Regression: the optimization view

Statistical Natural Language Processing 1

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

Winter Semester 2024/2025

version: 986f20e @2024-12-12

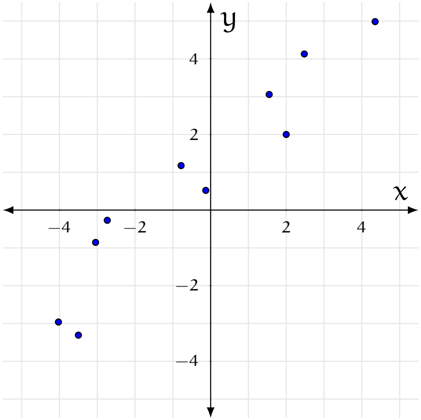
Linear regression

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



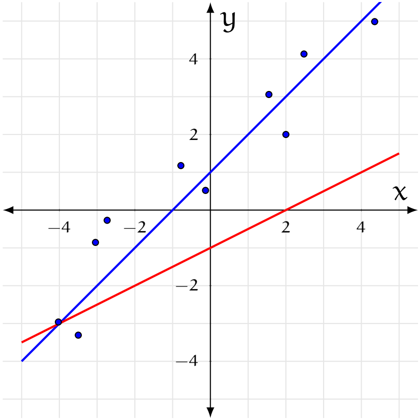
Linear regression

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



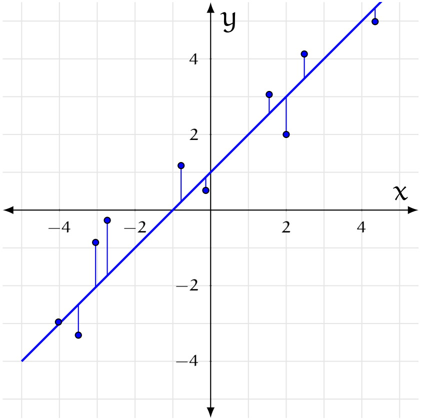
Linear regression

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



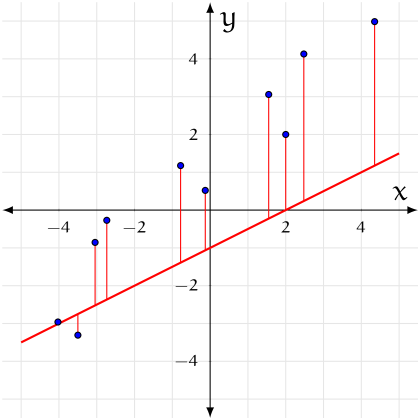
Linear regression

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

$$y = w_1 x + w_0$$

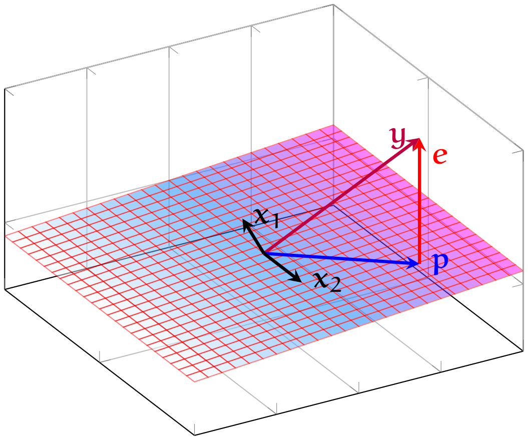
where,

- 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: the linear algebra approach

- We want to find $\mathbf{X}w = \mathbf{y}$, but the system is overdetermined, there is no unique solution
- Only possible solutions exists in the column space of $\mathbf{X}$
- The closest vector to $\mathbf{y}$, in the column space of $\mathbf{X}$ is the orthogonal projection $\mathbf{p}$
- The error $\mathbf{e} = \mathbf{y} - \mathbf{p}$



