Linear Models for Regression

Master's Degree in Bioinformatics and Computational Biology - Machine Learning

Carlos María Alaíz Gudín

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Academic Year 2024–25

Universidad Autónoma
de Madrid

Contents

Exercise ▷ **Questionnaire**

Please, fill in the questionnaire regarding your prior knowledge about this topic.

Introduction

Supervised Learning: Regression (I)

Definition (Supervised Learning)

Supervised learning is the Machine Learning (ML) task of learning a function that maps an input to an output based on example input–output pairs.

Definition (Regression Problem)

A **regression problem** is a supervised learning problem where the outputs are continuous.

Examples (Regression Problems)

- Predicting the wind energy production at a certain hour using Numerical Weather Predictions.
- Predicting the weight of a person based on the height, age, gender, etc.
- Predicting the future price of a stock based on its current value, the value of related stocks, the current trends, etc.

Supervised Learning: Regression (II)

Elements of a Supervised Learning Problem

Data Set of input–output pairs, $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$.

Features Vector of attributes (independent/input variables, covariates...), $\mathbf{x}_i \in \mathcal{X}$.

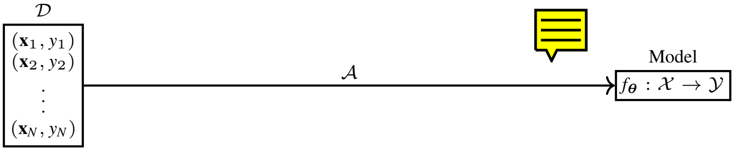
Target Label (dependent variable, outcome...), $y_i \in \mathcal{Y}$.

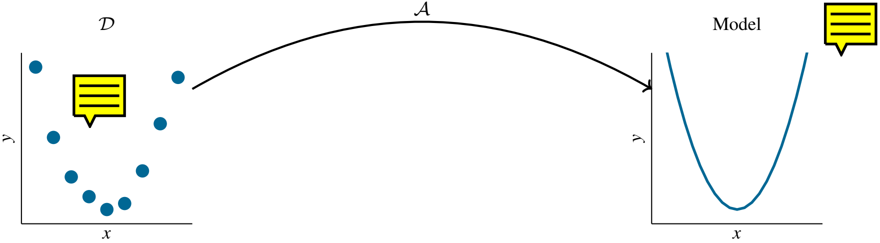
Model Mapping from the input to the output space, $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, with $\boldsymbol{\theta}$ the model parameters.

Learning Algorithm Procedure to obtain a model based on the data, $\mathcal{A} : \mathcal{D} \to f_{\boldsymbol{\theta}}(\cdot)$.

- In a regression setting usually $\mathcal{Y} = \mathbb{R}$.
- In many situations, specially after preprocessing the data, $\mathcal{X} = \mathbb{R}^d$.

Illustration





Linear Models for Regression

- "Simplest" approaches to regression:
  - Ignore the input: **constant model**.
  - Define the output as a linear combination of the inputs: **linear model**.

Advantages

- Simple.
- Robust (small variance).
- Interpretable.
- Easy to train.
- Easy to predict.

Disadvantages

- Limited flexibility.
- Under-fitting (large bias).

1-Dimensional Linear Regression

1-D Linear Model

- In the simplest case, $d = 1$ and $\mathcal{X} = \mathbb{R}$.
- The data becomes $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, with $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$.

- The corresponding linear model is simply a line, with parameters $\boldsymbol{\theta} = \{b, w\}$.
  - $b \in \mathbb{R}$ is the intercept or bias term.
  - $w \in \mathbb{R}$ is the slope of the line.
  - The model is defined as:

$$f_{\boldsymbol{\theta}}(x) = b + wx.$$

- The **learning algorithm** will determine b and w using $\mathcal{D}$.

1-D Linear Model - Exercise

Exercise

▷ **Questionnaire**

Given a 1-dimensional linear model with parameters $\boldsymbol{\theta} = \{b, w\}$, with $b = 1$ and $w = 2$.

1. Compute the output of the model for $x = 2$.
2. Compute the output of the model for $x = -1$.

Notebook

1-Dimensional Linear Regression: First Example

Quality of the Model

- A procedure is needed to determine the bias b and the slope w, optimizing the **quality** of the model.
- The quality of the model has to be defined. Usually from two points of view:

  Error An error term $\mathcal{E}_{\mathcal{D}}(\boldsymbol{\theta})$ measures how well the model fits the training data.

