EECS 445 – Lecture 9
Decision trees

Professor Maggie Makar

A bit about me

Assistant Professor @ UM, 2022
PhD @ MIT, 2021
BSc @ Umass Amherst, 2013
Research: Machine learning & Causality

Causally motivated multi-shortcut identification & removal

Zheng & Makar. NeurIPS 2022

Causally Motivated Shortcut Removal Using Auxiliary Labels

Makar, et al., AIStats 2022

Estimation of Bounds on Potential Outcomes For Decision Making

Makar, et al., ICML 2019

FAIRNESS AND ROBUSTNESS IN ANTI-CAUSAL PREDICTION

Makar, D'Amour. TMLR 2022

Announcements

- **Project 1** due this Tuesday 2/13 at 10 pm.
- **Homework 2** due Tuesday 2/20 @10pm will be released after Project 1 due date.
- Double-check the number of late days you have remaining under Assignments > Late days > Remaining Homework Late Days.
- **Quiz 5** due on Sunday, 2/18 at 10 pm was released
- Note an update in the **course syllabus**
- Prof. Makar office hours 10:45 – 11:45 in BBB 3769

Mondays

Outline

- Recap
  - Linear models
  - Kernel models
- Decision trees
  - Why decision trees?
  - What are they?
  - How do we train them
    - What's a good tree?
    - Measuring uncertainty
    - Training algorithm
    - Bias Variance tradeoff

Recap: linear models, classification

$\in \{T, F\}$
$\in \{0, 1\}$

- Training data:

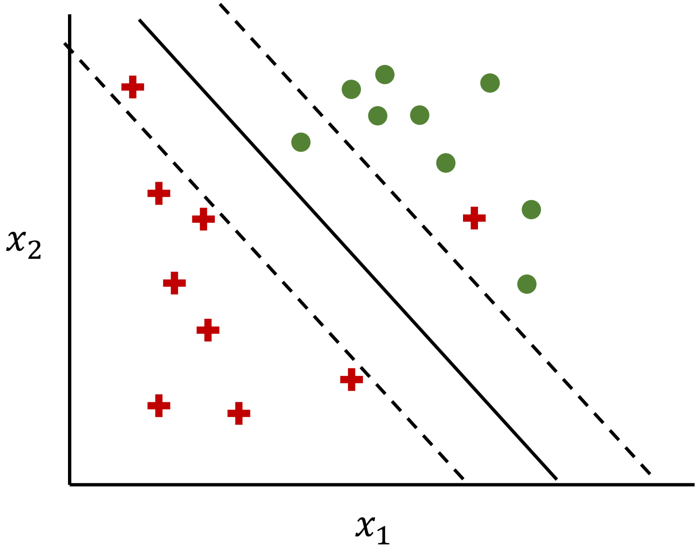
$$S_n = \{\bar{x}^{(i)}, y^{(i)}\}_{i=1}^n, \bar{x} \in \mathbb{R}^d, y \in \{-1, 1\}$$

- Recipe: SVM (soft margin)
  - Define the optimization problem (aka loss)

$$\min_{\bar{\theta}, \bar{\xi}, b} \frac{1}{2}\|\bar{\theta}\|^2 + C\sum_{i=1}^n \xi_i$$

$$\text{subject to } y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)} + b) \geq 1 - \xi_i$$

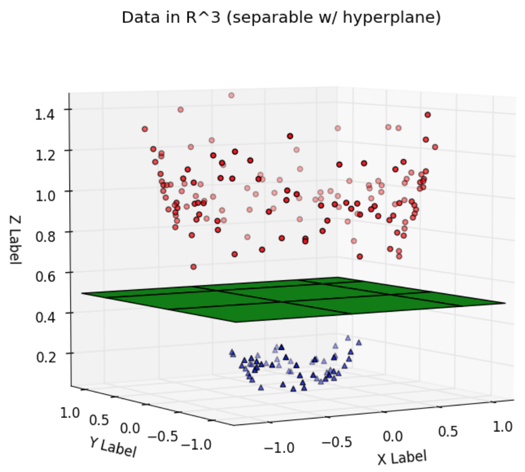
  - Find $\bar{\theta}^*, b^*, \bar{\xi}$ that minimize the loss

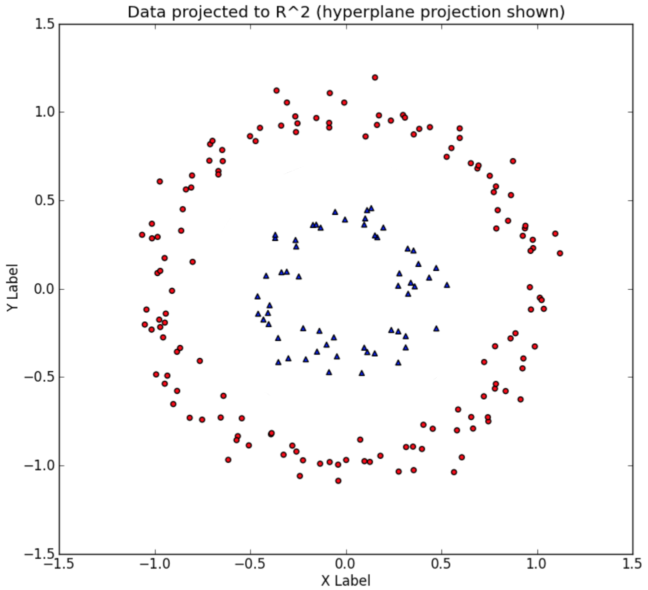


Non-linearly separable datasets

$x \in \mathbb{R}^1$

$\phi(x) = [x, 2x]$





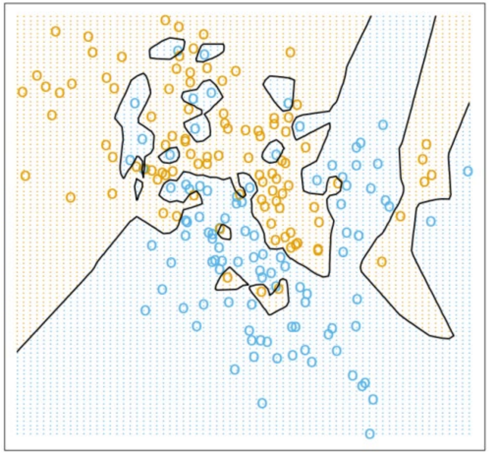
Kernel features:

- Map data to a higher dim. space in which there exists a separating hyperplane
- non-linear transformation of the feature space allows us to separate the data

Kernel trick: Avoid explicitly computing feature mappings

Non-linearly separable datasets

$$K(\bar{x}^{(i)}, \bar{x}^{(j)}) = \exp(-\gamma||\bar{x}^{(i)} - \bar{x}^{(j)}||^2)$$



Radial Basis Function
...aka Gaussian kernel
...aka Exponentiated Quadratic Kernel

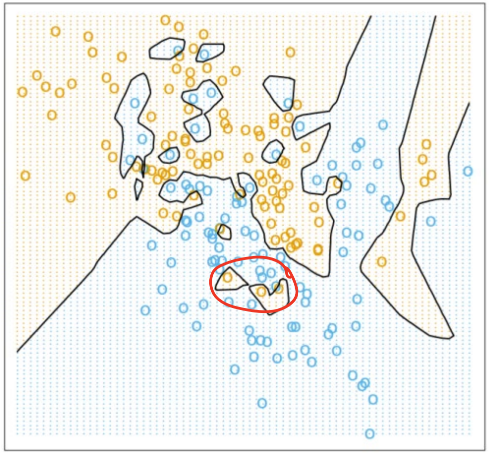
Kernel features:

- Map data to a higher dim. space in which there exists a separating hyperplane
- non-linear transformation of the feature space allows us to separate the data

Kernel trick: Avoid explicitly computing feature mappings

Non-linearly separable datasets

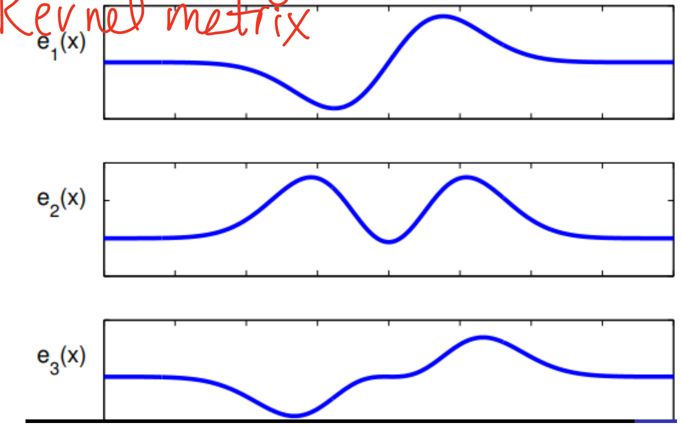
Will patient be admitted to ICU?



$$K(\bar{x}^{(i)}, \bar{x}^{(j)}) = \exp(-\gamma ||\bar{x}^{(i)} - \bar{x}^{(j)}||^2)$$

$$= \sum_{\ell=1}^{\infty} \phi_\ell(\bar{x}^{(i)}) \cdot \phi_\ell(\bar{x}^{(j)})$$

$$= \sum_{\ell=1}^{\infty} (\underbrace{\sqrt{(\lambda_\ell)}}_{\text{eigenvalue of Kernel matrix}}\underbrace{e_\ell(\bar{x}^{(i)})}_{\text{eigenfunction}}) \cdot (\sqrt{(\lambda_\ell)} e_\ell(\bar{x}^{(j)}))$$



Kernel features:

- Map data to a higher dim. space in which there exists a separating hyperplane
- non-linear transformation of the feature space allows us to separate the data

Kernel trick: Avoid explicitly computing feature mappings

Image from Arthur Gretton

Is accuracy all you need?