Deriving linear regression with linear algebra

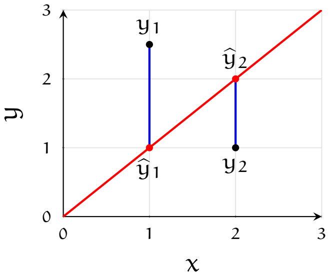
$$
\begin{aligned}
\mathbf{X}^{\mathsf{T}}(\mathbf{y}-\mathbf{p}) &= 0 && \text{Error vector is orthogonal to columns} \\
\mathbf{X}^{\mathsf{T}}(\mathbf{y}-\mathbf{X}w) &= 0 && \mathbf{p} \text{ is the weighted combination of columns} \\
\mathbf{X}^{\mathsf{T}}\mathbf{X}w &= \mathbf{X}^{\mathsf{T}}\mathbf{y} && \text{Note: } \mathbf{X}^{\mathsf{T}}\mathbf{X} \text{ is square (and invertible if } \mathbf{X} \text{ has indep. columns)} \\
w &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} && \text{The final solution}
\end{aligned}
$$

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

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

Estimating regression parameters

- We view learning as a search for the regression equation with least error
- The error terms are also called *residuals*
- We want error to be low for the whole training set: average (or sum) of the error has to be reduced
- Can we minimize the sum of the errors?



$$y_i = \underbrace{w_0 + w_1 x_i}_{\hat{y}_i} + e_i$$

$$e_i = y_i - \underbrace{w_0 + w_1 x_i}_{\hat{y}_i}$$

Least squares regression

In least squares regression, we want to find w_0 and w_1 values that minimize

$$E(\boldsymbol{w}) = \sum_i (y_i - (w_0 + w_1 x_i))^2$$

- Note that $E(\boldsymbol{w})$ is a *quadratic* function of $\boldsymbol{w} = (w_0, w_1)$
- As a result, $E(\boldsymbol{w})$ is *convex* and have a single extreme value
  - there is a unique solution for our minimization problem
- In case of least squares regression, there is an analytic solution
- Even if we do not have an analytic solution, if the error function is convex, a search procedure like *gradient descent* can still find the *global minimum*

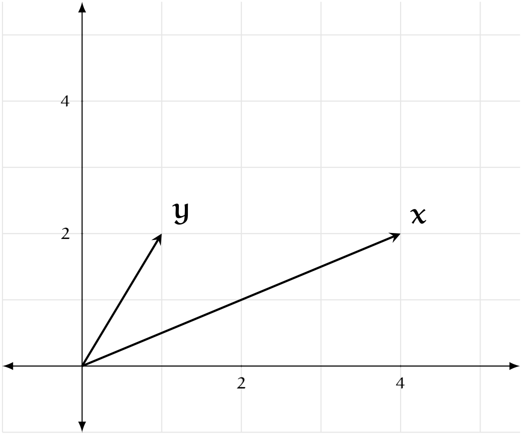
A simple example

earlier solution with linear algebra

- The data:

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

We want to solve, $\boldsymbol{x}w = \boldsymbol{y}$, but not solvable



A simple example

earlier solution with linear algebra

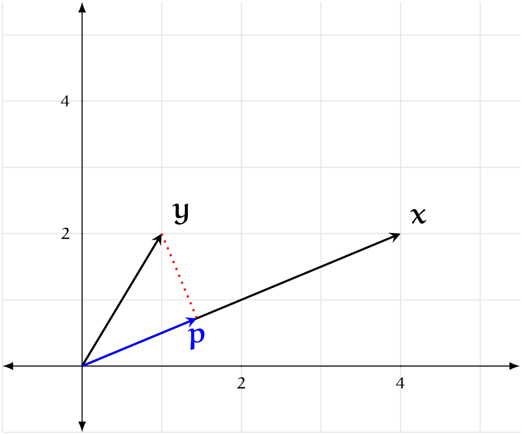
- The data:

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

  We want to solve, $\mathbf{x}w = \mathbf{y}$, but not solvable

- Instead we solve, $\mathbf{x}w = \mathbf{p}$,

$$w = \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\mathbf{x}^\mathsf{T}\mathbf{x}} = \frac{4 \times 1 + 2 \times 2}{4 \times 4 + 2 \times 2} = \frac{2}{5}$$