  Complexity A regularization term $\mathcal{R}(\boldsymbol{\theta})$ penalizes the complexity of the model.

Error Term for a 1-Dimensional Linear Model

Residual For the i-th pattern, $r_i = y_i - f_{\boldsymbol{\theta}}(x_i) = y_i - (b + wx_i)$.

Mean Squared Error (MSE) $\mathrm{MSE}(b, w) = \mathbb{E}\left[R^2\right] \approx \frac{1}{N}\sum_{i=1}^{N} r_i^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - (b + wx_i))^2$.

Mean Absolute Error (MAE) $\mathrm{MAE}(b, w) = \mathbb{E}[|R|] \approx \frac{1}{N}\sum_{i=1}^{N}|r_i| = \frac{1}{N}\sum_{i=1}^{N}|y_i - (b + wx_i)|$.

Quality of the 1-Dimensional Model - Exercise

Exercise

▷ **Questionnaire**

Given a 1-dimensional linear model with parameters $\boldsymbol{\theta} = \{b, w\}$, with $b = 1$ and $w = 2$, and for the following data:

x_i	y_i
2	4
-1	1

1. Compute the MAE.
2. Compute the MSE.

Notebook

1-Dimensional Linear Regression: Quality of the Model

Optimization

Definition (Optimization)

- **Optimize** (from Latin *optimus*, best) is to "make the best or most effective use of (a situation or resource)".
- Optimization is ubiquitous.
  - In nature.
  - In daily tasks.
  - As a strategy to design procedures.

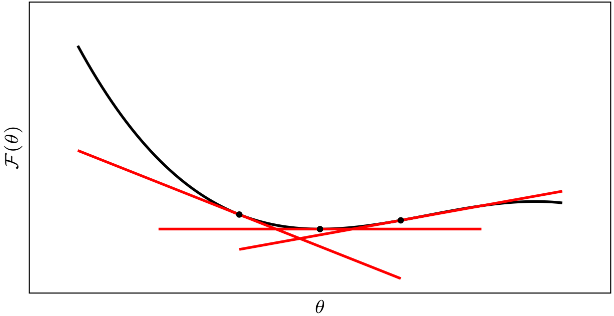
Formalization

- An optimization problem consists in finding the best element $\boldsymbol{\theta}^\star$ of a certain space $\mathcal{S}$ with respect to some criteria given by an objective function $\mathcal{F}$:

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta} \in \mathcal{S}} \{\mathcal{F}(\boldsymbol{\theta})\}.$$

Gradient-Based Optimization: 1-Dimensional Problems

- Given a 1-dimensional function $\mathcal{F}(\theta)$, its derivative $\mathcal{F}'(\theta)$ corresponds to the slope of the line which is tangent to the graph of $\mathcal{F}$ at θ.
  - If it is negative, $\mathcal{F}$ and its tangent go down.
  - If it is positive, $\mathcal{F}$ and its tangent go up.
  - If it is 0, the tangent is horizontal, hence there are two options:
    - $\mathcal{F}$ reaches a (local) minimum or maximum at θ.
    - $\mathcal{F}$ is flat at θ (plateau).



Training a 1-D Linear Model

- The most common choice for the error function is the MSE.
  - It is **differentiable**.
  - It corresponds to the **distance** between the vector of predictions and the vector of targets.
  - It is a natural choice when the observation noise is assumed to be **Gaussian**.

More detail in the appendix **Bayesian Perspective**.

- The learning algorithm for training the linear model consists in solving the problem:

$$\min_{b,w\in\mathbb{R}} \{\mathrm{MSE}(b,w)\} = \min_{b,w\in\mathbb{R}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - (b + wx_i))^2 \right\}.$$

- How is this problem solved?
  - It is **differentiable**: the optima are characterized by the zeros of the derivatives.
  - It is **convex**: there are no local minima.

