Regression: wrap up & MLE

Statistical Natural Language Processing 1

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

Winter Semester 2024/2025

version: 28a5d13 @2025-01-09

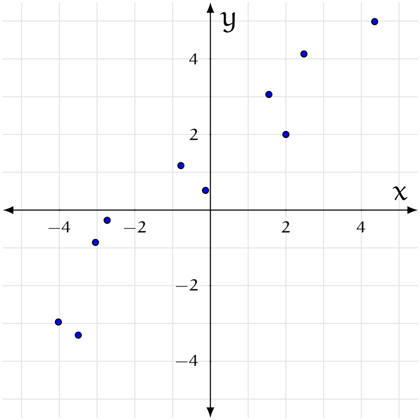
Linear regression

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



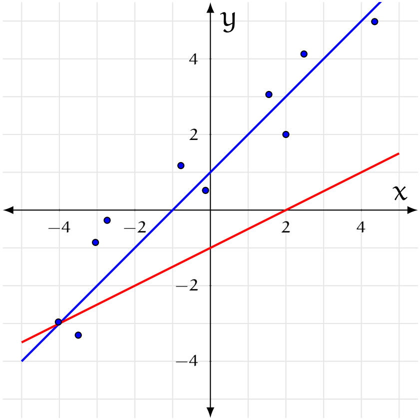
Linear regression

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



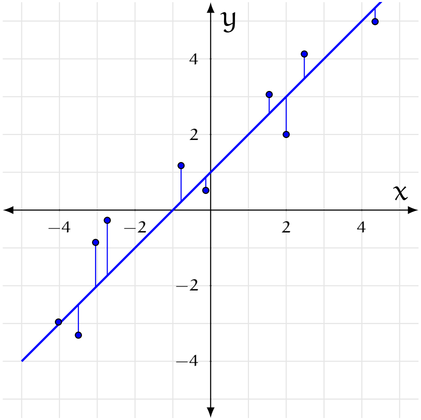
Linear regression

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



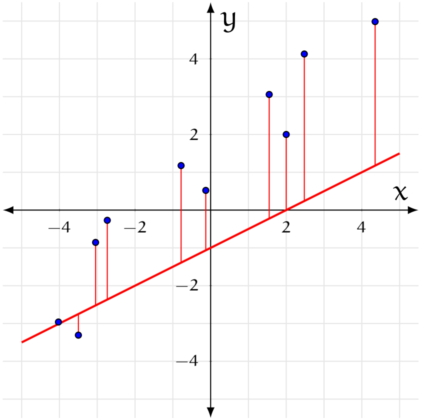
Linear regression

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

$$y = w_1 x + w_0$$

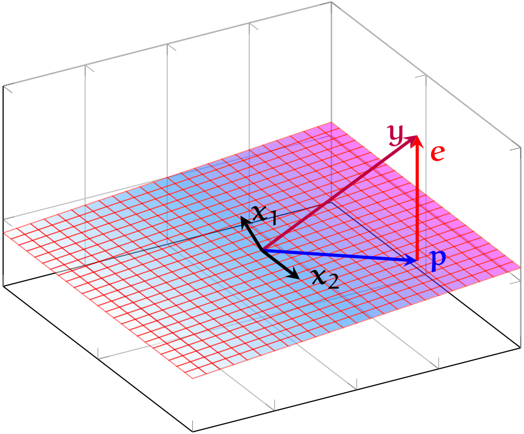
where,

- 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

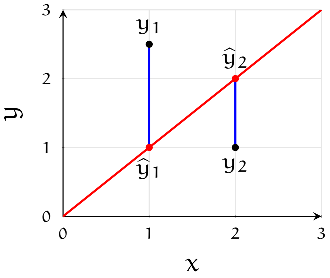
$$
\begin{aligned}
\mathbf{X}^T(\mathbf{y}-\mathbf{p}) &= 0 && \text{Error vector is orthogonal to columns} \\
\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{w}) &= 0 && \mathbf{p} \text{ is the weighted combination of columns} \\
\mathbf{X}^T\mathbf{X}\boldsymbol{w} &= \mathbf{X}^T\mathbf{y} && \text{Note: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

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

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

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?