A simple example

optimization approach

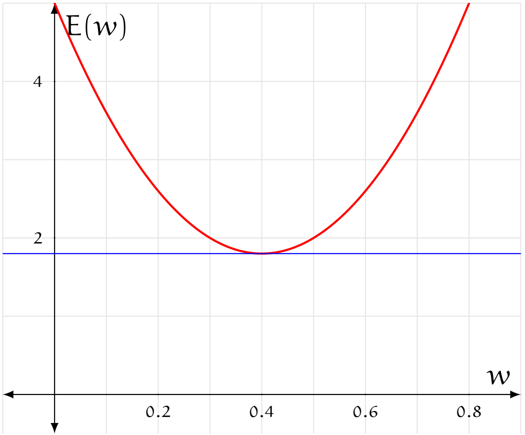
- Data: $\mathbf{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

  Model: $\hat{y} = wx$

A simple example
optimization approach

- Data: $\mathbf{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$
  Model: $\hat{y} = wx$
- Squared errors

$$\begin{aligned} E(w) &= (4w - 1)^2 + (2w - 2)^2 \\ &= 20w^2 - 16w + 5 \end{aligned}$$



A simple example

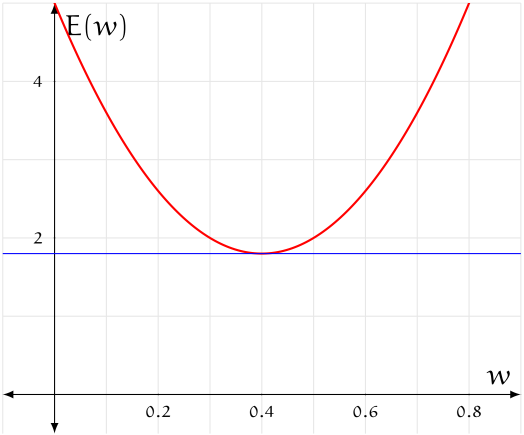
optimization approach

- Data: $\mathbf{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$
  Model: $\hat{y} = wx$
- Squared errors

$$\begin{aligned} E(w) &= (4w - 1)^2 + (2w - 2)^2 \\ &= 20w^2 - 16w + 5 \end{aligned}$$

- Setting the derivative to zero:

$$\frac{dE}{dw} = 40w - 16 = 0 \Rightarrow w = \frac{2}{5}$$



A simple example

extending with the bias term

- Data: $\boldsymbol{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

  Model: $\hat{y} = w_0 + w_1 x$

A simple example

extending with the bias term

- Data: $\boldsymbol{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

  Model: $\hat{y} = w_0 + w_1 x$
- Squared errors

$$\begin{aligned} E(w) &= (w_0 + 4w_1 - 1)^2 + (w_0 + 2w_1 - 2)^2 \\ &= 2w_0^2 + 20w_1^2 + 12w_0 w_1 - 6w_0 - 8w_1 + 5 \end{aligned}$$

A simple example

extending with the bias term

- Data: $\mathbf{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$
  Model: $\hat{y} = w_0 + w_1 x$
- Squared errors

$$
\begin{aligned}
E(w) &= (w_0 + 4w_1 - 1)^2 + (w_0 + 2w_1 - 2)^2 \\
&= 2w_0^2 + 20w_1^2 + 12w_0w_1 - 6w_0 - 8w_1 + 5
\end{aligned}
$$

- Partial derivatives

$$
\begin{aligned}
\frac{\partial E}{\partial w_0} &= 2w_0 + 12w_1 - 6 \\
\frac{\partial E}{\partial w_1} &= 12w_0 + 40w_1 - 16
\end{aligned}
$$

A simple example

extending with the bias term

- Data: $\boldsymbol{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

  Model: $\hat{y} = w_0 + w_1 x$
- Squared errors

$$\begin{aligned} E(w) &= (w_0 + 4w_1 - 1)^2 + (w_0 + 2w_1 - 2)^2 \\ &= 2w_0^2 + 20w_1^2 + 12w_0 w_1 - 6w_0 - 8w_1 + 5 \end{aligned}$$