Training a 1-D Linear Model: Optimization (I)

$$\min_{b,w\in\mathbb{R}} \{\mathrm{MSE}(b,w)\} = \min_{b,w\in\mathbb{R}} \left\{ \frac{1}{N}\sum_{i=1}^{N}(y_i - (b + wx_i))^2 \right\}.$$

$$\frac{\partial}{\partial b}\mathrm{MSE}(b,w)\Big|_{\substack{b=b^\star\\ w=w^\star}} = 0 \implies -\frac{2}{N}\sum_{i=1}^{N}(y_i - (b^\star + w^\star x_i)) = 0$$

$$\implies -\frac{2}{N}\sum_{i=1}^{N} y_i + \frac{2}{N}\sum_{i=1}^{N} b^\star + \frac{2}{N}\sum_{i=1}^{N} w^\star x_i = 0$$

$$\implies -\underbrace{\frac{1}{N}\sum_{i=1}^{N} y_i}_{\bar{y}} + b^\star + w^\star \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\bar{x}} = 0 \implies \boxed{b^\star = \bar{y} - w^\star \bar{x}}.$$

Training a 1-D Linear Model: Optimization (II)

$$\min_{b,w\in\mathbb{R}} \{\mathrm{MSE}(b,w)\} = \min_{b,w\in\mathbb{R}} \left\{ \frac{1}{N}\sum_{i=1}^{N}(y_i - (b + wx_i))^2 \right\}.$$

$$\frac{\partial}{\partial w}\mathrm{MSE}(b,w)\Big|_{\substack{b=b^\star\\ w=w^\star}} = 0 \implies -\frac{2}{N}\sum_{i=1}^{N} x_i(y_i - (b^\star + w^\star x_i)) = 0$$

$$\implies -\frac{2}{N}\sum_{i=1}^{N} x_i y_i + \frac{2}{N}\sum_{i=1}^{N} x_i b^\star + \frac{2}{N}\sum_{i=1}^{N} w^\star x_i^2 = 0$$

$$\overset{b^\star = \bar{y} - w^\star \bar{x}}{\implies} -\frac{2}{N}\sum_{i=1}^{N} x_i y_i + \frac{2}{N}\sum_{i=1}^{N} x_i \bar{y} - \frac{2}{N}\sum_{i=1}^{N} x_i w^\star \bar{x} + \frac{2}{N}\sum_{i=1}^{N} w^\star x_i^2 = 0$$

$$\implies -\sum_{i=1}^{N} x_i \underbrace{(y_i - \bar{y})}_{\hat{y}_i} + w^\star \sum_{i=1}^{N} x_i \underbrace{(x_i - \bar{x})}_{\hat{x}_i} \implies \boxed{w^\star = \frac{\sum_{i=1}^{N} x_i \hat{y}_i}{\sum_{i=1}^{N} x_i \hat{x}_i} = \frac{\sum_{i=1}^{N} \hat{x}_i \hat{y}_i}{\sum_{i=1}^{N} \hat{x}_i \hat{x}_i}}.$$

Training a 1-D Linear Model: Optimization (III)

- In summary, the Least Squares Regression Line is the solution of the following problem:

$$\min_{b,w \in \mathbb{R}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - (b + w x_i))^2 \right\}.$$

- These auxiliary elements are defined:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \text{ (Mean Target)}, \qquad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \text{ (Mean Feature)},$$

$$\hat{y}_i = y_i - \bar{y} \text{ (Centred Target)}, \qquad \hat{x}_i = x_i - \bar{x} \text{ (Centred Feature)}.$$

Least Squares Regression Line

$$w^\star = \frac{\sum_{i=1}^{N} \hat{x}_i \hat{y}_i}{\sum_{i=1}^{N} \hat{x}_i \hat{x}_i}; \quad b^\star = \bar{y} - w^\star \bar{x}.$$

Example in the appendix **Perfect 1-Dimensional Linear Model**.

Training a 1-Dimensional Linear Model - Exercise

Exercise ▷ **Questionnaire**

Given the following data:

x_i	y_i
0	4
1	6
2	4
3	6
4	8

1. Compute the value of $\bar{x}$ and $\bar{y}$.
2. Compute the value of $\hat{x}_i$ and $\hat{y}_i$.
3. Compute the value of $w^\star$ and $b^\star$.
4. Compute the corresponding MSE value.

Notebook

1-Dimensional Linear Regression: Optimization

Multiple Linear Regression

Linear Model

- For simplicity, $\mathcal{X} = \mathbb{R}^d$.
- The data becomes $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, with $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d}) \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.

- The corresponding linear model is a hyperplane, with parameters $\boldsymbol{\theta} = \{b, \mathbf{w}\}$.
  - $b \in \mathbb{R}$ is the intercept or bias term.
  - $\mathbf{w} = (w_1, w_2, \dots, w_d) \in \mathbb{R}^d$ is the normal vector of the hyperplane.
  - The model is defined as:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = b + \mathbf{w}^\intercal \mathbf{x} = b + \sum_{i=1}^{d} w_i x_i .$$

- The **learning algorithm** will determine b and $\mathbf{w}$ using $\mathcal{D}$.

