Supervised Machine Learning and Learning Theory

Lecture 2: Linear Regression, with Some Review of Linear Algebra

September 10, 2024

In-class quiz questions

- Given a data distribution D, a neural network f_W whose parameters are given by W, write down the mathematical definition of the test loss of f_W?

- Given n samples from D, denoted as $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, write down the mathematical definition of the training loss of f_W?

- What is representation learning? Could you name several methods for representation learning?

Matrices and vectors

- Matrices: A rectangular array of numbers

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \cdots & \cdots & \cdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}$$

- Vectors: An array consisting of a single column

$$a = \begin{bmatrix} a_1 \\ \cdots \\ a_n \end{bmatrix}$$

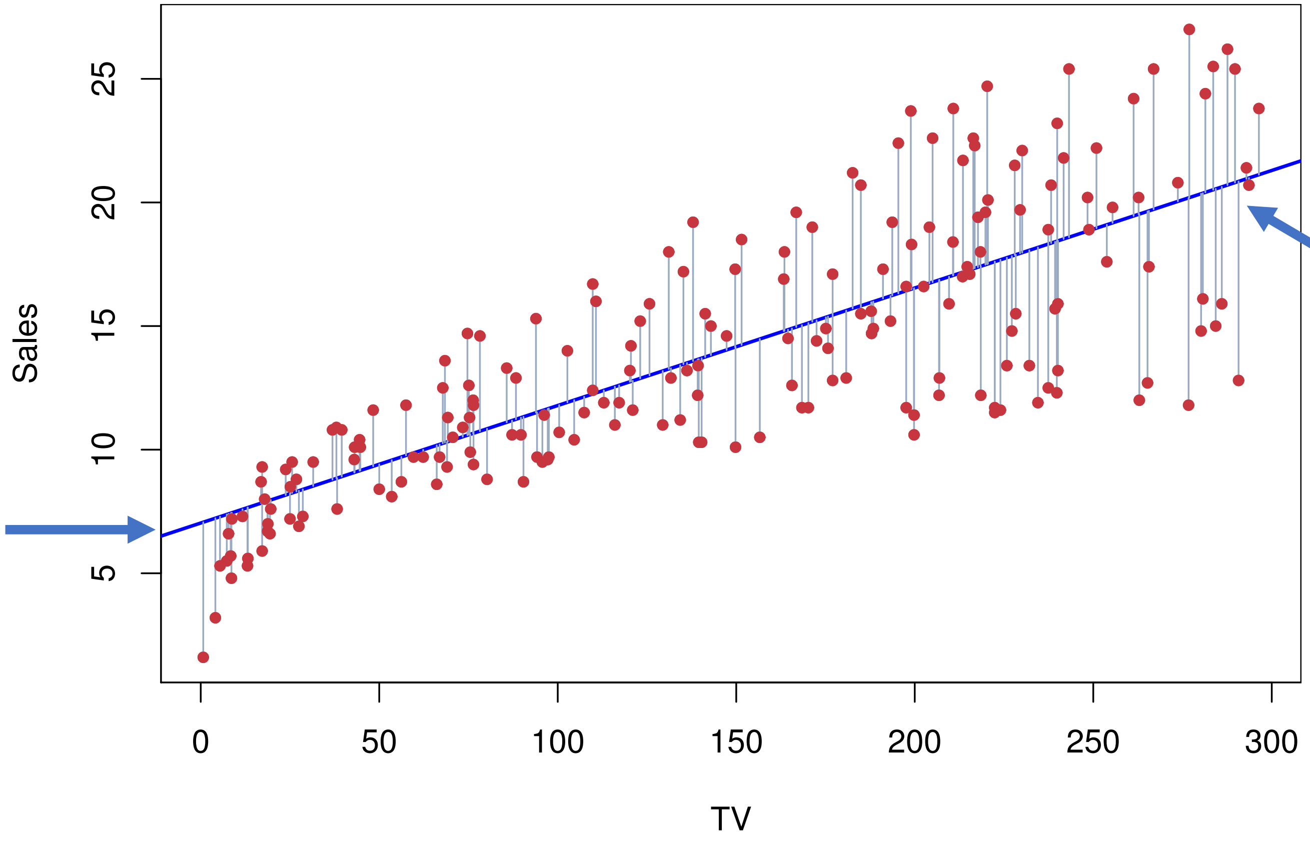
Simple linear regression

- Let us consider the simplest case of a linear regression problem: We are giving a list of one-dimensional features and their corresponding labels. We want to build a regression model to achieve that
  - Examples: Predicting housing values (last Friday), advertising, marketing, etc

- Input: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ (assume we have already done the training/test split)
- Output: a linear model parameterized by β_0 and β_1

Examples of β_0 and β_1

- Fitting a regression model mapping TA ad spending to Sales amount



β_0: Intercept

β_1: Slope

Setting up the linear model

- Recall the input to the problem: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ (this is the training data)
- Let us set up a predicted label for each sample:

$$\hat{y}_i = \beta_0 + x_i\beta_1, \text{ for } i = 1, 2, \ldots, n$$

- Next, let us set up the mean squared error metric:

$$\hat{L}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + x_i\beta_1 - y_i)^2$$

Where $\beta = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}$

Solving for β_0 and β_1

- Recall that $\hat{L}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + x_i\beta_1 - y_i)^2$; we would like to minimize the MSE metric
- We're going to set the derivatives of $\hat{L}$ with respect to β_0, β_1 as zero

$$\frac{\partial \hat{L}(\beta)}{\partial \beta_0} = \frac{2}{n}\sum_{i=1}^{n}(\beta_0 + x_i\beta_1 - y_i) = 0$$

$$\frac{\partial \hat{L}(\beta)}{\partial \beta_1} = \frac{2}{n}\sum_{i=1}^{n}x_i(\beta_0 + x_i\beta_1 - y_i) = 0$$

Solving for β_0 and β_1

- We can re-arrange the derivatives to be zero as follows

$$\beta_0 + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\beta_1 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)\beta_1 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Final solution