- Partial derivatives

$$\frac{\partial E}{\partial w_0} = 2w_0 + 12w_1 - 6$$
$$\frac{\partial E}{\partial w_1} = 12w_0 + 40w_1 - 16$$

- Gradient:

$$\nabla E(\boldsymbol{w}) = \begin{bmatrix} 4w_0 + 12w_1 - 6 \\ 12w_0 + 40w_1 - 16 \end{bmatrix}$$

A simple example

extending with the bias term

- Data: $\boldsymbol{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

  Model: $\hat{y} = w_0 + w_1 x$
- Squared errors

$$
\begin{aligned}
E(w) &= (w_0 + 4w_1 - 1)^2 + (w_0 + 2w_1 - 2)^2 \\
&= 2w_0^2 + 20w_1^2 + 12w_0w_1 - 6w_0 - 8w_1 + 5
\end{aligned}
$$

- Partial derivatives

$$
\begin{aligned}
\frac{\partial E}{\partial w_0} &= 2w_0 + 12w_1 - 6 \\
\frac{\partial E}{\partial w_1} &= 12w_0 + 40w_1 - 16
\end{aligned}
$$

- Gradient:

$$
\nabla E(\boldsymbol{w}) = \begin{bmatrix} 4w_0 + 12w_1 - 6 \\ 12w_0 + 40w_1 - 16 \end{bmatrix}
$$

- Settings $\nabla E(\boldsymbol{w}) = \boldsymbol{0}$,

$$
\begin{bmatrix} 4 & 12 \\ 12 & 40 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 6 \\ 16 \end{bmatrix}
$$

A simple example

extending with the bias term

- Data: $\boldsymbol{x} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

  Model: $\hat{y} = w_0 + w_1 x$
- Squared errors

$$
\begin{aligned}
E(w) &= (w_0 + 4w_1 - 1)^2 + (w_0 + 2w_1 - 2)^2 \\
&= 2w_0^2 + 20w_1^2 + 12w_0w_1 - 6w_0 - 8w_1 + 5
\end{aligned}
$$

- Partial derivatives

$$
\begin{aligned}
\frac{\partial E}{\partial w_0} &= 2w_0 + 12w_1 - 6 \\
\frac{\partial E}{\partial w_1} &= 12w_0 + 40w_1 - 16
\end{aligned}
$$

- Gradient:

$$\nabla E(\boldsymbol{w}) = \begin{bmatrix} 4w_0 + 12w_1 - 6 \\ 12w_0 + 40w_1 - 16 \end{bmatrix}$$

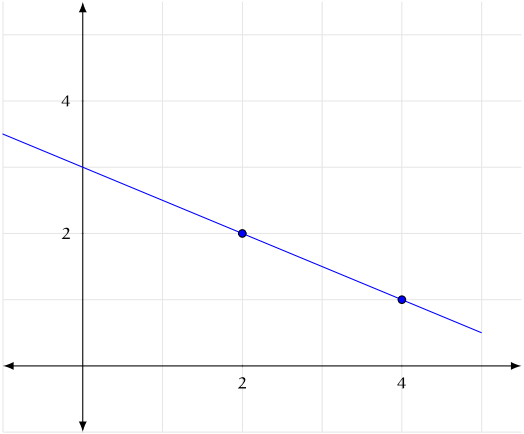
- Settings $\nabla E(\boldsymbol{w}) = \boldsymbol{0}$,

$$\begin{bmatrix} 4 & 12 \\ 12 & 40 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 6 \\ 16 \end{bmatrix}$$

