A simple ML model

MACHINE LEARNING

Dr. Temitayo Olugbade

UNIVERSITY OF SUSSEX

Learning outcome

After working through this mini-video,

- you'll see how one of the simplest machine learning models (linear regression) works.

Lecture outline

- ❑ The basic linear model

- ❑ How learning happens in ML

Recall from Week 1b mini-video

1. The most basic element of **machine learning** is a **model** that learns from data.
2. With **supervised learning**, data has both **features/input** $\boldsymbol{x}$ & **labels/output** $\boldsymbol{y}$.
3. When the label/output are real valued or continuous, it is a **regression** task. Otherwise, it is **classification**.
4. An important goal in machine learning is **generalizability** to data not seen by the model during its training.

The basic linear model

- ❑ **The basic linear model**
- ❑ How learning happens in ML

Basic linear model

is a function $f(\boldsymbol{x})$ defined as

$$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{w} + b = \hat{\boldsymbol{y}}$$

where

- $\boldsymbol{x}$ – features (i.e. model input)
- $\hat{\boldsymbol{y}}$ – real/continuous-valued predicted labels (i.e. model output)
- $\boldsymbol{w}$, b – weights, bias (these are model parameters)

Basic linear model

implies 'supervised learning'

is a function $f(\boldsymbol{x})$ defined as

$$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{w} + b = \hat{\boldsymbol{y}}$$

implies 'regression'

where

- $\boldsymbol{x}$ – features (i.e. model input)
- $\hat{\boldsymbol{y}}$ – real/continuous-valued predicted labels (i.e. model output)
- $\boldsymbol{w}$, b – weights, bias (these are model parameters)

Applying the model to toy data

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w} + b$$
$$= x_1 w_1 + x_2 w_2 + \cdots + x_D w_D + b$$

see alternative math expressions in last slide pages

Applying the model to toy data

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w} + b = x_1 w_1 + x_2 w_2 + \cdots + x_D w_D + b$$

Temperature (°C)	Relative humidity (%)	Wind speed (km/h)	Rain (mm)	Fire weather index
23	21	10	0	0
40	89	6	1	30
35	60	23	15	15

Applying the model to toy data

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w} + b = x_1 w_1 + x_2 w_2 + \cdots + x_D w_D + b$$

Temperature (°C)	**Relative humidity (%)**	**Wind speed (km/h)**	**Rain (mm)**	**Fire weather index**
23	21	10	0	0
40	89	6	1	30
35	60	23	15	15

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_{N=3} \end{bmatrix} = \begin{bmatrix} 23 & 21 & 10 & 0 \\ 40 & 89 & 6 & 1 \\ 35 & 60 & 23 & 15 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_D \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix}$$

Applying the model to toy data

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w} + b = x_1 w_1 + x_2 w_2 + \cdots + x_D w_D + b$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_{N=3} \end{bmatrix} = \begin{bmatrix} 23 & 21 & 10 & 0 \\ 40 & 89 & 6 & 1 \\ 35 & 60 & 23 & 15 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix}$$

$$\Rightarrow \hat{y}_1 = 23w_1 + 21w_2 + 10w_3 + 0w_4 + b$$

$$\hat{y}_2 = 40w_1 + 89w_2 + 6w_3 + 1w_4 + b$$

$$\hat{y}_3 = 35w_1 + 60w_2 + 23w_3 + 15w_4 + b$$

Applying the model to toy data

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w} + b = x_1 w_1 + x_2 w_2 + \cdots + x_D w_D + b$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_{N=3} \end{bmatrix} = \begin{bmatrix} 23 & 21 & 10 & 0 \\ 40 & 89 & 6 & 1 \\ 35 & 60 & 23 & 15 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_D \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix}$$

$$\Rightarrow \hat{y}_1 = 23w_1 + 21w_2 + 10w_3 + 0w_4 + b$$

$$\hat{y}_2 = 40w_1 + 89w_2 + 6w_3 + 1w_4 + b$$

$$\hat{y}_3 = 35w_1 + 60w_2 + 23w_3 + 15w_4 + b$$

but what are the best values of $\boldsymbol{w}$ and b, i.e. the optimal model parameters?

How learning happens in ML

- ❑ The basic linear model
- ❑ **How learning happens in ML**

Learning the optimal model parameters

- The optimal model parameters would minimize the error between the model prediction $\hat{\boldsymbol{y}}$ and the true label $\boldsymbol{y}$

- Since $\boldsymbol{y}$ is real/continuous-valued, one possible way to measure the model error is mean-squared error L_2, i.e.

$$L_2 = \frac{1}{N} \sum_{n=1}^{N} (\hat{\boldsymbol{y}}_n - \boldsymbol{y}_n)^2$$

Learning the optimal model parameters

- The optimal model parameters would minimize the error between the model prediction $\widehat{\boldsymbol{y}}$ and the true label $\boldsymbol{y}$

- Since $\boldsymbol{y}$ is real/continuous-valued, one possible way to measure the model error is **mean-squared error** L_2, i.e.

squared

$$L_2 = \frac{1}{N}\sum_{n=1}^{N}(\widehat{\boldsymbol{y}}_n - \boldsymbol{y}_n)^2$$

difference (aka error)

mean

Learning the optimal model parameters

- The optimal model parameters would minimize the error between the model prediction $\hat{\boldsymbol{y}}$ and the true label $\boldsymbol{y}$

- Since $\boldsymbol{y}$ is real/continuous-valued, one possible way to measure the model error is mean-squared error L_2, i.e.

