

# Decision Trees, Linear Algebra and Bias-Variance Decomposition

---

CSC311 Fall 2024

University of Toronto

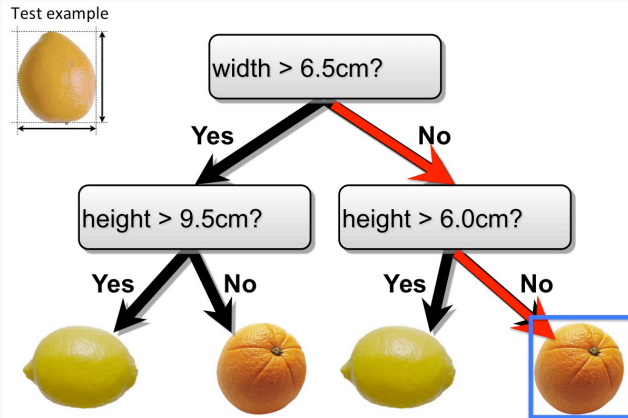
# Decision Trees Review

---

# Decision Trees Review

- A non-linear algorithm for classification and regression.
- Represents features of data in a tree-structure.
- Each node corresponds to one feature and thresholds that cover its possible values.
- Each branch from a node divides the data into bins based on its feature and thresholds.
- Leaves of the tree correspond to targets or outputs.

# Decision Trees Review



# Features

Features may be discrete or continuous.

- Discrete: Takes values in some discrete finite set. “Thresholds” just assign each branch to a different value. For example, a feature may be boolean and take values in

$\{\text{True}, \text{False}\}$

.

- Continuous: Takes a range of continuous values. “Thresholds” divide the range based on some value. For example, a feature like height may have thresholds 6, 9.5, dividing the data into the bins:

$\{\text{Height} \leq 6, 6 \leq \text{Height} \leq 9.5, \text{Height} \geq 9.5\}$

Outputs may be discrete or continuous.

- Discrete: Classification Tree
- Continuous: Regression Tree

We need some heuristic to determine good splits that guide decision making.

- Choose feature that will maximize *information gain* greedily.
- Repeat at every node.
- Stop when leaves are empty or contain examples of the same class.

# Linear Algebra

---



# Linear Algebra

We will use linear algebra tools to concisely depict data, parameters and measure different quantities like norms, similarity, projections, etc.

Some basic elements:

- Scalar: A number. Denoted by lowercase letters like  $a$ .
- Vector: A 1-D array of numbers. Denoted by bold lowercase  $\mathbf{a}$ .
- Matrix: A 2-D array of numbers. Denoted by bold uppercase  $\mathbf{A}$ .

# Norms

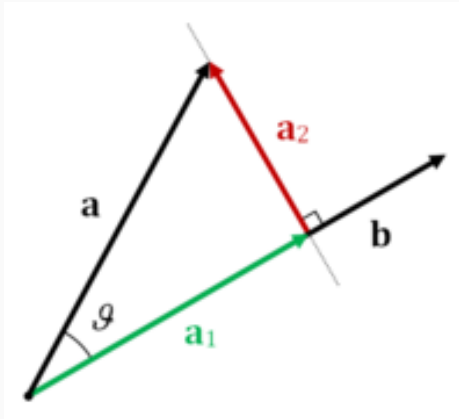
Norm is a measure of how “large” a vector is.

$$l_p\text{-norm } ||x||_p = \left[ \sum_i |x_i|^p \right]^{1/p}$$

- $l_2$ -norm is called the Euclidean norm:  $\sqrt{\sum_i x_i^2}$ .
- $l_1$ -norm is called the Manhattan norm:  $\sum_i |x_i|$ .
- $l_\infty$ -norm is called the max norm:  $\max_i |x_i|$ .

# Projections

When studying linear models, we will encounter vector projections<sup>1</sup>.



