Review of Linear Algebra and Probability

Vectors, Matrices and Solvability
Probability, Expectation, Variance

Outline

- Why do we need to review linear algebra?
- Vectors and Matrices.
- Matrix Multiplication
- Inverses and Transpose
- When do solutions exist, and when are they unique?
- What is a probability? What is Expectation and Variance?
- Conditional probabilities and joint distributions.

Why do we need linear algebra?

- All of data science is built around constructing models using a collection of ‘features’ or ‘predictors’ (ie. predict income from GPA). It is almost always the case that **there are several variables involved.** If there is **more than one, we need linear algebra to make sense of any equations.**
- For example, let y be the number of rooms that will be booked at a hotel tomorrow, and X_i be the ith observation of N variables, which could include **location, cost, rating** , etc. We can write a simple linear model as:

```
In [53]: df.head()
```

Out[53]:

	account	const	hotel_rating	location	price_per_night_avg	purchase_velocity_lastweek	rooms_left	sellouts_total
0	72722	1	1.6	9.5	584	60	198	2
1	20627	1	8.2	9.2	503	326	439	8
2	55924	1	8.0	0.9	467	327	240	8
3	14773	1	9.5	9.9	543	102	286	4
4	60469	1	1.8	5.7	144	42	335	2

$$y_i = \sum_{k=1}^{N} \beta_k X_{ik} + c \qquad \text{or} \qquad \mathbf{y} = \beta \cdot \mathbf{X} + \mathbf{c}$$

Things I’m going to gloss over

- The abstract definitions of vector spaces
- Detailed rules about existence and uniqueness of solutions.
- Proofs about any linear algebra related theorems.
- My goal is to remind people of the basic properties of matrices and notation we will use. Any deeper results (ie. PCA) will be presented in the context of a concrete problem.

Basic Properties of Matrices

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned}$$

Matrices are used to express systems of equations. They can be interpreted as representations of linear operators in a particular basis.

The above can be written as

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{A} = [a_{ij}]$$

$$\mathbf{b} = [4, 2, 8]$$

$$a_{11} = 4, a_{12} = -2, a_{23} == -1, \text{etc}$$

Here, **b** is considered a **vector.**

Matrix Addition - Done by components

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned} \qquad \begin{aligned} 5x - 3y + 4z &= 6 \\ 2x + y - z &= 4 \\ 5x + 7y - 3z &= 1 \end{aligned}$$

$$\mathbf{Ax = b} \qquad \mathbf{Cx = d}$$

$$(\mathbf{A + C})\mathbf{x} := \mathbf{Dx} = \mathbf{b + d} := \mathbf{e}$$

$$\begin{aligned} 9x - 5y + 7z &= 10 \\ 7x + 4y - 2z &= 6 \\ 8x - 5y - 9z &= 9 \end{aligned}$$

Matrix Multiplication

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned}$$

$$\begin{aligned} 5x - 3y + 4z &= 6 \\ 2x + y - z &= 4 \\ 5x + 7y - 3z &= 1 \end{aligned}$$

$$\mathbf{Ax = b}$$

$$\mathbf{Cx = d}$$

$$\mathbf{AB} := \mathbf{M} = [m_i j] := \sum_{i,j} a_{ik} b_{kj}$$

Matrix Multiplication

$$\begin{bmatrix} 4 & -2 & 3 \\ 5 & 3 & -1 \\ 3 & 2 & -6 \end{bmatrix} \begin{bmatrix} 5 & -3 & 4 \\ 2 & 1 & -1 \\ 5 & 7 & -3 \end{bmatrix}$$

$\mathbf{A}$ $\mathbf{B}$

$$\mathbf{AB} := \mathbf{M} = [m_{ij}] := \sum_k a_{ik} b_{kj}$$

Ith row of A, dotted with j th column of B

$$m_{11} = 4 \cdot 5 - 2 \cdot 2 + 3 \cdot 5$$
$$m_{12} = 4 \cdot (-3) - 2 \cdot 1 + 3 \cdot 7$$

Note: $\mathbf{AB} \neq \mathbf{BA}$

Matrix Multiplication

$$\begin{bmatrix} 4 & -2 & 3 \\ 5 & 3 & -1 \\ 3 & 2 & -6 \end{bmatrix} \begin{bmatrix} 5 & -3 & 4 \\ 2 & 1 & -1 \\ 5 & 7 & -3 \end{bmatrix}$$

$\mathbf{A}$ $\mathbf{B}$

$$\mathbf{AB} := \mathbf{M} = [m_i j] := \sum_{i,j} a_{ik} b_{kj}$$

ith row of A, dotted with j th column of B

$$\mathbf{M} = \begin{bmatrix} 31 & 7 & 9 \\ 26 & -19 & 20 \\ -11 & -49 & 28 \end{bmatrix}$$

Following this, we obtain
(**Exercise:** Verify this if you are rusty)

How does this relate to us?