$$L_2 = \frac{1}{N}\sum_{n=1}^{N}(\hat{\boldsymbol{y}}_n - \boldsymbol{y}_n)^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n\boldsymbol{w} + b - \boldsymbol{y}_n)^2$$

Minimizing the error

- Error $L_2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n\boldsymbol{w} + b - \boldsymbol{y}_n)^2$

- From math principles, the minimum of a function is when its gradient (derivative) is zero, so

$$0 = \frac{dL_2}{d\boldsymbol{w}}$$

- When you expand, apply the derivative, and make $\boldsymbol{w}$ the subject of the equation

$$\boldsymbol{w} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$$

see math details (i.e. proof) in last slide pages

Minimizing the error

- Error $L_2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n\boldsymbol{w} + b - \boldsymbol{y}_n)^2$

- From math principles, the minimum of a function is when its gradient (derivative) is zero, so

$$0 = \frac{dL_2}{d\boldsymbol{w}}$$

- When you expand, apply the derivative, and make $\boldsymbol{w}$ the subject of the equation

inverse

transpose

$$\boldsymbol{w} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$$

ML with the basic linear model

- Get training data, i.e. $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ pairs, $1 \leq n \leq N$

- Choose an error metric, e.g. mean-squared error L_2

- Find the optimal model parameters, i.e. the best values for $\boldsymbol{w}^*$ and b^*

- Plug this in your model and apply to obtain $\hat{\boldsymbol{y}}$

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w}^* + b^*$$

ML with the basic linear model

- Get training data, i.e. $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ pairs, $1 \leq n \leq N$

loss function

- Choose an error metric, e.g. mean-squared error L_2

- Find the optimal model parameters, i.e. the best values for $\boldsymbol{w}^*$ and b^*

model training

inference

- Plug this in your model and apply to obtain $\hat{\boldsymbol{y}}$

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w}^* + b^*$$

trained model

Other loss function – Mean absolute error

- Mean absolute error L_1

$$L_1 = \frac{1}{N}\sum_{n=1}^{N} |\widehat{\boldsymbol{y}}_n - \boldsymbol{y}_n|$$

- Demerit

  Its gradient $\frac{dL_1(\boldsymbol{w})}{d\boldsymbol{w}}$ is a constant and not a function of $\boldsymbol{w}$, so the optimal $\boldsymbol{w}$ can't be obtained as easily

see math details (i.e. proof) in last slide pages

- Merit

  L_1 is less influenced by outliers

Summary

1. The basic linear ML model is $f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w} + b$. It is a regression model. This makes this a **linear regression** model.

2. Training a ML model involves optimizing the model weights based on a **loss function**.

Any questions???

scan the QR code to ask questions

Math details and proofs

Basic linear model reframed

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w} + b$$

- In matrix form

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{ND} \end{bmatrix} \times \begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}$$

- Rewriting to absorb b in $\boldsymbol{w}$

$$f(\boldsymbol{x}) = \hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{w}$$

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1D} & 1 \\ \vdots & \ddots & \vdots & \\ x_{N1} & \cdots & x_{ND} & 1 \end{bmatrix} \times \begin{bmatrix} w_1 \\ \vdots \\ w_D \\ b \end{bmatrix}$$

Minimizing the error – MATH DETAIL (1)

$$0 = \frac{d\mathrm{L}_2(\mathbf{w})}{d\boldsymbol{w}}$$

if you substitute for $\mathrm{L}_2(\mathbf{w})$ *with its value*

$$0 = \frac{1}{N} \times \frac{d\left(\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n\boldsymbol{w} + b - \boldsymbol{y}_n)^2\right)}{d\boldsymbol{w}}$$

a neater (shorthand) way to write this is

$$0 = \frac{1}{N} \times \frac{d(\|\boldsymbol{x}\boldsymbol{w} - \boldsymbol{y}\|^2)}{d\boldsymbol{w}}$$

when you expand the numerator of the right hand side

$$0 = \frac{d\left((\boldsymbol{x}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{x}\boldsymbol{w} - \boldsymbol{y})\right)}{d\boldsymbol{w}}$$

Minimizing the error – MATH DETAIL (2)

$$0 = \frac{d\left((\boldsymbol{xw} - \boldsymbol{y})^T(\boldsymbol{xw} - \boldsymbol{y})\right)}{d\boldsymbol{w}}$$

when you further expand and collect like terms

$$0 = \frac{d\left(\boldsymbol{w}^T\boldsymbol{x}^T\boldsymbol{xw} - 2\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}\right)}{d\boldsymbol{w}}$$

*when you apply the derivative with respect to **w** to the right hand side*

$$0 = 2\boldsymbol{x}^T\boldsymbol{xw} - 2\boldsymbol{x}^T\boldsymbol{y}$$

*when you make **w** the subject of the formula*

$$\boldsymbol{w} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$$

Derivative of the L1 loss

$$\frac{d\mathrm{L}_1(\mathbf{w})}{d\boldsymbol{w}} = \frac{1}{N} \times \frac{d(|\boldsymbol{x}\boldsymbol{w} - \boldsymbol{y}|)}{d\boldsymbol{w}}$$

$$\frac{d\mathrm{L}_1(\mathbf{w})}{d\boldsymbol{w}} = \frac{1}{N} \times \frac{d(\boldsymbol{x}\boldsymbol{w} - \boldsymbol{y})}{d\boldsymbol{w}}$$

$$\frac{d\mathrm{L}_1(\mathbf{w})}{d\boldsymbol{w}} = \frac{1}{N} \times \boldsymbol{x}$$

L1 loss gradient is a constant!

→ optimal $\boldsymbol{w}$ cannot be obtained analytically