- This is a two-by-two linear system, which can be solved explicitly

$$\beta_0 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

$$\beta_1 = \frac{\left(1 - \frac{1}{n}\sum_{i=1}^{n} x_i\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

Takeaways

- In order to have a valid solution, we need that

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \neq 0$$

This is true as long as the x_i's are not all the same!

- We can use the explicit expressions of β_0, β_1 to derive confidence intervals
  - This is a bit advanced, but the high-level idea is we assume the x_i's are Gaussian, from which we could derive the distribution of β_0, β_1

Summary of simple linear regression

- After solving $\hat{\beta}_0, \hat{\beta}_1$, we could use the estimated coefficients to make predictions on unseen regions

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

Evaluation metrics

- **R^2 statistic** measures the proportion of variance explained

$$\text{RSS (Residual sum of squares)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{TSS (Total sum of squares)} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 always takes on a value between 0 and 1

Evaluation metrics

- **Correlation** between two random variables is another measure of linear relationship between X and Y

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- **Example:** in the linear regression example, we may take the uniform distribution of $y_1, y_2, \dots, y_n$ as the 1st random variable, and the uniform distribution of $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ as the 2nd random variable

- **Example**: If X and Y are independent, then $Cor(X,Y) = 0$
    - Recall $E[X \cdot Y] = E[X] \cdot E[Y]$

Lecture plan

- **Multiple linear regression**

Multiple linear regression

- Multiple features
- Quantitative inputs
- Transformations of quantitative inputs: log, square-root, or square
- Basis expansion: $x_2 = x_1^2, x_3 = x_1^3$
- Numeric coding of qualitative inputs
- Interactions between inputs: $x_3 = x_1 \cdot x_2$

Setting up the problem

- We're giving a training set $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Let us assume that each x has p features in total

- We want to learn a linear regression model to map x's to y's: the linear model has $p + 1$ variables in total, $\beta_0, \beta_1, \dots, \beta_p$

Let us introduce several matrix notations

- Feature matrix (note that we have added a column of ones):

$$X = \begin{bmatrix} 1 & x_{1,1}, \dots, x_{1,p} \\ 1 & x_{2,1}, \dots, x_{2,p} \\ & \vdots \quad \vdots \\ 1 & x_{n,1}, \dots, x_{n,p} \end{bmatrix}$$

- Label vector:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Exercise: what is the dimension of $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}$, respectively?

