Linear algebra: regression

Statistical Natural Language Processing 1

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

Winter Semester 2025/2026

version: 2f4834d @2025-11-11

Quick recap

So far we reviewed:

- Vectors, matrices
- Operations on vectors and matrices: scalar multiplication, addition, dot product, matrix multiplication
- Matrices as operators (linear functions / transformations)
- Linearity and linear combinations
- Solving systems of linear equations, elimination
- Finding matrix inverse

Recap: solutions to systems of linear equations

For a $n \times m$ matrix $\mathbf{A}$

- Square, $n = m$
  - Unique solution if $\mathbf{A}$ is full rank $n = r$
  - Otherwise,
    - Infinite solutions if $\mathbf{b}$ is in the column space of $\mathbf{A}$
    - No solutions otherwise
- Rectangular, $n < m$ (wide matrix)
  - Infinite solutions if $\mathbf{b}$ is in the column space of $\mathbf{A}$
  - No solutions otherwise
- Rectangular, $n > m$ (tall/thin matrix)
  - Unique solution if $\mathbf{b}$ is in the column space of $\mathbf{A}$
  - No solutions otherwise

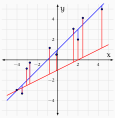
Linear regression

Linear regression is about finding a linear *model* of the form,

$$y = w_1 x + w_0$$

where,

- y is a numeric quantity we want to predict
- x is a measurement/value helpful for predicting y
- w_0 and w_1 are the parameters that we want to learn from data
- both x and y can be vector valued



Linear regression: and alternative view

this lecture

- Linear regression is also about finding the closest solution to a system of equations without a solution
- Given a dataset like

x_1	x_2	y
250.39	5.21	4913.19
332.18	3.77	59.67
312.47	1.26	154.42
272.01	7.01	166.27

- Find the closest solution to $\mathbf{X}w = \mathbf{y}$
- In other words, we solve $\mathbf{X}w = \mathbf{p}$, where $\mathbf{p}$ is a vector that allows the system to be solved, and it the closest such vector to $\mathbf{y}$

A simple example

- Let's take

$$\mathbf{x} = \begin{bmatrix}4\\2\end{bmatrix} \quad \mathbf{y} = \begin{bmatrix}1\\2\end{bmatrix}$$

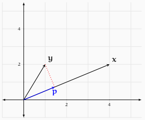
- We want to solve,

$$\mathbf{x}w = \mathbf{y}$$

- Instead we solve,

$$\mathbf{x}w = \mathbf{p}$$

where $\mathbf{p}$ is the orthogonal projection of $\mathbf{y}$ onto the line defined by $\mathbf{x}$



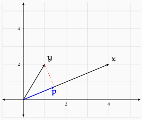
Finding the projection

- $\mathbf{p}$ is a scalar multiple (linear combination) of $\mathbf{x}$: $\mathbf{p} = \mathbf{x}w$
- We know that the length of $\mathbf{p}$ is the normalized dot product $\mathbf{x}^T\mathbf{y}/\|\mathbf{x}\|$
- We get the projection, if we multiply this with the unit vector in $\mathbf{x}$ direction

$$p = \frac{\mathbf{x}}{\|\mathbf{x}\|}\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|} = \frac{\mathbf{x}\mathbf{x}^T}{\|\mathbf{x}\|^2}\mathbf{y} = \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{x}}\mathbf{y}$$

- w, in this case is also easy:

$$w = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}}$$



Finding the projection

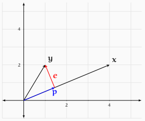
a slightly different explanation

- Note that $\mathbf{e} = \mathbf{y} - \mathbf{p}$
- Since $\mathbf{x}$ and $\mathbf{e}$ are orthogonal

$$\mathbf{x}^T(\mathbf{y} - \mathbf{x}w) = 0$$
$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}w$$
$$w = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}}$$

- Since we defined $\mathbf{p} = \mathbf{x}w$,

$$\mathbf{p} = \mathbf{x}\frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}} = \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{x}}\mathbf{y}$$



Solution to the simple regression example

- Our 'training' gives us

$$w = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}}$$

- For future x values, the prediction of y is

$$y = wx$$

For our example,

$$\mathbf{x} = \begin{bmatrix}4\\2\end{bmatrix} \quad \mathbf{y} = \begin{bmatrix}1\\2\end{bmatrix}$$

- $w = \frac{2}{5}$
- The model:

$$y = \frac{2}{5}x$$

Questions:

- what is the error $\mathbf{e}$ on the training instances?
- what is $\mathbf{e}^T\mathbf{x}$?

The other picture of the solution

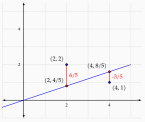
- The model: $y = \frac{2}{5}x$
- Predictions:

$$p = \begin{bmatrix}4 \times 2/5\\2 \times 2/5\end{bmatrix} = \begin{bmatrix}8/5\\4/5\end{bmatrix}$$

- Error:

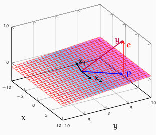
$$e = \begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}8/5\\4/5\end{bmatrix} = \begin{bmatrix}-3/5\\6/5\end{bmatrix}$$

- Is this a good model?



Linear regression in higher dimensions

- In higher dimensional spaces we want the projection onto the column space of $\mathbf{X}$
- The error vector $\mathbf{e}$ is perpendicular to all column vectors of $\mathbf{X}$, $\mathbf{x}_i$
- Again, note that $\mathbf{e} = \mathbf{y} - \mathbf{p}$



Deriving linear regression on higher dimensions

$\mathbf{X}^T(\mathbf{y} - \mathbf{p}) = 0$	Error vector is orthogonal to columns
$\mathbf{X}^T(\mathbf{y} - \mathbf{X}w) = 0$	$\mathbf{p}$ is the weighted combination of columns
$\mathbf{X}^T\mathbf{X}w = \mathbf{X}^T\mathbf{y}$	Note: $\mathbf{X}^T\mathbf{X}$ is square
$w = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	The final solution

The projection of $\mathbf{y}$ onto columns space of $\mathbf{X}$ is

$$\mathbf{p} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The intercept (bias) term

- The models we fit so far are 'linear',

$$y = w_1x_1 + w_2x_2 + \ldots + w_mx_m$$

they are forced to include $y = 0$ for $\mathbf{x} = \mathbf{0}$
- In most (almost all) cases, this is too restrictive, we also want to learn an intercept term

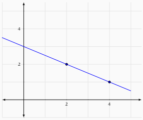
$$y = w_0 + w_1x_1 + w_2x_2 + \ldots + w_mx_m$$

- A straightforward solution is to include an artificial column of 1s in the input matrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & 2 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Solution with the intercept term

- Solution: $w_0 = 3$, $w_1 = -1/2$
- The model: $y = 3 - 1/2x$
- Is this a better model?



Regression in the real world

- In this lecture, we focused on finding the best fit to the data
- This may (very likely) result in *overfitting*
- To prevent overfitting, we
  - use *regularization*
  - **never rely on performance on the training set**, success should only be measured on a *held-out* data set
- We will return to these concepts later

Summary / next

- We reviewed regression as a way to find an approximate solution to a system of linear equations
- We will come back to regression multiple times

Next:
- Determinant, eigenvalues/eigenvectors, SVD

Further reading

Any of the linear algebra references provided earlier.