- So far we've focused on ERM
- But interpretability is important:
  - For the end users: trust and legal requirements

A Review of Challenges and Opportunities in Machine Learning for Health

Marzyeh Ghassemi, PhD[1], Tristan Naumann, PhD[2], Peter Schulam, PhD[3], Andrew L. Beam, PhD[4], Irene Y. Chen, SM[5], Rajesh Ranganath, PhD[6]

"In a clinical setting, black box methods present new challenges...clinical staff must [be able to] justify deviations in treatment to satisfy both clinical and legal requirements"

  - For developers: debugging

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

Interpretability *and* accuracy

- Decision trees are non-linear and interpretable
- They are also accurate *

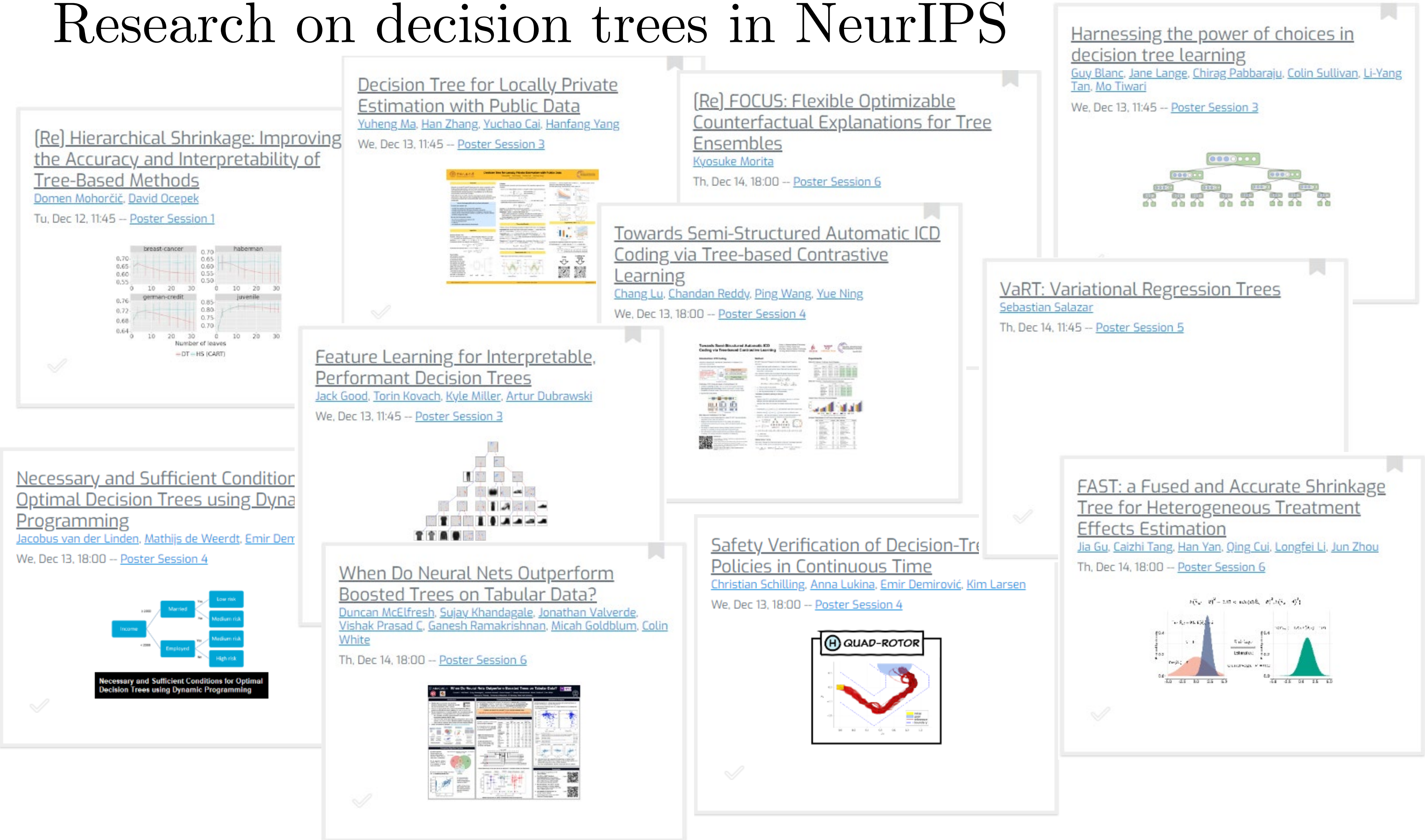
kaggle

A look at Mathurin's toolkit, which he keeps coming back to:

- **Packages**: scikit learn, pandas, numpy
- **Frameworks**: Keras, Tensorflow, Pytorch and Fastai
- **Algorithms**: lightgbm, xgboost, catboost
- **AutoML tools:** Prevision.io, h2o and other open sources such as TPOT, auto sklearn
- **Cloud services:** Google colab and kaggle kernels

https://analyticsindiamag.com/kaggle-master-mathurin-ache-competition-data-science/
Example due to Tamara Broderick

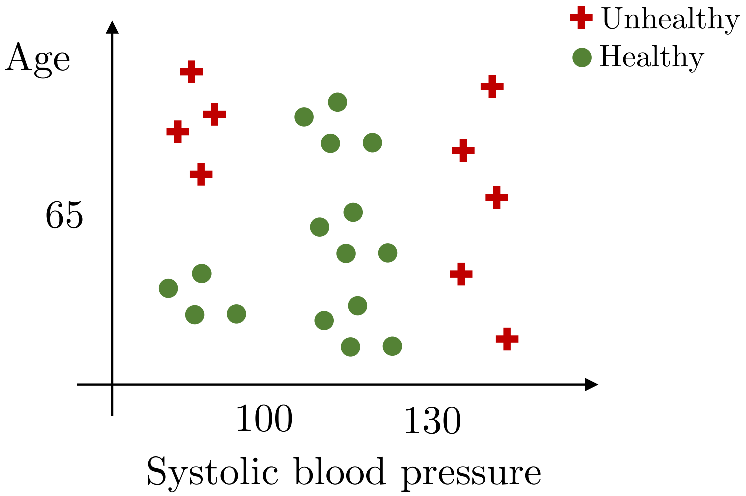
Research on decision trees in NeurIPS



TL;DPA: Decision trees are interpretable, and can be very accurate (especially as a building block for more complicated methods)

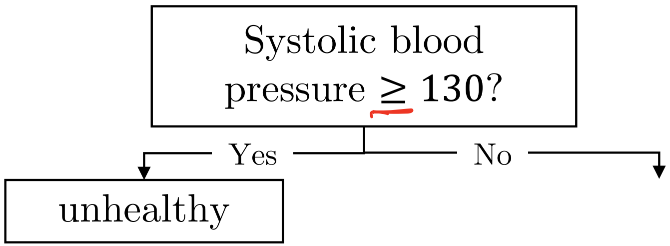
Decision trees: what are they? Classification trees

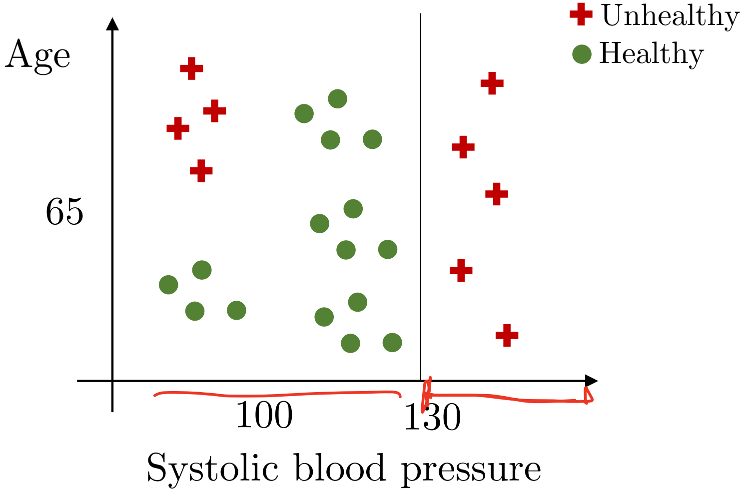
- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)



Decision trees: what are they?