- Predicted label:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}, \text{ for } i = 1,2, \dots, n$$

More matrix notations

- Let us stack the variables we need to estimate together

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{bmatrix}$$

- Using matrix multiplication rule, we shall verify that

$$\hat{y} = X\beta$$

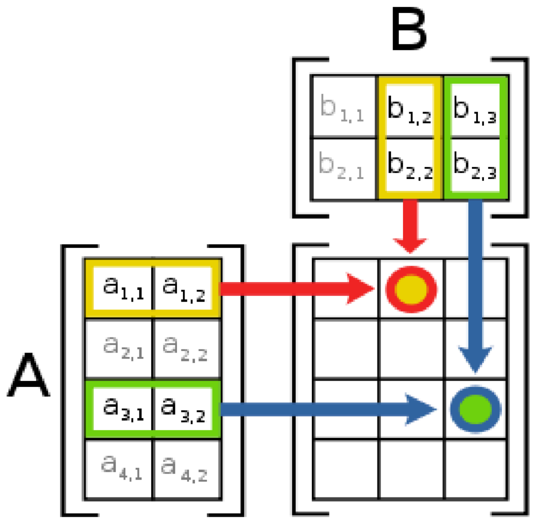
Where $\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$

One slide about matrix multiplication

- Let $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, their product $C = AB \in \mathbb{R}^{m \times p}$
- Number of columns of A must be equal to the number of rows of B
- Compute the product $C = AB$ using

$$C_{i,j} = \sum_{k=1}^{n} A_{i,k} B_{k,j}$$

- An illustration



- Exercise: multiply $A = [1,2]$ with $B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

Start with the one-dimensional case

- **Fitting a line** with coefficient $\beta_1 \in \mathbb{R}$ and intercept $\beta_0 \in \mathbb{R}$

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- **Recall matrix notation:** $\hat{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

- **Exercise:** verify that $\hat{y} = X\beta$

Move to the multi-dimensional case

- **Fitting a hyperplane** with coefficients $\beta_1, \beta_2, \ldots, \beta_p$ and intercept β_0
- **Exercise:** First verify that the predicted labels are $\hat{y} = X\beta$
- Recall that MSE metric:

$$\hat{\boldsymbol{L}}(\boldsymbol{\beta}) = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{x}_{\boldsymbol{i}}^{\top}\boldsymbol{\beta} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}} = \frac{\mathbf{1}}{\boldsymbol{n}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

- We'll set the derivatives to zero: $\frac{\partial \hat{L}(\beta)}{\partial \beta_0}, \frac{\partial \hat{L}(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \hat{L}(\beta)}{\partial \beta_p}$
- There's an easier way to write this in the multi-dimensional case

Defining the gradient

- **Definition:** let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be a multi-dimensional function, which takes a vector of d variables X as input, and outputs a real value $y = f(X)$

- Suppose f is differentiable at every coordinate, then, the gradient of f, denoted as ∇f, is defined as

$$\nabla f(X) = \begin{bmatrix} \frac{\partial f(X)}{\partial X_1}, \\ \frac{\partial f(X)}{\partial X_2}, \\ \cdots, \\ \frac{\partial f(X)}{\partial X_d} \end{bmatrix}$$