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

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

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*

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 $\mathbf{x}$ such that $P(m|\mathbf{x})$ is the largest

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

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

Maximum Likelihood Estimation (MLE)

- In MLE the task is to find the model m that assigns the maximum ~~probability~~ *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

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

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

$$\boldsymbol{w}_{\text{MLE}} = \arg\max_{w} \log \mathcal{L}(w; \boldsymbol{x}) = \arg\min_{w} \log -\mathcal{L}(w; \boldsymbol{x})$$

MLE: simple example with coin flips

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

MLE: simple example with coin flips

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

$$p(x|p = 0.5) = 0.5^6(1 - 0.5)^4 = \frac{1}{1024} = 0.000\,977$$

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

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?

$$p(x|p = 0.5) = 0.5^6(1 - 0.5)^4 = \frac{1}{1024} = 0.000\,977$$

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

$$p(x|p = 0.4) = 0.4^6(1 - 0.4)^4 = \frac{1}{1024} = 0.000\,531$$

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

MLE: example with coin flips

finding the maximum likelihood

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

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

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

$$p_{MLE} = \underset{p}{\arg\max} \ln p^{n_H}(1-p)^{n_T} = \underset{p}{\arg\max}\, n_H \ln p + n_T \ln(1-p)$$

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

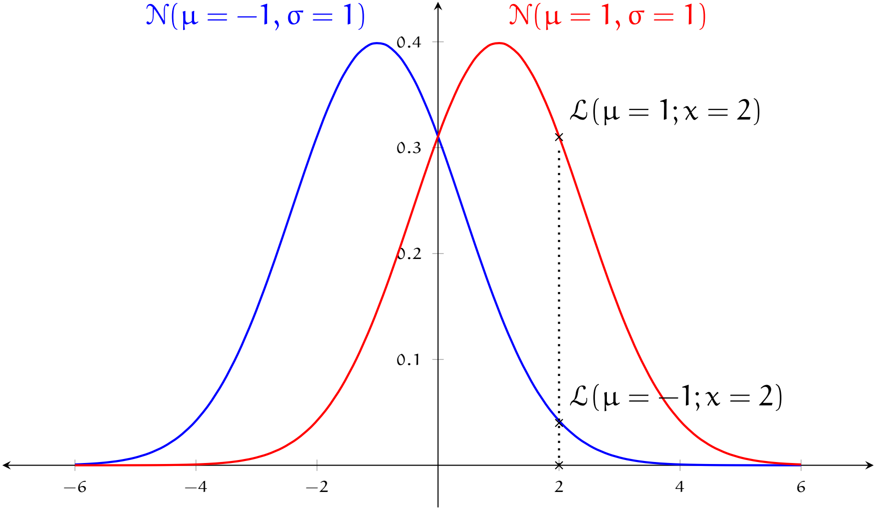
$$p_{MLE} = \arg\max_p \ln p^{n_H}(1-p)^{n_T} = \arg\max_p n_H \ln p + n_T \ln(1-p)$$

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

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

Another example: the mean of the Normal distribution

with known/equal variance



MLE for the parameters of Normal distribution

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

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

MLE for the parameters of Normal distribution

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

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

Log likelihood: $\mathcal{L}\mathcal{L}(\mu, \sigma; \boldsymbol{x}) = n \ln \frac{1}{\sqrt{2\pi}} + n \ln \frac{1}{\sigma} + \frac{1}{2\sigma^2} \sum_{i=0}^{n} (x - \mu)^2$

MLE for the parameters of Normal distribution

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

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

Log likelihood: $\mathcal{LL}(\mu, \sigma; \mathbf{x}) = n \ln \frac{1}{\sqrt{2\pi}} + n \ln \frac{1}{\sigma} + \frac{1}{2\sigma^2} \sum_{i=0}^{n} (x - \mu)^2$

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

MLE for the parameters of Normal distribution

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

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

Log likelihood: $\mathcal{LL}(\mu, \sigma; \mathbf{x}) = n \ln \frac{1}{\sqrt{2\pi}} + n \ln \frac{1}{\sigma} + \frac{1}{2\sigma^2} \sum_{i=0}^{n} (x - \mu)^2$

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

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

Properties of MLE

- In the limit ($n \to \infty$), 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*

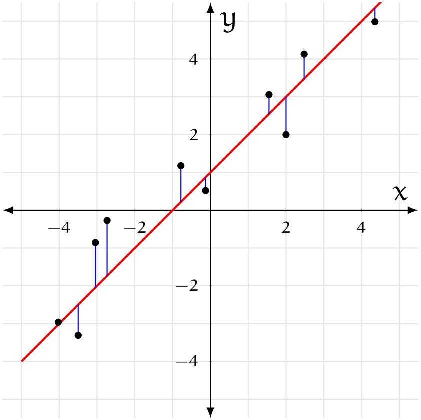
MLE for simple regression

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

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

- We additionally assume that 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)

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

- 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

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

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