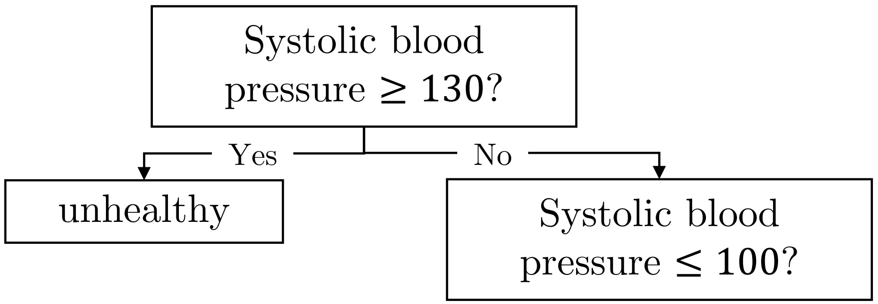
- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)

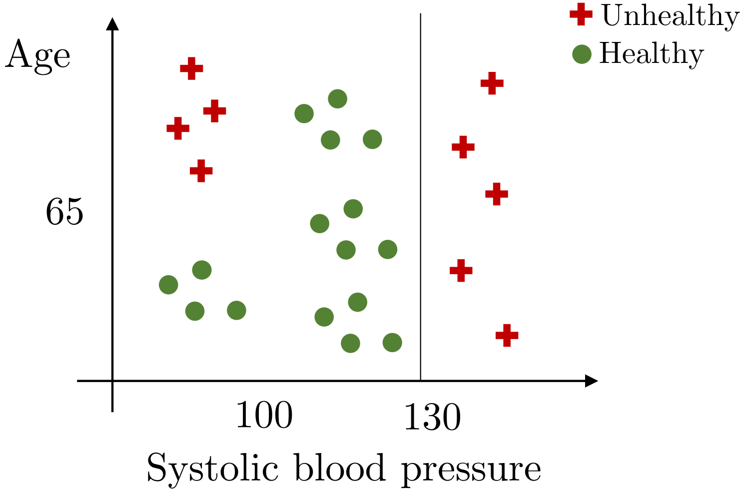




Decision trees: what are they?

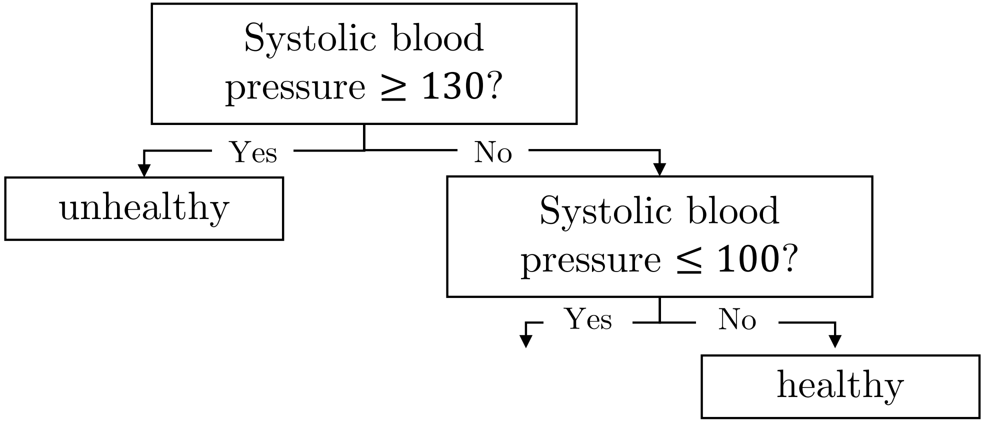
- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)

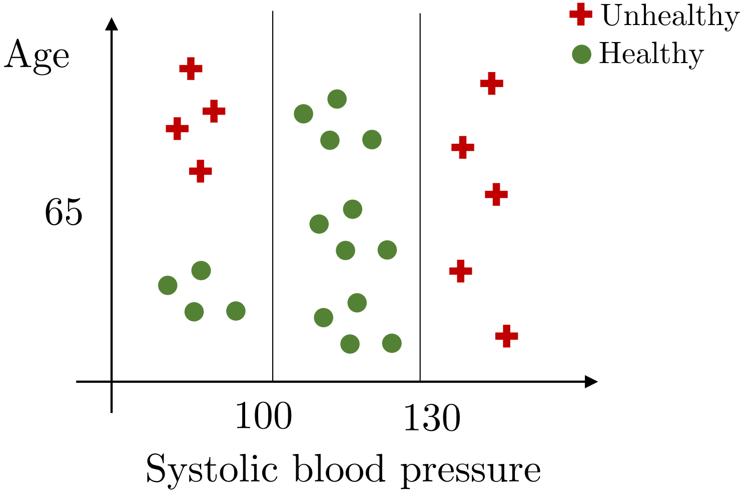




Decision trees: what are they?

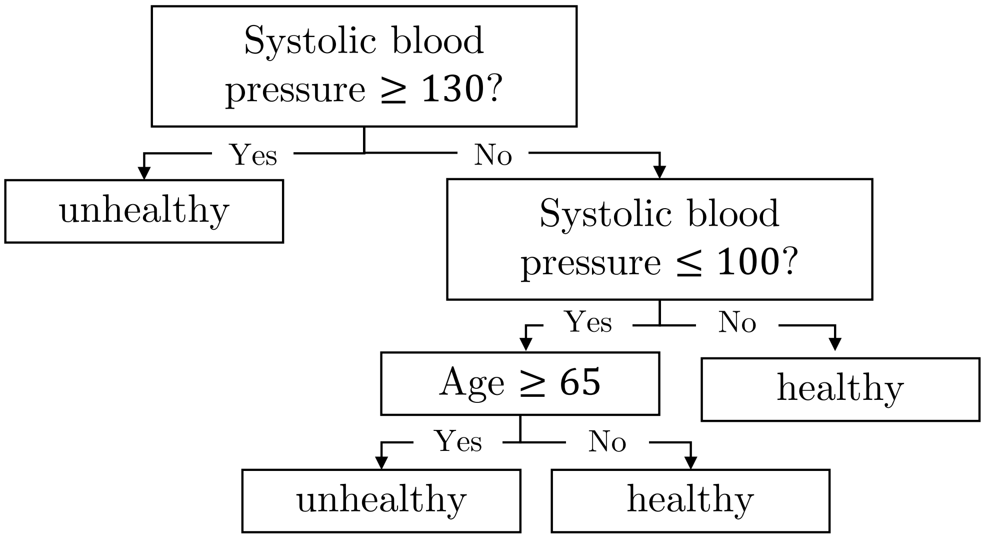
- $f : \bar{x} \to y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)

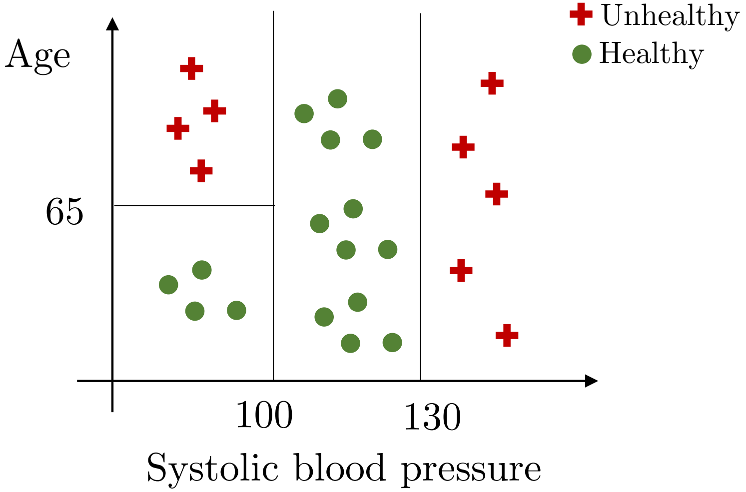




Decision trees: what are they?

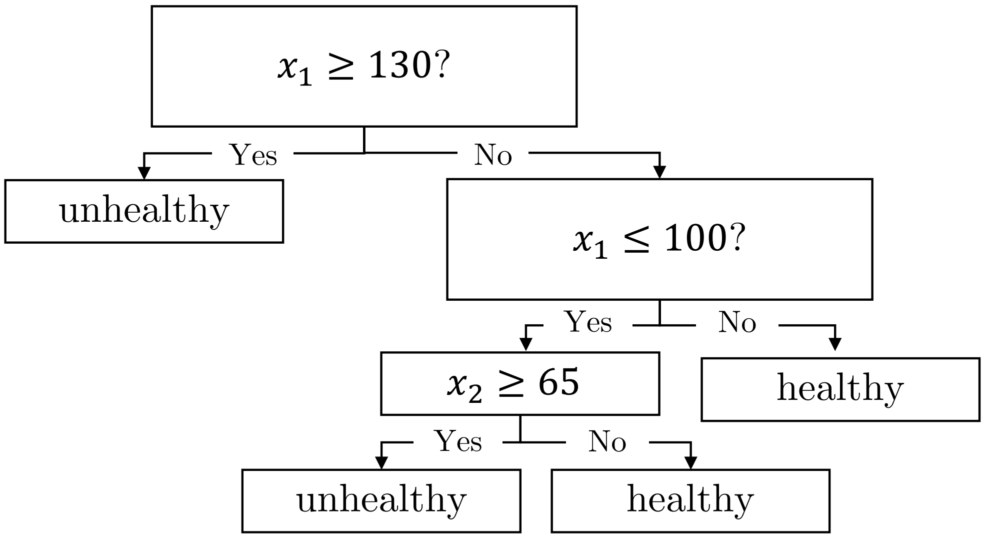
- $f : \bar{x} \to y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)