Back to estimating the coefficients

- The condition for setting all of the derivatives of $\hat{L}(\beta)$ to zero amounts to the following

$$\nabla\hat{L}(\beta) = 0$$

- **Claim:**

$$\nabla\hat{L}(\beta) = \frac{2}{n}X^\top(X\beta - y)$$

  - **Exercise:** Verify the dimension of the right-hand side
  - Now, we want to set the gradient as zero
  - This means we have $X^\top(X\beta - y) = 0$
  - This leads to the following equation for β

$$\boldsymbol{\hat{\beta} = (X^T X)^{-1} X^T y}$$

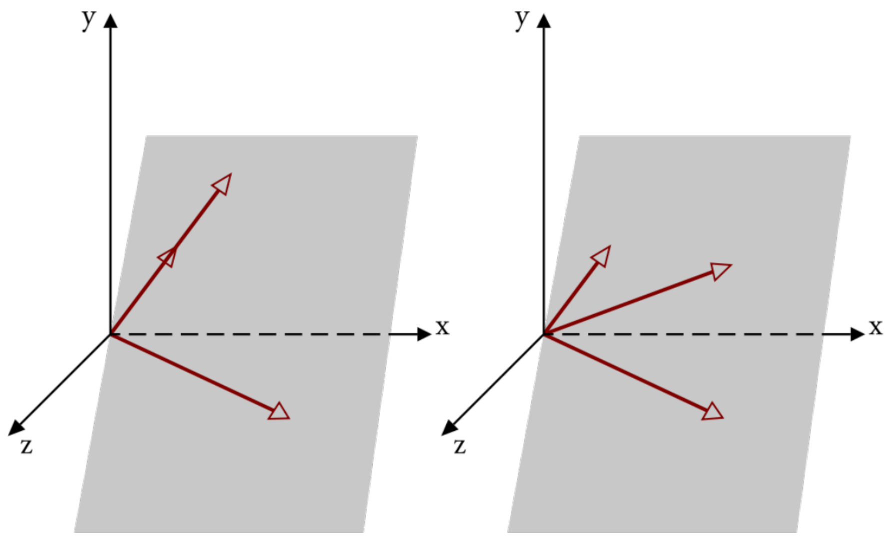
This is called the Ordinary Least Squares (OLS) estimator

Takeaways

- We want $X^\top X$ to be invertible (what does it mean?)
- Let's first explain linear combinations: Given a set of vectors $S = \{x_1, \dots, x_n\}$ where $x_i \in \mathbb{R}^n$, a **linear combination** of S is

$$\sum_{i=1}^{n} \boldsymbol{a_i x_i} \text{ where } \boldsymbol{a_i} \in \mathbb{R}$$

  - The **vector span** of S, denoted as $\mathrm{Span}(S)$, is the set of all **linear combinations** of the elements of S



Linearly independent vs. not linearly independent

- A set of vectors $S = \{x_1, x_2, \dots, x_n\}$ is **linearly independent** if the following holds

$$\sum_{i=1}^{n} \boldsymbol{a_i x_i} = \mathbf{0} \textbf{ if and only if } \boldsymbol{a_1} = \boldsymbol{a_2} = \cdots = \boldsymbol{a_n} = \mathbf{0}$$