- Solution: $\boldsymbol{w} = \begin{bmatrix} 3 \\ -1/2 \end{bmatrix}$

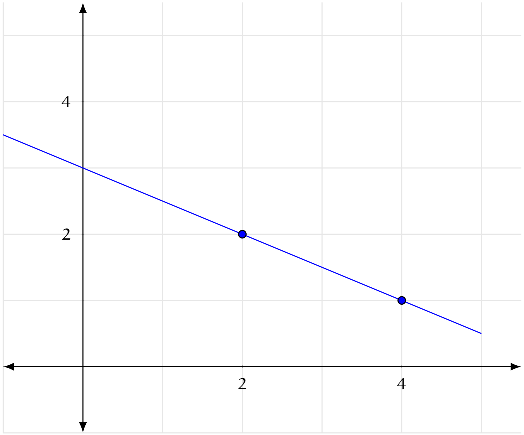
Solution with the intercept term

- Solution: $w_0 = 3, w_1 = -1/2$



Solution with the intercept term

- Solution: $w_0 = 3, w_1 = -1/2$
- The model: $y = 3 - 1/2x$



Regression with multiple predictors

$$y_i = \underbrace{w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \ldots + w_k x_{i,k}}_{\hat{y}} + e_i = \boldsymbol{w}\boldsymbol{x}_i + e_i$$

w_0 is the intercept (as before).

$w_{1..k}$ are the coefficients of the respective predictors.

e is the error term (residual).

- using the vector notation the equation becomes:

$$y_i = \boldsymbol{w}\boldsymbol{x}_i + e_i$$

where $\boldsymbol{w} = (w_0, w_1, \ldots, w_k)$ and $\boldsymbol{x}_i = (1, x_{i,1}, \ldots, x_{i,k})$

Note that the least square error, $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}$ is still quadratic in $\boldsymbol{w}$.

Evaluating machine learning systems

- Any (machine learning) system needs a way to measure its success
- For measuring success (or failure) in a machine learning system we need quantitative measures
- Remember that we need to measure the success outside the training data

Measuring success in Regression

- *Root-mean-square error* (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2}$$

measures average error in the units compatible with the outcome variable.

- Another well-known measure is the *coefficient of determination*

$$R^2 = \frac{\sum_i^n (\hat{y}_i - \mu_y)^2}{\sum_i^n (y_i - \mu_y)^2} = 1 - \left(\frac{\text{RMSE}}{\sigma_y}\right)^2$$

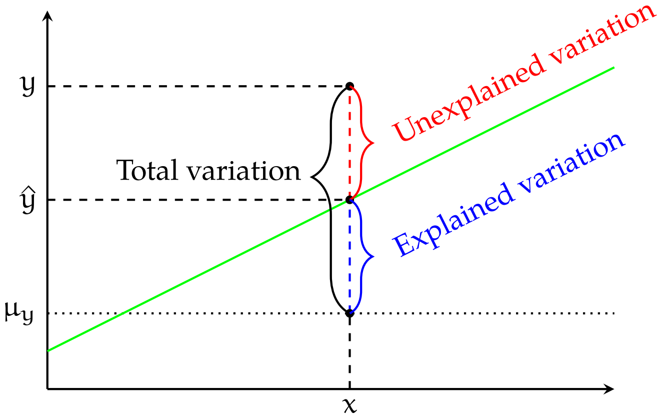
Assessing the model fit: R^2

We can express the variation explained by a regression model as:

$$\frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum_i^n (\hat{y}_i - \mu_y)^2}{\sum_i^n (y_i - \mu_y)^2}$$

- In simple regression, it is the square of the correlation coefficient between the outcome and the predictor
- The range of R^2 is $[0, 1]$
- $100 \times R^2$ is interpreted as 'the percentage of variance explained by the model'
- R^2 shows how well the model fits to the data: closer the data points to the regression line, higher the value of R^2

Explained variation



Total variation	=	Unexplained variation	+	Explained variation
$y - \mu_y$	=	$y - \hat{y}$	+	$\hat{y} - \mu_y$

Some cautionary notes

- Least-square regression is sensitive to *outliers*, large errors contribute more when minimizing squares
- It is always a good idea to inspect the data
- Other (robust) methods are also available (e.g., least absolute deviations)
- Other (robust) methods are also available

Summary / next

- We reviewed regression as finding the minimum error through differentiation
- We will come back to regression multiple times

Summary / next

- We reviewed regression as finding the minimum error through differentiation
- We will come back to regression multiple times

Next:

- Probability theory
- Reading: probability theory tutorial by Goldwater (2018)