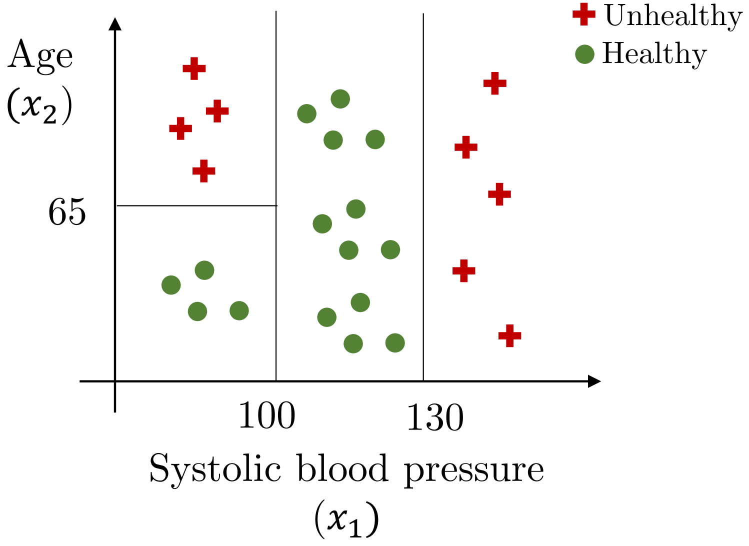




Decision trees: what are they

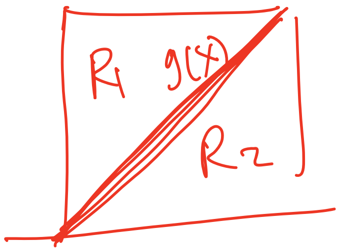
- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)

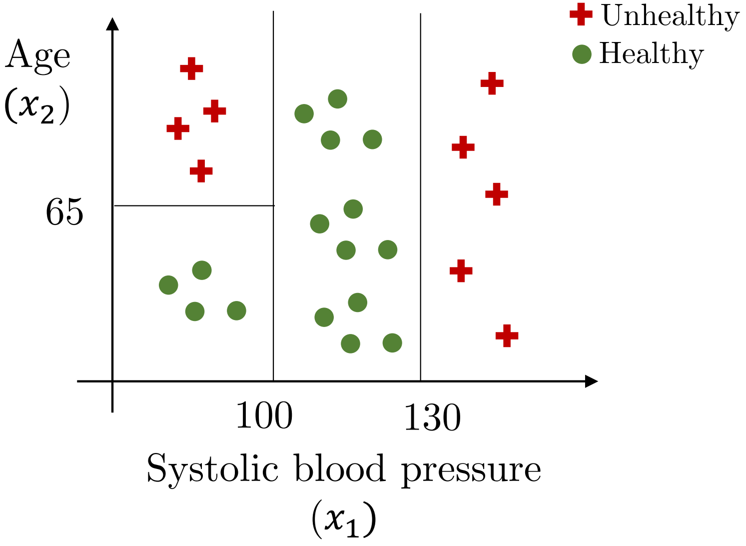




Decision trees: definitions

- $f : \bar{x} \to y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)
- Decision trees are:
  - Non-linear: decision boundary is non-linear
  - Axis aligned partitions:
    - Partition the input space $\bar{x}$
    - Axis aligned: parallel to the x and y axis



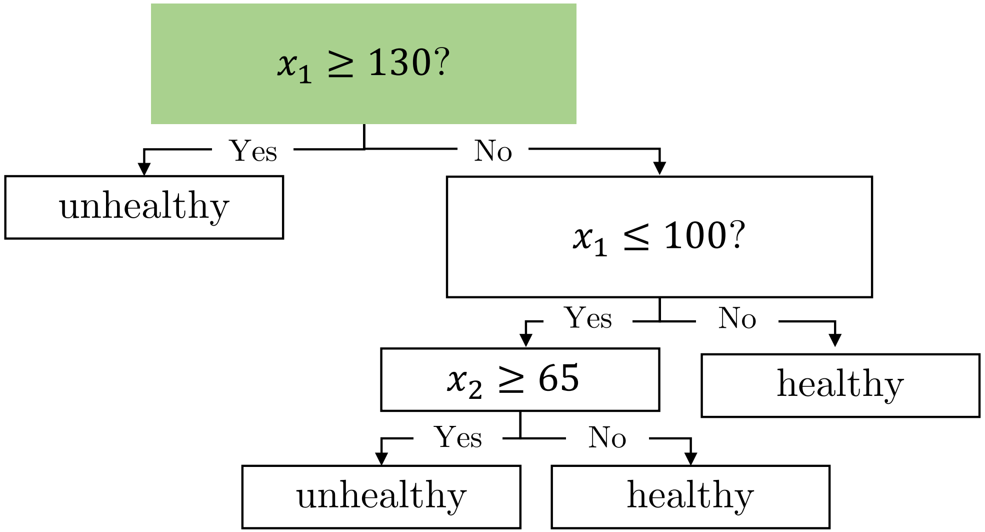


Decision trees: definitions

- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)

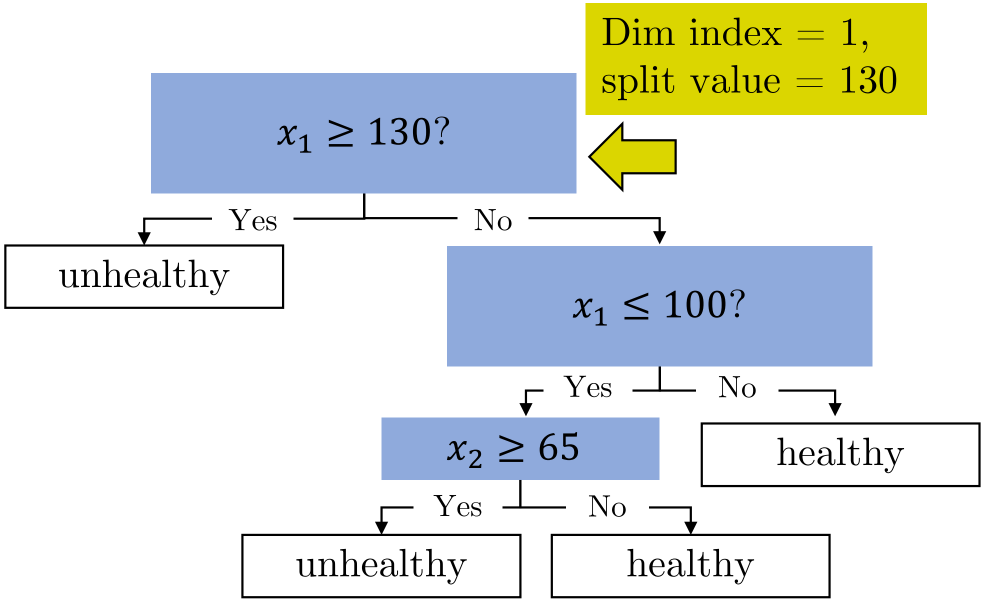
Root node:

- The first node



Decision trees: definitions

- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure $(\bar{x})$



Root node:

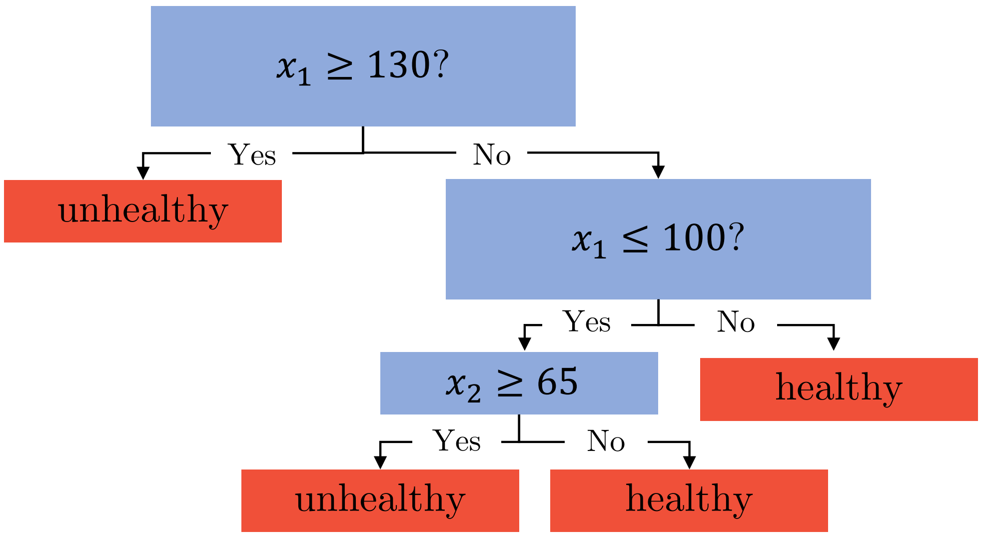
- The first node

Internal node:

- Defined by dimension index j and split value s
- Has 2 child nodes: either internal nodes or leaves

Decision trees: definitions

- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)



Root node:

- The first node

Internal node:

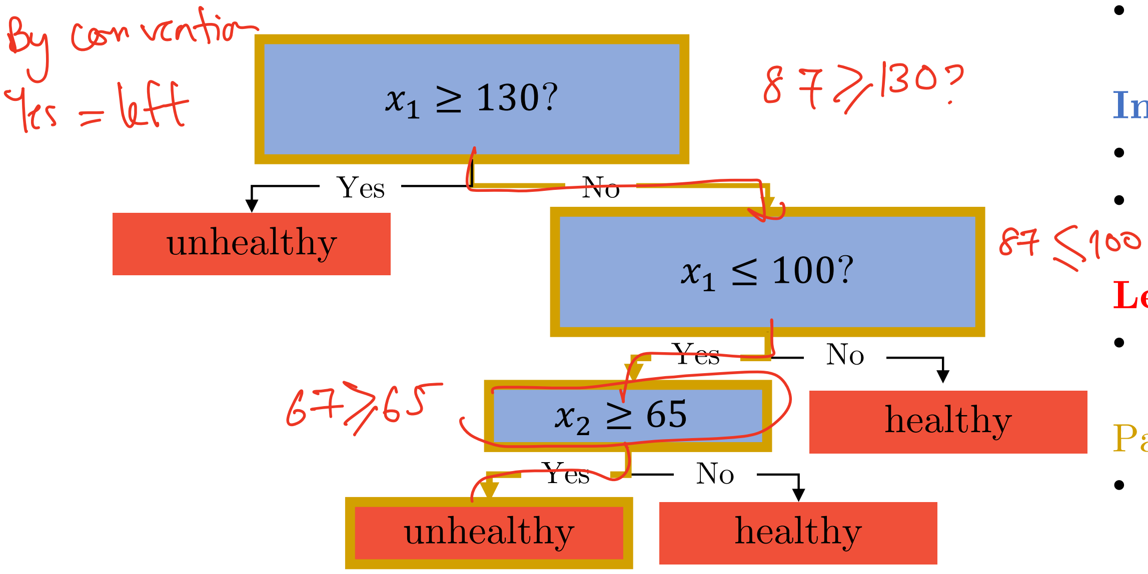
- Defined by dimension index j and split value s
- Has 2 child nodes: either internal nodes or leaves

Leaf:

- Assigns a label (discrete/continuous)

Decision trees: definitions

- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Predict if a patient is healthy or not (y) using their age and systolic blood pressure ($\bar{x}$)



Root node:

- The first node

Internal node:

- Defined by dimension index j and split value s
- Has 2 child nodes: either internal nodes or leaves

Leaf:

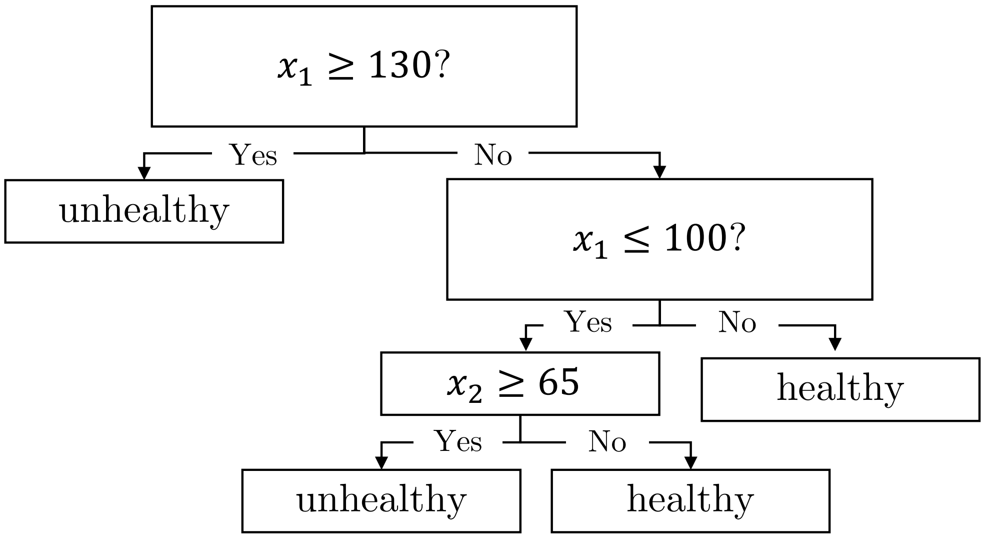
- Assigns a label (discrete/continuous)

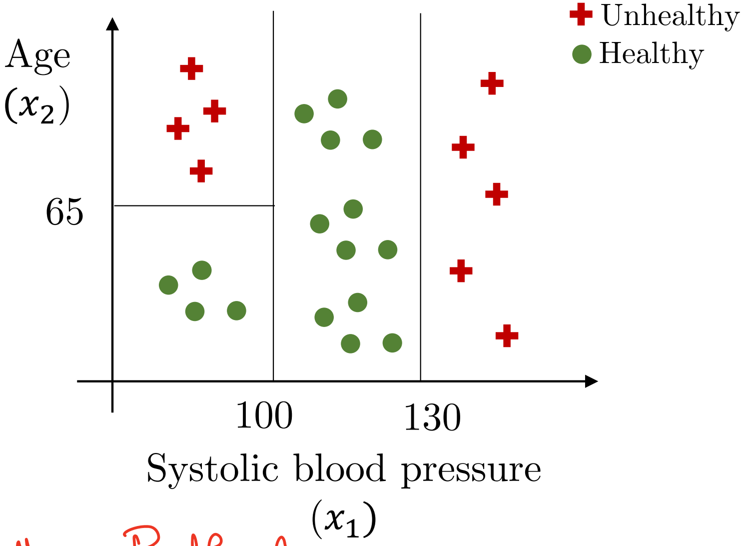
Path

- The nodes "traversed" by a group of data points. Highlighted: patient with systolic BP = 87, and age 67.

Decision trees: Prediction rule = most likely label

- $f : \bar{x} \rightarrow y, y \in \{0, 1\}$
- Classification setting: prediction rule is the "majority" label in a terminal leaf in the training data





① Follow Path for a given $\bar{x}$

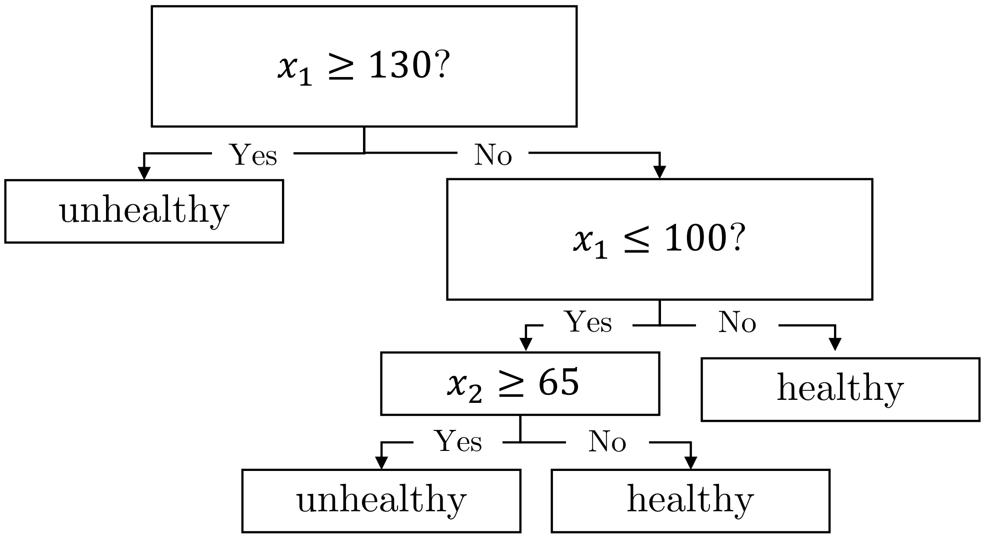
② get the majority label from training data.