---

<sup>1</sup>Image from Wikipedia

# Projections

- Each vector is determined by its magnitude and direction.
- Projection of one vector on another can be thought of as dropping a perpendicular from one to the other.
- The magnitude of the projection is determined by the magnitude of the first vector and the angle between the two vectors.
- The direction of the projection is the same as that of the second vector.
- Mathematically, the projection of **a** on **b** is given by  $\frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{b}\|_2}$ .

## Exercise: Linear Algebra Notation

Suppose we are trying to predict commute times based on the distance traveled and day of the week. We have the following data:

dist	day	commute time
2.7	1	25
3.4	1	31
5.2	2	45
1.0	3	16
2.8	5	22

We estimate that commute times have the following relationship:

$$\text{commute time} = 10 \times \text{dist} - \text{day}$$

What are our predicted commute times? How can we use matrices to compute this quickly?

## Exercise: Linear Algebra Notation

## Exercise: Linear Algebra Notation

Suppose we want to calculate the average mean squared error between the predictions and the ground truth. How do we do this?

# Bias-Variance Decomposition

---



# Bias-Variance Decomposition

For training, we choose datapoints by sampling i.i.d. from some data distribution. This introduces randomness into the outputs of the model.

- Consider the squared error loss between outputs and targets,  $(y - t)^2$ .
- Treat both  $y$  and  $t$  as random variables.
- We saw in lecture that the expected loss can be decomposed into the bias and variance of  $y$ , the outputs.
- Recall that bias is the deviation of a random variable from its expectation.

# Bias-Variance Decomposition

Let's revisit the proof.

- Let  $y_* = \mathbb{E}[t]$ .
- From lecture, we have

$$\mathbb{E}[(y - t)^2] = \mathbb{E}[(y - y_*)^2] + \text{Var}(t)$$

- Here,  $\text{Var}(t)$  is called the Bayes error.

# Bias-Variance Decomposition

- We expand the first term and use linearity of expectation:

$$\mathbb{E} [(y - y^*)^2] = y_*^2 - 2y_*\mathbb{E} [y] + \mathbb{E} [y^2]$$

# Bias-Variance Decomposition

- Next, recall the definition of

$$\text{Var}(y) = \mathbb{E}[y^2] - \mathbb{E}[y]^2$$

to get

$$\mathbb{E}[(y - y_*)^2] = y_*^2 - 2y_*\mathbb{E}[y] + \mathbb{E}[y]^2 + \text{Var}(y)$$

# Bias-Variance Decomposition

- Note that

$$(y_* - \mathbb{E}[y])^2 = y_*^2 - 2y_*\mathbb{E}[y] + \mathbb{E}[y]^2$$

- Putting all this together, we have

$$\mathbb{E}[(y - t)^2] = (y_* - \mathbb{E}[y])^2 + \text{Var}(y) + \text{Var}(t)$$

- In words, *expected loss* = *bias* + *variance* + *Bayes error*.

## Exercise: Bias, Variance and Bayes Error

Assume we have  $N$  scalar-valued observations  $\{x^{(i)}\}_{i=1}^N$  sampled independently from some distribution with known variance 2 and unknown mean  $\mu$ .

We'd like to estimate the mean parameter  $\mu$ , or equivalently, choose a  $\hat{\mu}$  which minimizes the squared error risk  $\mathbb{E}[(x - \hat{\mu})^2]$ .

We will estimate the unknown mean parameter  $\mu$  by taking the empirical mean, or average, of the observations:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x^{(i)}$$

Compute the different terms from the bias-variance decomposition.

## Exercise: Bias, Variance and Bayes Error

Bayes Error:  $\mathbb{E} [(x - \mu)^2]$

## Exercise: Bias, Variance and Bayes Error

Bias:  $(\mathbb{E} [\hat{\mu}] - \mu)^2$



## Exercise: Bias, Variance and Bayes Error

Variance:  $\text{Var}(\hat{\mu})$