Linear Model - Exercise

Exercise

▷ **Questionnaire**

Given a 2-dimensional linear model with parameters $\boldsymbol{\theta} = \{b, \mathbf{w}\}$, with $b = 1$ and $\mathbf{w} = (1, 2)^\intercal$.

1. Compute the output of the model for $\mathbf{x} = (1, 1)^\intercal$.
2. Compute the output of the model for $\mathbf{x} = (-1, 0)^\intercal$.

Notebook

Multiple Linear Regression: First Example

Linear Equations (I)

- A procedure is needed to determine the bias b and the vector $\mathbf{w}$.
- A first approach is to try to match all input–output pairs $(\mathbf{x}_i, y_i)$, $i = 1, \dots, N$. Specifically:

$$\begin{cases} b + \mathbf{w}^\intercal \mathbf{x}_1 = y_1 \\ b + \mathbf{w}^\intercal \mathbf{x}_2 = y_2 \\ \dots \\ b + \mathbf{w}^\intercal \mathbf{x}_N = y_N \end{cases} \equiv \begin{cases} b + w_1 x_{1,1} + w_2 x_{1,2} + \cdots + w_d x_{1,d} = y_1 \\ b + w_1 x_{2,1} + w_2 x_{2,2} + \cdots + w_d x_{2,d} = y_2 \\ \dots \\ b + w_1 x_{N,1} + w_2 x_{N,2} + \cdots + w_d x_{N,d} = y_N \end{cases}.$$

- The following matrix notation can simplify the equations:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,d} \end{pmatrix}; \; \tilde{\mathbf{X}} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,d} \\ 1 & x_{2,1} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \dots & x_{N,d} \end{pmatrix}; \; \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}; \; \tilde{\mathbf{w}} = \begin{pmatrix} b \\ w_1 \\ \vdots \\ w_d \end{pmatrix},$$

where $\mathbf{X} \in \mathbb{R}^{N \times d}$ is the data matrix, $\tilde{\mathbf{X}} \in \mathbb{R}^{N \times (d+1)}$ is the data matrix with a constant term, $\mathbf{y} \in \mathbb{R}^N$ is the vector of targets and $\tilde{\mathbf{w}} \in \mathbb{R}^{d+1}$ is the weight vector including the intercept.

Linear Equations (II)

- The system of equations becomes:

$$\tilde{\mathbf{X}}\tilde{\mathbf{w}} = \mathbf{y}.$$

- Since $\tilde{\mathbf{X}} \in \mathbb{R}^{N\times(d+1)}$, $\tilde{\mathbf{w}} \in \mathbb{R}^{d+1}$ and $\mathbf{y} \in \mathbb{R}^N$:
  - N equations.
  - $d+1$ unknowns.
- Usually, $N \gg d+1$ and the system is **overdetermined**.
- The inverse of $\tilde{\mathbf{X}}$ is not defined.

- The Moore-Penrose pseudo-inverse can be used instead, $\tilde{\mathbf{X}}^\dagger = \left(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\intercal$.
- A **different approach** also justifies this method.

Quality of the Model

- An alternative procedure is needed to determine the bias b and the vector $\mathbf{w}$.
- The solution is to optimize the **quality** of the model, probably not fitting exactly the training data.
- The quality of the model has to be defined. Usually from two points of view:
  - Error: An error term $\mathcal{E}_{\mathcal{D}}(\boldsymbol{\theta})$ measures how well the model fits the training data.
  - Complexity: A regularization term $\mathcal{R}(\boldsymbol{\theta})$ penalizes the complexity of the model.

Error Term for a Linear Model

- Residual: For the i-th pattern, $r_i = y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i - (b + \mathbf{w}^\intercal \mathbf{x}_i)$.
- Mean Squared Error (MSE): $\mathrm{MSE}(b, \mathbf{w}) = \mathbb{E}\left[R^2\right] \approx \frac{1}{N} \sum_{i=1}^{N} (y_i - (b + \mathbf{w}^\intercal \mathbf{x}_i))^2$.
- Mean Absolute Error (MAE): $\mathrm{MAE}(b, \mathbf{w}) = \mathbb{E}[|R|] \approx \frac{1}{N} \sum_{i=1}^{N} |y_i - (b + \mathbf{w}\mathbf{x}_i)|$.