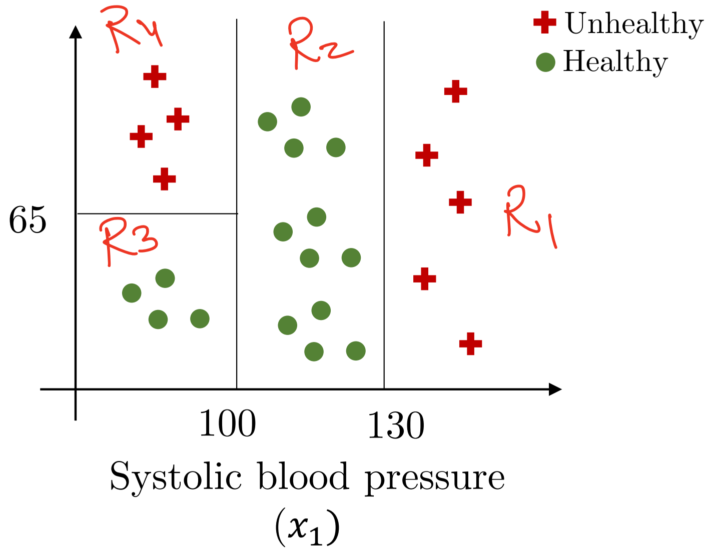
Decision trees: functional form

μ_1 = majority label in R_1

$$f(\bar{x}) = \mu_1[\![\bar{x} \in R_1]\!] + \mu_2[\![\bar{x} \in R_2]\!] + \mu_3[\![\bar{x} \in R_3]\!] + \mu_4[\![\bar{x} \in R_4]\!]$$

$$\mu_1 = \arg\max\Big(\sum_i y_i[\![\bar{x}_i \in R_1]\!], \sum_i (1 - y_i)[\![\bar{x}_i \in R_1]\!]\Big)$$

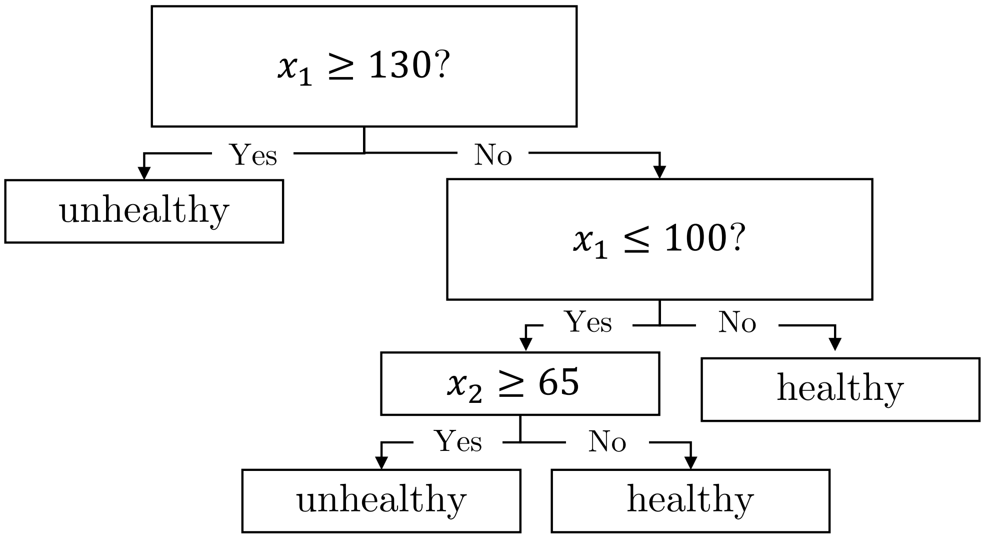


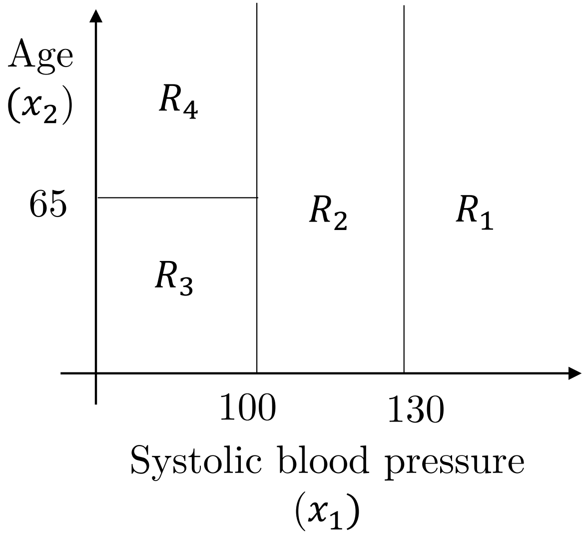


Decision trees: functional form

$$f(\bar{x}) = \sum_{m=1}^{M} \mu_m [\![\bar{x} \in R_m]\!]$$

$$\mu_m = \arg\max\left(\sum_i y_i [\![\bar{x}_i \in R_m]\!], \sum_i (1 - y_i)[\![\bar{x}_i \in R_m]\!]\right)$$

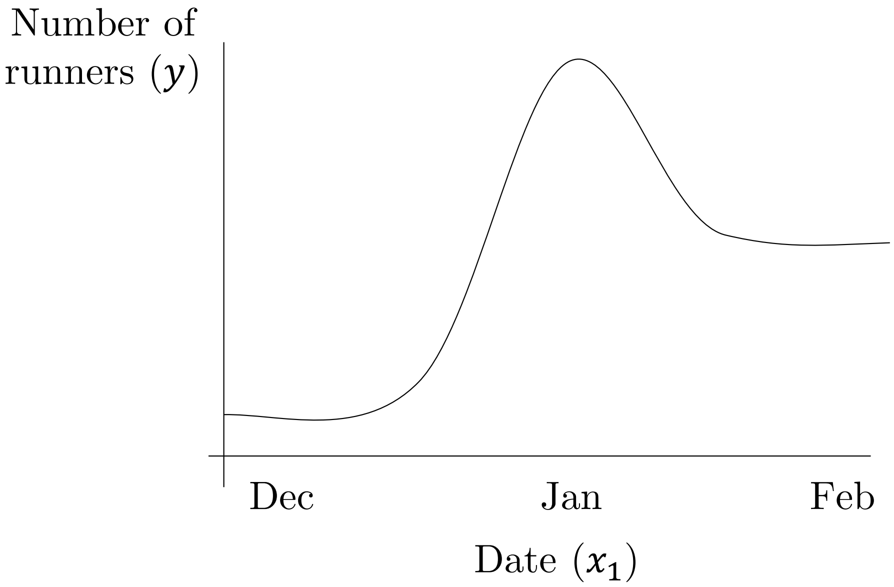




Decision trees: input data

Regression tree, $y \in \mathbb{R}$

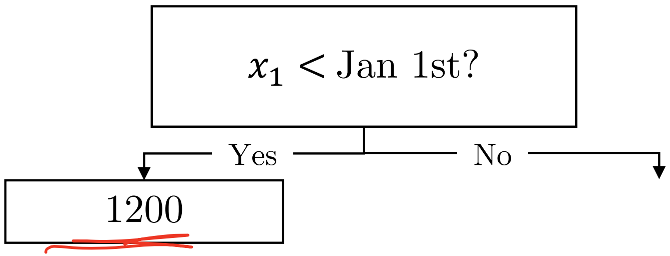
- Predict the number of winter runners (y) based on the date $(\bar{x})$

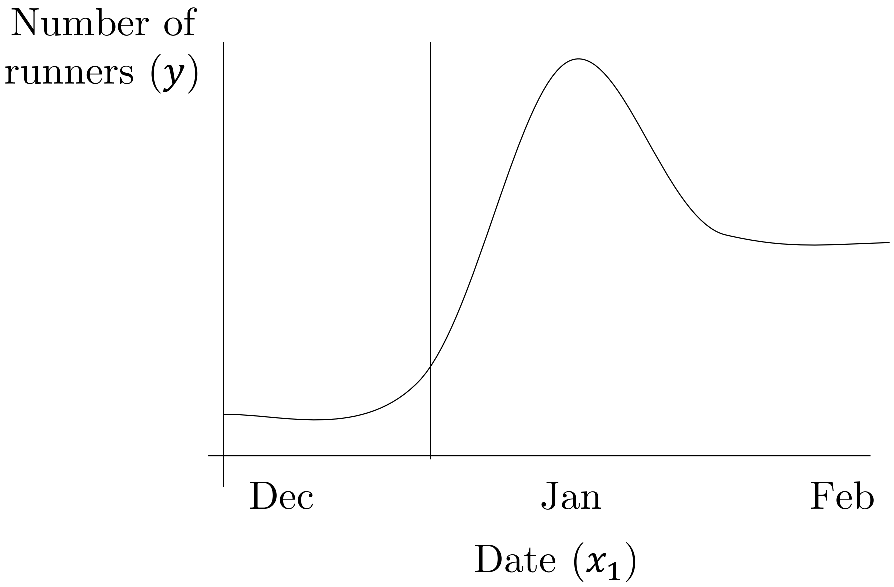


Decision trees: input data

Regression tree, $y \in \mathbb{R}$

- Predict the number of winter runners (y) based on the date ($\bar{x}$)