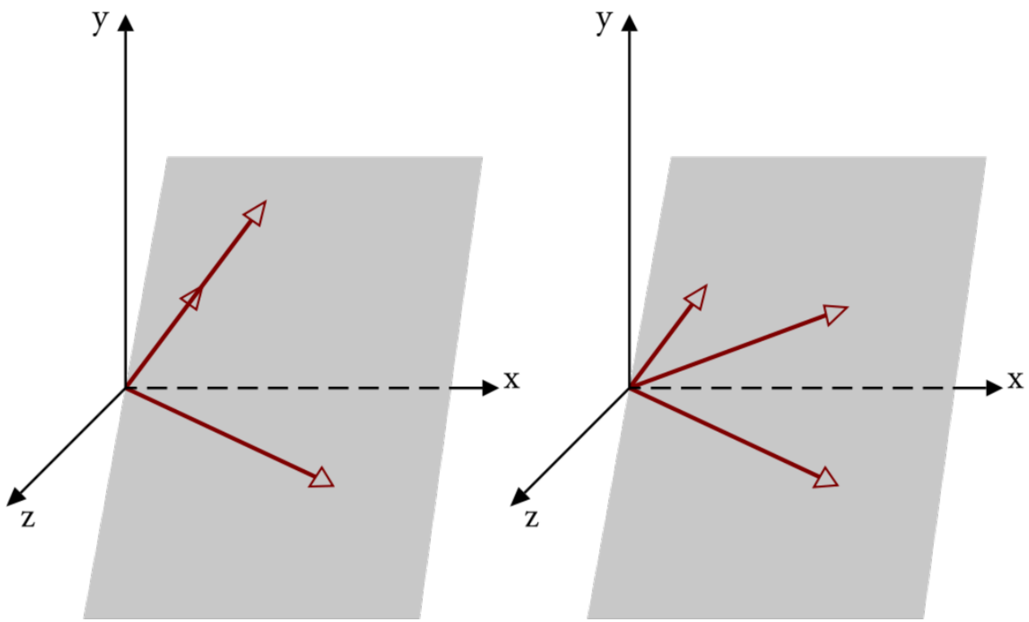
- On the other hand, S is **not linearly independent** if there exists $\boldsymbol{a_1, a_2, \dots, a_n}$ **that are not all zeros** such that

$$\sum_{i=1}^{n} \boldsymbol{a_i x_i} = \mathbf{0}$$

- **Back to the previous example, which one is linearly independent and which one is not?**

Examples of linearly independent vectors

- **Left:** The two vectors are **linearly independent**
- **Right:** The three vectors are **not linearly independent**



Rank

- **Rank:** For $A \in \mathbb{R}^{m \times n}$, the rank of A is the **maximum** number of linearly independent columns or rows

- **Exercises (after class)**

$$rank(A) \leq \min(m, n)$$
$$rank(A) = rank(A^\top)$$
$$rank(AB) \leq \min(rank(A), rank(B))$$
$$rank(A + B) \leq rank(A) + rank(B)$$

Metrics

- Mean squared error (MSE) is the average amount that the response will deviate from the true regression line

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Normalized MSE: Divide MSE by $\frac{1}{n}\sum_{i=1}^{n} y_i^2$
- Root mean squared error: $\text{RMSE} = \sqrt{\text{MSE}}$
  - RMSE measures the average deviation between $\hat{y}_i$ and y_i
- $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$
  - $\hat{y}_i$ is the fitted y_i, for example, in the linear model, $\hat{y}_i = \hat{\beta}_0 + x_i \cdot \hat{\beta}_1$
  - More generally, let $\hat{f}$ be the fitted function (e.g., quadratic), and then $\hat{y}_i = \hat{f}(x_i)$
  - $0 \leq R^2 \leq 1$

Setting confidence intervals

- Are the estimated coefficients statistically significant?
- **Construct confidence intervals:** With 95% probability, the range will contain the true value of the parameter

$$\beta_0 \in [\hat{\beta}_0 - 2 \cdot \mathrm{SE}(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot \mathrm{SE}(\hat{\beta}_0)]$$

$$\cdots$$

$$\beta_p \in [\hat{\beta}_p - 2 \cdot \mathrm{SE}(\hat{\beta}_p), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_p)]$$

Statsmodel package provides estimated coefficients and standard errors

https://www.statsmodels.org/stable/index.html

Hypothesis testing and significance values

- Null hypothesis: $\beta_1 = 0$, there is no relationship between X and Y
- Expected outcome: $\beta_1 \neq 0$, there is relationship between X and Y

- **T-statistic**: number of standard errors between $\hat{\beta}_1$ and 0

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$$

- ***p*-value**: probability of observing at least $|t|$ under null hypothesis

Announcements

- **Office hour:** 12:30 PM – 1:30 PM, 177 Huntington Ave FL 22, Room 2211
  - Also accessible via Zoom, see link on Canvas
- 1st homework will be released on Friday
- **TAs:** Deb Roy, Michael Zhang