Quality of the Model - Exercise

Exercise

▷ **Questionnaire**

Given a 2-dimensional linear model with parameters $\boldsymbol{\theta} = \{b, \mathbf{w}\}$, with $b = 1$ and $\mathbf{w} = (1, 2)^\intercal$, and for the following data:

$x_{i,1}$	$x_{i,2}$	y_i
1	1	4
-1	0	2

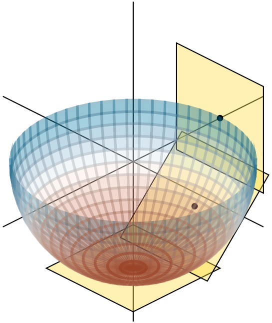
1. Compute the MAE.
2. Compute the MSE.

Gradient-Based Optimization: Multdimensional Problems

- In several dimensions, the derivative is generalized to the gradient (the vector of partial derivatives):

$$\nabla_{\boldsymbol{\theta}} \mathcal{F} = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \mathcal{F} & \frac{\partial}{\partial \theta_2} \mathcal{F} & \dots & \frac{\partial}{\partial \theta_d} \mathcal{F} \end{pmatrix}^\intercal .$$

  - The gradient defines the tangent hyperplane.
  - It points in the direction of greatest increase of $\mathcal{F}$.
  - If it is $\mathbf{0}$ at $\boldsymbol{\theta}$, then $\boldsymbol{\theta}$ is a stationary point.



Training a Linear Model

- The most common choice for the error function is the MSE.
  - It is **differentiable**.
  - It corresponds to the **distance** between the vector of predictions and the vector of targets.
  - It is a natural choice when the observation noise is assumed to be **Gaussian**.

More detail in the appendix **Bayesian Perspective**.

- The learning algorithm for training the linear model consists in solving the problem:

$$\min_{\substack{b \in \mathbb{R} \\ \mathbf{w} \in \mathbb{R}^d}} \{\mathrm{MSE}(b, \mathbf{w})\} = \min_{\substack{b \in \mathbb{R} \\ \mathbf{w} \in \mathbb{R}^d}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - (b + \mathbf{w}^\intercal \mathbf{x}_i))^2 \right\}.$$

- How is this problem solved?
  - It is **differentiable**: the optima are characterized by the zeros of the gradient.
  - It is **convex**: there are no local minima.

Training a Linear Model: Optimization (I)

$$\min_{\substack{b \in \mathbb{R} \\ \mathbf{w} \in \mathbb{R}^d}} \{\mathrm{MSE}(b, \mathbf{w})\} = \min_{\substack{b \in \mathbb{R} \\ \mathbf{w} \in \mathbb{R}^d}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - (b + \mathbf{w}\mathbf{x}_i))^2 \right\} \equiv \min_{\tilde{\mathbf{w}} \in \mathbb{R}^{d+1}} \left\{ \left(\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}\right)^\intercal \left(\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}\right) \right\}.$$

$$\begin{aligned}
\nabla_{\tilde{\mathbf{w}}} \mathrm{MSE}(\tilde{\mathbf{w}})\big|_{\tilde{\mathbf{w}} = \tilde{\mathbf{w}}^\star} = \mathbf{0} &\implies 2\tilde{\mathbf{X}}^\intercal \left(\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}^\star\right) = \mathbf{0} \\
&\implies \tilde{\mathbf{X}}^\intercal \mathbf{y} - \tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}}\tilde{\mathbf{w}}^\star = \mathbf{0} \\
&\implies \tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}}\tilde{\mathbf{w}}^\star = \tilde{\mathbf{X}}^\intercal \mathbf{y} \\
&\implies \boxed{\tilde{\mathbf{w}}^\star = \left(\tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^\intercal \mathbf{y} = \tilde{\mathbf{X}}^\dagger \mathbf{y}}.
\end{aligned}$$

Training a Linear Model: Optimization (II)

- In summary, the Least Squares Linear Model is the solution of the following problem:

$$\min_{\substack{b\in\mathbb{R}\\ \mathbf{w}\in\mathbb{R}^d}} \left\{ \frac{1}{N}\sum_{i=1}^{N}\left(y_i - (b + \mathbf{w}^\intercal \mathbf{x}_i)\right)^2 \right\}.$$

Least Squares Linear Model

$$\begin{pmatrix} b^\star \\ \mathbf{w}^\star \end{pmatrix} = \tilde{\mathbf{w}}^\star = \tilde{\mathbf{X}}^\dagger \mathbf{y} = \begin{bmatrix} \mathbf{1} & \mathbf{X} \end{bmatrix}^\dagger \mathbf{y}.$$

Example in the appendix **Perfect Linear Model**.