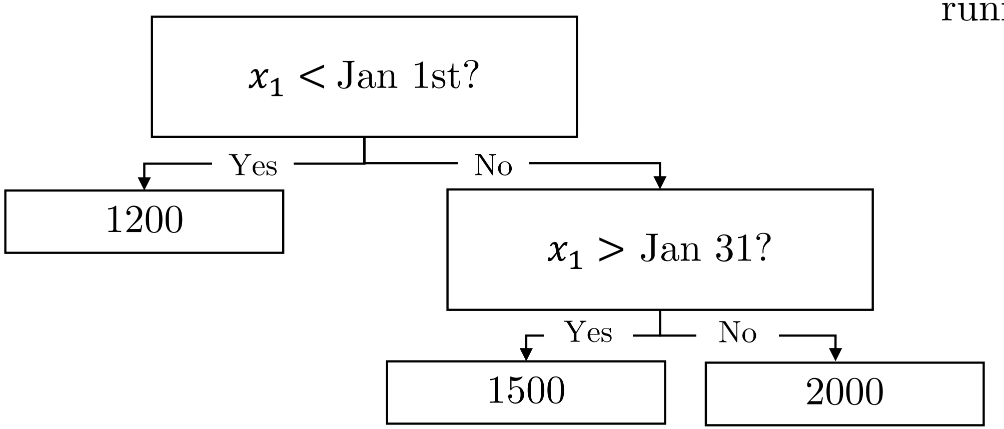


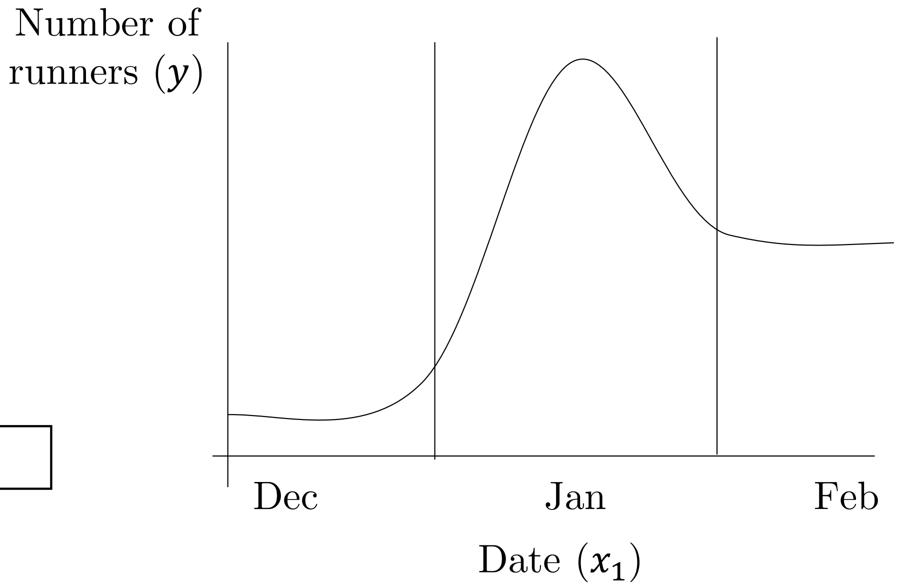


Decision trees: input data

Regression tree

- Predict the number of winter runners (y) based on the date ($\bar{x}$)

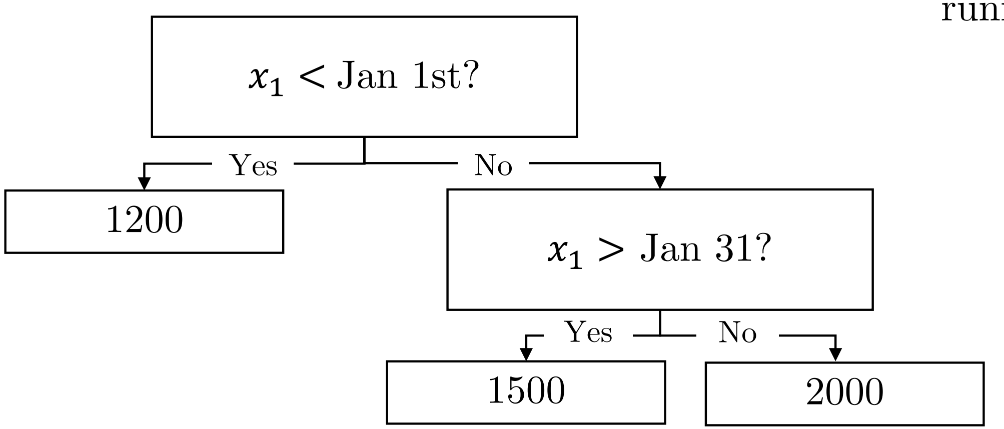


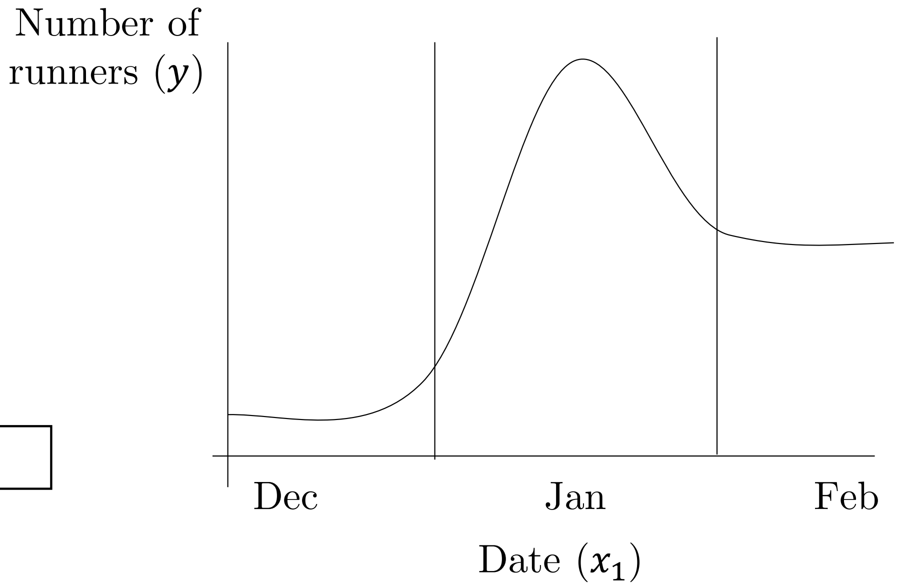


Decision trees: input data

Regression tree

- Predict the number of winter runners (y) based on the date ($\bar{x}$)





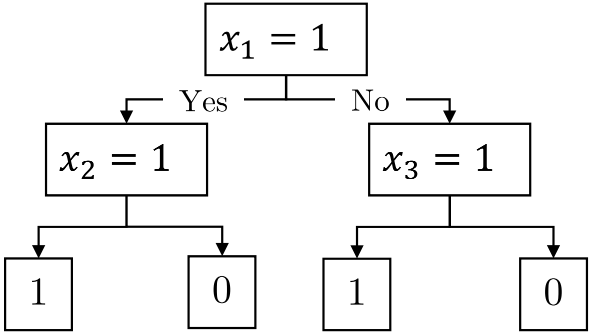
- Prediction rule: Most likely (=average) label for data points falling in leaf

TL;DPA: Decision trees are axis aligned partitions of the input data. The prediction rule is the most likely label in the leaf that the example falls into.

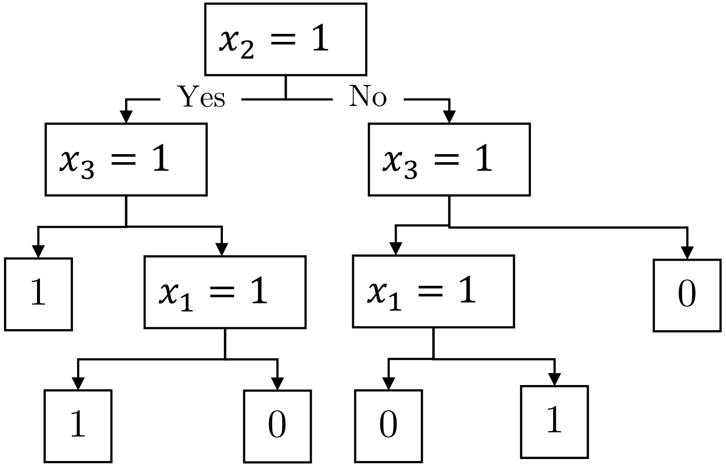
Desiderata for a good decision tree

1. Accurate
2. Smaller is better

Tree 1

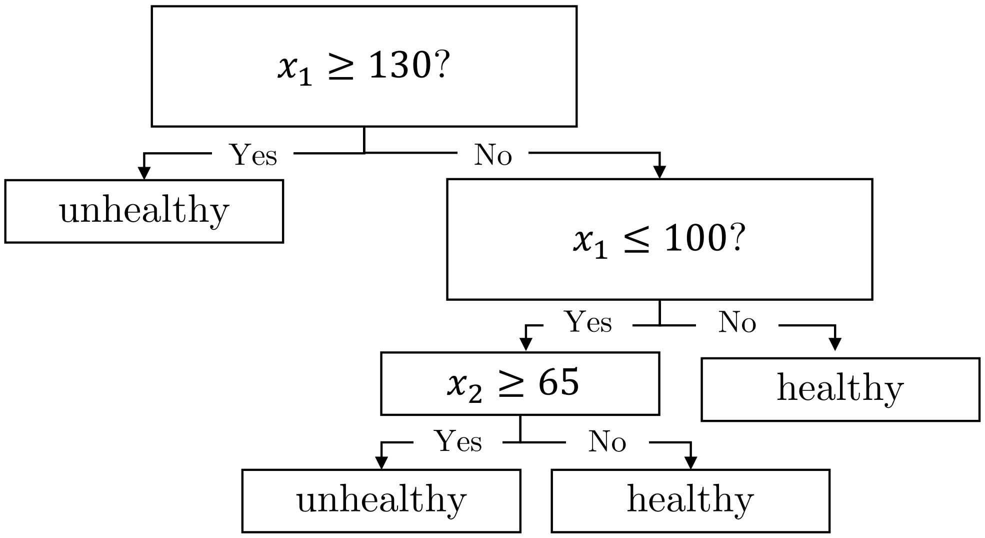


Tree 2



Training decision trees

- Want: Simplest (smallest) accurate decision tree
- Recall our training recipe:
  - Define a loss
  - Pick the parameters that minimize this loss



$$f(\bar{x}) = \sum_{m}^{M} \mu_m [\![\bar{x} \in R_m]\!]$$

- Parameters are not fixed in advance

Training decision trees

- Want: Simplest (smallest) accurate decision tree
- Possible new recipe: a brute force approach
  - Try all possible trees
  - Pick the best

