ e_2^2 = (-0.77)^2 = 0.59 \]

\[ e_1^2 + e_2^2 + e_3^2 = 0.96 \]

We minimized sum of squared errors!!!
Performance of Regression

\[ MSE = \frac{1}{n} \sum_{i=1}^{n} (f_i - y_i)^2 \]

\( f_i \): predicted target for data point \( i \)
\( y_i \): true target for data point \( i \)
\( n \): total number of data points

\[ MSE \in [0, \infty) \]

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (f_i - y_i)^2}{\sum_{i=1}^{n} (m - y_i)^2} \]

\( m \): mean value for the target variable

\[ R^2 \in (-\infty, 1] \]

\( R^2 \) is the percentage of variance “explained” by the regression method. Values close to 1 indicate perfect predictor.
Trivial Classifiers for Regression

Trivial classifier: always predicts target mean

R² for the trivial classifier is 0.