Notebook

Multiple Linear Regression: Optimization

Summary

Linear Models for Regression: Summary

- A **regression** problem is a supervised problem with continuous targets.

- A simple yet useful regression model is the **linear model**.
  - The prediction is a linear combination of the features.

- In order to train the linear model, an **optimization problem** is usually solved.

- The **MSE** is often used to measure the quality of the model.
  - It is a natural choice.
  - The resultant problem can be solved in closed-form using the pseudo-inverse of the data matrix.

Linear Models for Regression

Carlos María Alaíz Gudín

Additional Material

Training a 1-Dimensional Linear Model - Example

Example (Perfect Case)

- In the perfectly linear case, $y_i = wx_i + b$.
- This implies:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}(wx_i + b) = w\bar{x} + b,$$
$$\hat{y}_i = y_i - \bar{y} = wx_i + b - w\bar{x} - b = w(x_i - \bar{x}) = w\hat{x}_i.$$

- Therefore, the regression lines becomes:

$$w^\star = \frac{\sum_{i=1}^{N} \hat{x}_i \hat{y}_i}{\sum_{i=1}^{N} \hat{x}_i \hat{x}_i} = \frac{w\sum_{i=1}^{N} \hat{x}_i^2}{\sum_{i=1}^{N} \hat{x}_i^2} = w;$$
$$b^\star = \bar{y} - w^\star \bar{x} = w\bar{x} + b - w\bar{x} = b.$$

Training a Linear Model - Example

Example (Perfect Case)

- In the perfectly linear case, $y_i = \mathbf{w}^\intercal \mathbf{x}_i + b$.
- In matrix notation, $\mathbf{y} = \tilde{\mathbf{X}}\tilde{\mathbf{w}}$.
- Therefore, the linear model becomes:

$$\begin{aligned}
\tilde{\mathbf{w}}^\star &= \tilde{\mathbf{X}}^\dagger \mathbf{y} \\
&= \left(\tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^\intercal \mathbf{y} \\
&= \left(\tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^\intercal \left(\tilde{\mathbf{X}}\tilde{\mathbf{w}}\right) \\
&= \left(\tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}}\right)^{-1} \left(\tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}}\right) \tilde{\mathbf{w}} \\
&= \tilde{\mathbf{w}}.
\end{aligned}$$

Training a Linear Model: Bayesian Perspective (I)

- There is an additional justification for using the MSE in a linear model.
- The output is assumed to be a linear transformation of the input corrupted with Gaussian noise:

$$y_i = \mathbf{w}^\intercal \mathbf{x}_i + \epsilon_i,$$

  where $\epsilon_i \sim \mathcal{N}(0, \sigma)$.
- The likelihood of the data becomes:

$$\mathrm{p}(\mathcal{D}|\mathbf{w}) \propto \prod_{i=1}^{N} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \prod_{i=1}^{N} \exp\left(-\frac{(y_i - \mathbf{w}^\intercal \mathbf{x}_i)^2}{2\sigma^2}\right).$$

- $\mathbf{w}^\star \in \mathbb{R}^d$ is selected as the maximizer of the likelihood:

$$\max_{\mathbf{w} \in \mathbb{R}^d} \left\{ \prod_{i=1}^{N} \mathrm{p}(\mathcal{D}|\mathbf{w}) \right\} = \max_{\mathbf{w} \in \mathbb{R}^d} \left\{ \prod_{i=1}^{N} \exp\left(-\frac{(y_i - \mathbf{w}^\intercal \mathbf{x}_i)^2}{2\sigma^2}\right) \right\}.$$

Training a Linear Model: Bayesian Perspective (II)

- Equivalently, instead of maximizing the likelihood, the minus log-likelihood is minimized:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \sum_{i=1}^{N} (y_i - \mathbf{w}^\intercal \mathbf{x}_i)^2 \right\},$$

which coincides with the least squares problem for a linear model.

- Bayesian Linear Regression is more than this.
- The **prior** can be used to impose structure, use prior knowledge, etc.