

Çağrı Çöltekin

University of Tübingen  
Seminar für Sprachwissenschaft

Winter Semester 2024/2025

https://nlp1.seminar.uni-tuebingen.de/

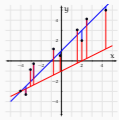
## Linear regression

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

$$y = w_1 x + w_0$$

where,

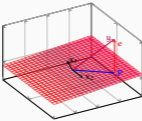
- $y$  is a numeric quantity we want to predict
- $x$  is a measurement/value helpful for predicting  $y$
- $w_0$  and  $w_1$  are the parameters that we want to learn from data
- both  $x$  and  $y$  can be vector valued



Recap: MLE: MLE for regression

## Linear regression: the linear algebra approach

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



Recap: MLE: MLE for regression

## Deriving linear regression with linear algebra

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

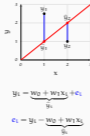
The projection of  $y$  onto column space of  $X$  is

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

Recap: MLE: MLE for regression

## Regression as optimization: finding minimum error

- 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?



Recap: MLE: MLE for regression

## Least squares regression

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

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

- Note that  $E(w)$  is a quadratic function of  $w = (w_0, w_1)$
- As a result,  $E(w)$  is convex and has a unique 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*

Recap: MLE: MLE for regression

## Learning as finding the best model

- In most ML problems, learning is viewed as finding the best (parametric) model among a family of models
- The task is finding  $m$  given the input  $x$  such that  $P(m|x)$  is the largest

$$P(m|x) = \frac{P(m)P(x|m)}{P(x)}$$

- A Bayesian learner, learns a (proper) distribution for the posterior  $P(m|x)$
- Estimating only the model with the highest posterior is called *maximum a posteriori* (MAP) estimation
- Finding the model with the highest likelihood,  $P(x|m)$  is called *maximum likelihood estimation* (MLE)

Recap: MLE: MLE for regression

## Maximum Likelihood Estimation (MLE)

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

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

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

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

Recap: MLE: MLE for regression

## MLE: simple example with coin flips

- Assume we observed  $x = 0110110011$  (0 = tail, 1 = head)
- If coin is fair (parameter  $p = 0.5$ ), what is the likelihood of obtaining the sample above?

$$p(x|p=0.5) = 0.5^8 (1-0.5)^4 = \frac{1}{1024} = 0.000977$$

- If coin is biased towards  $T$  with  $p = 0.4$ , what is the likelihood of obtaining the sample?

$$p(x|p=0.4) = 0.4^8 (1-0.4)^4 = \frac{1}{1024} = 0.000531$$

- What is the model (specified with parameter  $p$ ) with the maximum likelihood?

Recap: MLE: MLE for regression

## MLE: example with coin flips

finding the maximum likelihood

- For a trial with  $n_H$  heads and  $n_T$  tails, the likelihood function is

$$\mathcal{L}(p;x) = p^{n_H} (1-p)^{n_T}$$

- Working with logarithms is easier

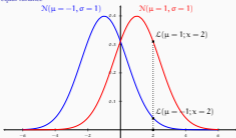
$$p_{MLE} = \arg \max_p \log p^{n_H} (1-p)^{n_T} = \arg \max_p n_H \ln p + n_T \ln (1-p)$$

- Taking the partial derivative with respect to  $p$ , and setting it to 0

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{n_H}{p} - \frac{n_T}{1-p} = 0 \Rightarrow p = \frac{n_H}{n_H + n_T}$$

Recap: MLE: MLE for regression

## Another example: the mean of the Normal distribution with known/regular variance



Recap: MLE: MLE for regression

## MLE for the parameters of Normal distribution

Given  $n$  independent samples,  $x = \{x_1, \dots, x_n\}$ ,

Likelihood:  $\mathcal{L}(\mu, \sigma; x) = \prod_{i=1}^n p(x_i) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$ , we want  $\arg \max_{\mu, \sigma} \mathcal{L}(\mu, \sigma; x)$

$$\text{Log likelihood: } \mathcal{L}(\mu, \sigma; x) = n \ln \frac{1}{\sigma \sqrt{2\pi}} + n \ln \frac{1}{\sigma^2} + \frac{n}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2} \left( \sum_{i=1}^n x_i - n\mu \right), \quad \frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{n}{\sigma} + \frac{n}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2$$

$$\mu_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i \quad \sigma_{MLE} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{MLE})^2$$

## Properties of MLE

- In the limit ( $n \rightarrow \infty$ ), MLE estimate is (asymptotically) correct
- MLE estimate is consistent, more data results in more accurate estimate
- MLE estimates are asymptotically normal: estimates from a large number of samples is distributed normally
- MLE estimate can be *biased*

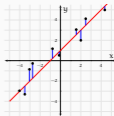
## MLE for simple regression

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

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

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

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



## MLE for simple regression (2)

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

- 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

## Summary / next

- We revisited three different (but equivalent) approaches to regression:
  - Best approximation to solving systems of linear equations
  - Minimizing sum of squared errors
  - MLE with Gaussian error
- Regression is the fundamental component of many ML methods: we will see similarities to regression in others

Next:

- Estimation, evaluation, bias, variance