- Matrices are basis representations of linear operators.
- The rows for us will always be observations, and the columns will be the variables themselves.

$$y_i = \sum_{k=1}^{N} \beta_k X_{ik} + c$$

or

$$\mathbf{y} = \beta \cdot \mathbf{X} + \mathbf{c}$$

Existence and Uniqueness of Solutions

Inverse, Transpose and Fredholm Alternative

When do systems of equations have solutions?

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned}$$

$$\mathbf{Ax} = \mathbf{b}$$

$$x = \mathbf{A}^{-1}\mathbf{b}$$

$$\begin{bmatrix} 4 & -2 & 3 \\ 5 & 3 & -1 \\ 3 & 2 & -6 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 16/115 & 6/115 & 7/115 \\ -27/115 & 33/115 & -19/115 \\ -1/115 & 14/115 & -22/115 \end{bmatrix}$$

Solution is computed using by reducing to Reduced Row Echelon form. Since this isn't essential for this course, we are going to omit the details.

When do systems of equations have solutions?

$$\mathbf{Ax} = \mathbf{b}$$

$$x = \mathbf{A}^{-1}\mathbf{b}$$

- Inverses exist only for square matrices (ie. column dimension is the same as the row dimension)

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{b} = [4, 2]$$

No solution

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{b} = [4, 0]$$

Infinitely many solutions

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \mathbf{b}$$

Exactly one solution

- Review this if you are not familiar. Will be important when we cover **regularization and collinearity.**
- The main point to take home here is to understand that **these three possibilities exist and only these**.

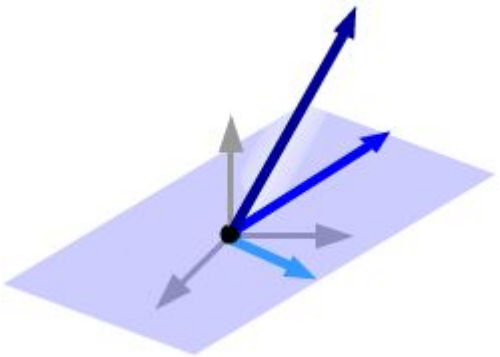
Eigenvalues, Eigenvectors and Linear Dependence.

Important for dimensionality reduction and stability.

Linear Independence

A collection of vectors (or features in our case), are **linearly dependent** if and only if

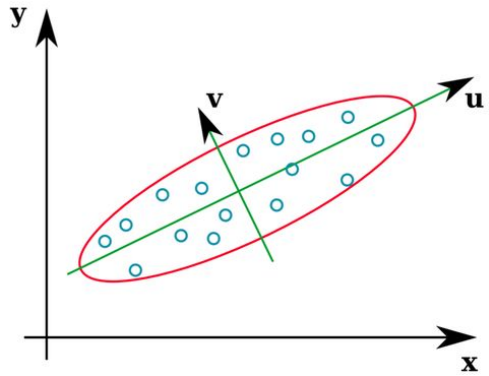
$$\sum_{j=1}^{N} \alpha_j X_j = 0 \text{ for some } \{\alpha_j\} \in \mathbb{R}$$



Otherwise, they are **linearly independent.**

Important since linear dependent vectors cause instability in solutions, and take up unnecessary space/resources.

Eigenvalues and Eigenvectors

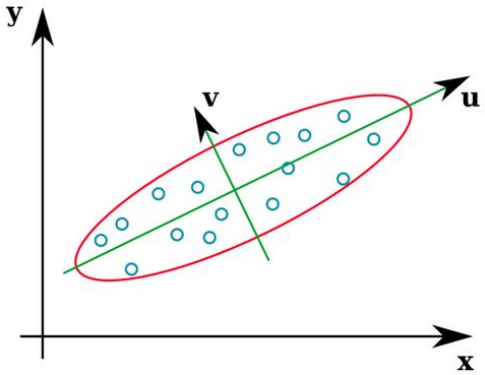


$$Ax = \lambda x \quad \lambda \in \mathbb{R}$$

$$x \in \mathbb{R}^n$$

- If A is invertible, there exists an orthogonal set of eigenvectors x and eigenvalues $\lambda \in \mathbb{R}$
- These are important for dimensionality reduction and for stability of solutions to ordinary least squares (next deck).

Eigenvalues and Eigenvectors



$$X^T X$$

- The matrix here will be common in the course. It is positive semi-definite.

$$y^T X^T X y \geq 0$$

- The matrix measures the dependence between features when X is your data. If your data is mean centered, it's the same as the correlation matrix.

