Recap: regression

Statistical Methods in NLP 2
ISCL-BA-08

Çağrı Çöltekin
ccoltekin@sfs.uni-tuebingen.de

University of Tübingen
Seminar für Sprachwissenschaft

Summer Semester 2025

version: 1718995 @2025-04-30

Regression

- In ML, regression refers to a problem where outcome variable is a number (numeric quantity, continuous random variable)
- In practice, regression most commonly refers to *linear regression*, solved using *least squares* optimization
- We will review solving the linear regression problem from different viewpoints

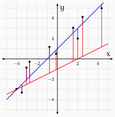
Linear regression

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

$$y = w_1 x + w_0$$

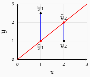
where,

- 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



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*

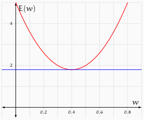
A simple example

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

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



Regression with multiple variables

- The example generalizes to more parameters
- Instead of derivative with respect to a single variable, we calculate the gradient
- The solution is where the gradient is 0
- This leads to a system of linear equations, whose solution vector is the best parameters

Maximum Likelihood Estimation (MLE)

- In MLE the task is to find the model m that assigns the maximum *likelihood* to the observed data $\boldsymbol{x}$
- To emphasize that likelihood is a function of model parameters, $\boldsymbol{w}$, we indicate it as $\mathcal{L}(w;\boldsymbol{x})$
- Formally, the task is finding

$$\boldsymbol{w}_{MLE} = \arg\max_w \mathcal{L}(w;\boldsymbol{x})$$

- In most cases, working with $\log$ likelihood is easier, since $\log$ is a monotonically increasing function,

$$\boldsymbol{w}_{MLE} = \arg\max_w \log \mathcal{L}(w;\boldsymbol{x}) = \arg\min_w -\log \mathcal{L}(w;\boldsymbol{x})$$

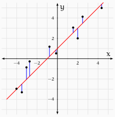
MLE for simple regression

$$y_i = w_0 + w_1 x_i + \epsilon_i$$

where $\epsilon \sim \mathcal{N}(0, \sigma)$

- We additionally assume that σ is independent of x
- This means $y \sim \mathcal{N}(w_0 + w_1 x, \sigma)$
- Now the likelihood function becomes,

$$\prod_{i=1}^{n} \frac{e^{-\frac{(y_i - (w_0 + w_1 x_i))^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$



MLE for simple regression (2)

Log likelihood: $\quad -n \ln \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$

- Note that maximizing log likelihood is equivalent to minimizing

$$\sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

- This is the squared error (the same as what we did before)
- MLE estimate of the regression parameters is equivalent to least-squares regression

Approximate solutions to systems of linear equations

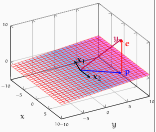
$$\begin{bmatrix} 4 & 2 \\ 2 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 4 \\ 4 \end{bmatrix}$$

- Can we solve the equation above?
- Can we find the 'best' approximation?
- Reminder: finding the solution means

$$\begin{bmatrix} 4 \\ 2 \\ 4 \end{bmatrix} w_1 + \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} w_2 = \begin{bmatrix} 10 \\ 4 \\ 4 \end{bmatrix}$$

Picture of the (non)solution

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



Deriving linear regression on higher dimensions

$$\mathbf{X}^\mathsf{T}(\mathbf{y}-\mathbf{p}) = 0$$ Error vector is orthogonal to columns

$$\mathbf{X}^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{w}) = 0$$ $\mathbf{p}$ is the weighted combination of columns

$$\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{w} = \mathbf{X}^\mathsf{T}\mathbf{y}$$ Note: $\mathbf{X}^\mathsf{T}\mathbf{X}$ is square

$$\boldsymbol{w} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$ The final solution

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

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

Pseudo inverse

- We want matrix multiplication to get as close to $\mathbf{I}$ as possible. Consider the 3×4 diagonal matrix:

$$\begin{bmatrix} 1/a & 0 & 0 \\ 0 & 1/b & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- For an $n \times n$ diagonal matrix $\mathbf{\Sigma}$, $\mathbf{\Sigma}^+ = \mathbf{\Sigma}^{-1}$
- For any invertible $n \times n$ matrix $\mathbf{X}$, $\mathbf{X}^+ = \mathbf{X}^{-1}$
- Singular value decomposition (SVD) provides the general solution:

$$\mathbf{X}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\mathsf{T}$$

Regression through pseudo inverse

- Pseudo inverse is another method to find the regression parameters
- We want

$$X\boldsymbol{w} = \mathbf{y}$$

but there is no general solution. Multiplying both sides with the pseudo inverse results in the best approximation

$$X^+X\boldsymbol{w} \approx X^+\mathbf{y}$$
$$\boldsymbol{w} \approx X^+\mathbf{y}$$

- This also allows regression with 'wide' matrices, in which case we get lowest L2-norm solution

Final remarks

- Regression is probably the most popular method for all (scientific) research
- Many statistical methods are variations/extensions of regression
- Regression is also part of almost any ML method

Next:

- Recap: classification / evaluation

Some sources of information

- Any modern linear algebra book (e.g., Strang, 2009) would cover regression
- For a more ML-focused introductions, also see James et al. (2024) or any machine learning textbook

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor (2024). *An introduction to statistical learning*. Springer. ISBN: 9783031391897. URL: https://www.statlearning.com/.

Strang, Gilbert (2009). *Introduction to Linear Algebra, Fourth Edition*. 4th ed. Wellesley Cambridge Press. ISBN: 9780980232714.