- Number of possible trees grows exponentially with the dimension of the features and # of distinct values
- If we have d possible "tests" on a pathway, there are $d!$ different ways to order that test
- This problem is NP-hard (Hyafil and Rivest, 1976)

Training decision trees

- Want: Simplest (smallest) accurate decision trees
- Possible new recipe: Greedy approach
  - Idea: Use heuristics
    - Greedy approach to picking the best feature to split on
    - Want our splits to minimize uncertainty in the label

Training decision trees: Example 1

Choice 1

$X_1 = 1$

Y: $P(Y=1) = 0.5$ N: $P(Y=1) = 0.5$

Choice 2

$X_2 = 1$

Y: $\boxed{1}$ N: $\boxed{0}$

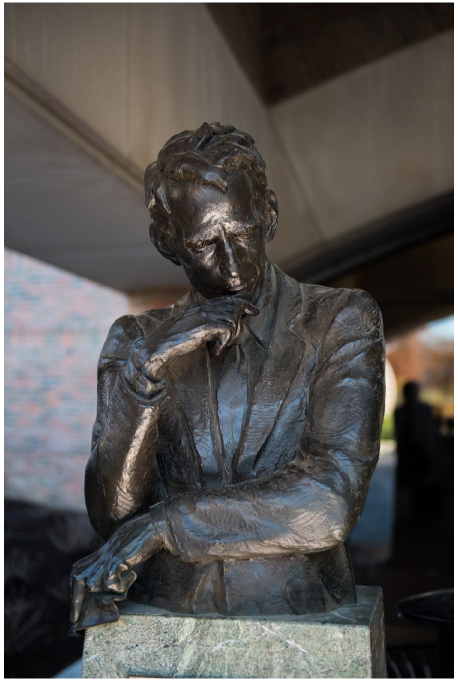
x_1	x_2	y
1	0	0
0	0	0
1	1	1
0	1	1

Training decision trees

- Want: Simplest (smallest) accurate decision trees
- Possible new recipe: Greedy approach
    - Idea: Use heuristics
        - Greedy approach to picking the best feature to split on
        - Want our splits to minimize uncertainty in the label

Measuring uncertainty using Shannon's entropy

- Shannon entropy: a measure of the amount of "uncertainty" in a variable



Measuring uncertainty using Shannon's entropy

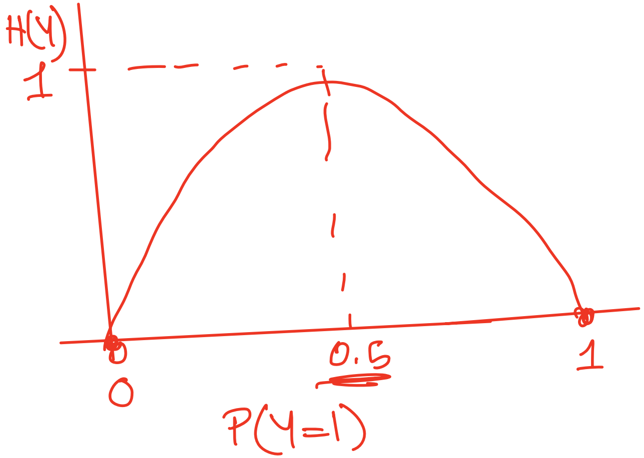
- Shannon entropy: a measure of the amount of "uncertainty" in a variable
- For a variable $Y \in \{0, 1\}$

$$\underbrace{H(Y)}_{\text{entropy}} = -(\underbrace{p(Y=1)}_{\text{Prop of data that has } Y=1}\log_2 p(Y=1) + \underbrace{p(Y=0)}_{\text{Prop of data that has } Y=0}\log_2 p(Y=0))$$

$P(Y=1)=1$

$H(Y) = -1\log_2 1 - \underbrace{0\log_2 0}_{=0}$

$\lim_{p_+ \to 0} p\log_2 p \to 0$



Measuring uncertainty using Shannon's entropy

- Shannon entropy: a measure of the amount of "uncertainty" in a variable
- For a variable $Y \in \{0, 1\}$

$$H(Y) = -(p(Y = 1) \log_2 p(Y = 1) + p(Y = 0) \log_2 p(Y = 0))$$

- More generally, for a discrete $Y \in [y_1, y_2, \ldots, y_k, \ldots, y_K]$:

$$H(Y) = -\sum_{k=1}^{K} p(Y = y_k) \log_2 p(Y = y_k)$$

- Check your understanding:

Which has a higher entropy: biased or fair dice?

Marginal and conditional entropy

- Entropy: Uncertainty in the value of Y

$$H(Y) = -\sum_{k=1}^{K} p(Y = y_k) \log_2 p(Y = y_k)$$

x_1	x_2	y
Sunny	Raining	0
Sunny	Dry	1
Cloudy	Raining	0
Cloudy	Dry	0
Cloudy	Dry	1

Marginal and conditional entropy

- Entropy: $H(Y) = -\sum_{k=1}^{K} p(Y = y_k) \log_2 p(Y = y_k)$
- The entropy of Y conditioned on $X = x$: Uncertainty in the value of Y among the "sub-dataset" defined by X=x

x_1	x_2	y
Sunny	Raining	0
Sunny	Dry	1
Cloudy	Raining	0
Cloudy	Dry	0
Cloudy	Dry	1

$$H(Y|\,X = x) = -\sum_{k=1}^{K} p(Y = y_k|\,X = x) \log_2 p(Y = y_k|\,X = x)$$

$H(Y|X_1 = \text{cloudy}) = -[P(Y=1|X_1=c)\log_2 P(Y=1|X_1=c) + P(Y=0|X_1=c)\log_2 P(Y=0|X_1=c)]$

$= -[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}] =$

$H(Y|X_2 = R) = -[1\log_2 1 + 0\log_2 0]$

Marginal and conditional entropy

- Entropy: $H(Y) = -\sum_{k=1}^{K} p(Y = y_k) \log_2 p(Y = y_k)$
- The entropy of Y conditioned on $X = x$:

$$H(Y \mid X = x) = -\sum_{k=1}^{K} p(Y = y_k \mid X = x) \log_2 p(Y = y_k \mid X = x)$$

- Conditional entropy:
  Uncertainty in Y after learning the value of X

$$\underbrace{H(Y \mid X)}_{\text{conditional entropy} \neq H(Y \mid X=x)} = \sum_{x} \underbrace{p(X = x)}_{\text{weighting}} \underbrace{H(Y \mid X = x)}_{\text{entropy Y condition on } X=x}$$

x_1	x_2	y
Sunny	Raining	0
Sunny	Dry	1
Cloudy	Raining	0
Cloudy	Dry	0
Cloudy	Dry	1

What is $H(Y|X_1)$?

- Entropy: $H(Y) = -\sum_{k=1}^{K} p(Y = y_k) \log_2 p(Y = y_k)$
- The entropy of Y conditioned on $X = x$:

$$H(Y|\ X = x) = -\sum_{k=1}^{K} p(Y = y_k|\ X = x) \log_2 p(Y = y_k|\ X = x)$$

- Conditional entropy:

$$H(Y \mid X) = \sum_{x} p(X = x) H(Y \mid X = x)$$

x_1	x_2	y
Sunny	Raining	0
Sunny	Dry	1
Cloudy	Raining	0
Cloudy	Dry	0
Cloudy	Dry	1

$$H(Y|X_1) = P(X_1 = \text{Sunny})\, H(Y|X_1 = \text{Sunny}) + P(X_1 = \text{Cloudy})\, H(Y|X_1 = \text{cloudy})$$

$$= \frac{2}{5}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] + \frac{3}{5}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]$$

$$= \frac{2}{5} + 0.55 = 0.95$$

What is $H(Y|X_2)$? Your turn!

- Entropy: $H(Y) = -\sum_{k=1}^{K} p(Y = y_k) \log_2 p(Y = y_k)$
- The entropy of Y conditioned on $X = x$:

$$H(Y|\ X = x) = -\sum_{k=1}^{K} p(Y = y_k|\ X = x) \log_2 p(Y = y_k|\ X = x)$$

- Conditional entropy:

$$H(Y \mid X) = \sum_{x} p(X = x) H(Y \mid X = x)$$

x_1	x_2	y
Sunny	Raining	0
Sunny	Dry	1
Cloudy	Raining	0
Cloudy	Dry	0
Cloudy	Dry	1

$$H(Y|X_2) = P(X_2 = \text{Rain})\, H(Y|X_2 = \text{Rain}) + P(X_2 = \text{Dry})\, H(Y|X_2 = \text{Dry})$$

$$= \frac{2}{5}\left[-0 \log_2 0 - 1 \log_2 1\right] + \frac{3}{5}\left[-\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3}\right]$$

$$= 0.55$$