Review of Probability

Expectation, Variance and Conditional Distributions

Outline

- Definitions of a probability, expectation, variance.
- Basic examples.
- Conditional probability distributions

Goals of these slides

The goal of this section is to **briefly review definitions and some basic examples** - only so that we can understand a few statements made about decision trees. We will explain more advanced concepts later in the course such as limit theorems.

Some definitions

- A probability p takes events $\{A_i\}$ as inputs, and outputs a value between 0 and 1 inclusive.
- The closer to 1, the more probable the event. The closer to 0, the less probable.
- Y and X here denote **Random Variables** - they are just the outcome of experiments, such as rolling a die, or how many newspapers will be purchased at Starbucks.
- The events are outcomes of ‘experiments’ where uncertainty is involved, such as flipping a coin, or whether or not a user will click on an ad.

$$\sum_i p(A_i) = 1$$

Some definitions

- The expected value of a random variable Y is the sum over all outcomes x the probability of that outcome - it's the same as the mean, but has a different interpretation.

$$\mathbb{E}(Y) = \sum_{i=1}^{N} y_i p(y_i)$$

- The variance of a random variable is the L_2 distance to the mean with respect to the probability distribution. It's the same as the standard deviation when calculating on a list of numbers.

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} (y_i - \mathbb{E}(y))^2 p(y_i)$$

Special case for uniformly distributed data:

$$p(y_i) = \frac{1}{N}$$

An Example - Coin Flipping

$$p(Y = k) = \begin{cases} p & \text{if } k \text{ is } 1 \\ 1 - p & \text{if } k \text{ is } 0 \end{cases}$$

$$\mathbb{E}(Y) = p \cdot (1) + (1 - p) \cdot 0 = p$$

$$\text{Var}(Y) = (1 - p)^2 p + (0 - p)^2 (1 - p) = p - 2p^2 + p^2 = p(1 - p)$$

This is known as the Bernoulli distribution, and is very important in machine learning and for A/B tests.

Joint Distributions & Conditional Probability

Suppose the discrete random variables X and Y have supports S_x and S_y. Then we require P to satisfy

$$\sum_{x \in S_x, y \in S_y} P(X = x, Y = y) = 1$$

$$P(X|Y = y) = \frac{P(X \text{ and } Y = y)}{P(y)}$$

	$Y=-1$	$Y=0$	$Y=1$
$X=0$	0	$\frac{1}{3}$	0
$X=1$	$\frac{1}{3}$	0	$\frac{1}{3}$

$$P(X = 0|Y = -1) = \frac{0}{0 + \frac{1}{3}} = 0$$

$$P(X = 1|Y = -1) = \frac{1/3}{0 + \frac{1}{3}} = 1$$

Can be read as - probability of X = x and Y = y is X conditioned on Y=y times the probability that Y=y.

Conditional Expectation

	$Y=-1$	$Y=0$	$Y=1$
$X=0$	0	$\frac{1}{3}$	0
$X=1$	$\frac{1}{3}$	0	$\frac{1}{3}$

$$\mathbb{E}(X|Y=y) = \sum_x xP(X=x|Y=y)$$

$$P(X=0|Y=-1) = \frac{0}{0+\frac{1}{3}} = 0$$

$$P(X=1|Y=-1) = \frac{1/3}{0+\frac{1}{3}} = 1$$

$$\mathbb{E}(X|Y=-1) = 0 \cdot 0 + 1 \cdot 1 = 1$$

Why is this intuitive?

Conditional Variance

	$Y=-1$	$Y=0$	$Y=1$
$X=0$	0	$\frac{1}{3}$	0
$X=1$	$\frac{1}{3}$	0	$\frac{1}{3}$

$$\mathrm{Var}(X|Y=y) = \sum_x (x - \mathbb{E}(X|Y=y))^2 p(X=x|Y=y)$$

$$P(X=0|Y=-1) = \frac{0}{0+\frac{1}{3}} = 0$$

$$P(X=1|Y=-1) = \frac{1/3}{0+\frac{1}{3}} = 1$$

$$\mathbb{E}(X|Y=-1) = 0 \cdot 0 + 1 \cdot 1 = 1$$

Why is the result below obvious as well?

$$\mathrm{Var}(X|Y=-1) = (0-1)^2 \cdot 0 + (1-1)^2 \cdot 1 = 0$$

Summary

- **If the above was quick, don't worry! -** we will cover it again, and in more detail.
- Conditional probabilities are extremely important in data science / machine learning, since in all cases we want to maximize some reward function, subject to some constraints (usually attributes of what we are trying to predict). In fact..

- In **predictive machine learning** we try to solve:

$$\max_{\theta} p(Y|X,\theta)$$

- In **prescriptive machine learning**, we try to solve:

$$\max_{a,\theta} \mathbb{E}(Y|X,a,\theta)$$