Week 4

Machine Learning and Big Data - DATA622

Fall 2023

CUNY School of Professional Studies

Week 4

1. Discussion Board Week 4
2. Reading materials:
   - PMLiR Chapter 7: Naïve Bayes (20 mins)
   - PMLiR Chapter 9: Evaluating Performance (40 mins)
3. Reading materials:
   - Comparison of approaches: https://mdav.ece.gatech.edu/ece-6254-spring2022/notes/10-LR-NB.pdf
   - (Optional) Naïve Bayes Classifier:
     The PMLiR textbook provides a good overview on Naïve Bayes. If you want to know more review:
     - Read: ESL Section 6.6.3 "Naïve Bayes classifier" (page 210) - (10 mins)
     - Video: Simple explanation of Naïve Bayes: https://www.youtube.com/watch?v=O2L2Uv9pdDA
     - Video: Simple explanation of Gaussian Naïve Bayes: https://youtu.be/H3EjCKtlVog
   - Lab: R resource. https://uc-r.github.io/naive_bayes
     (If you prefer Python: https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/)

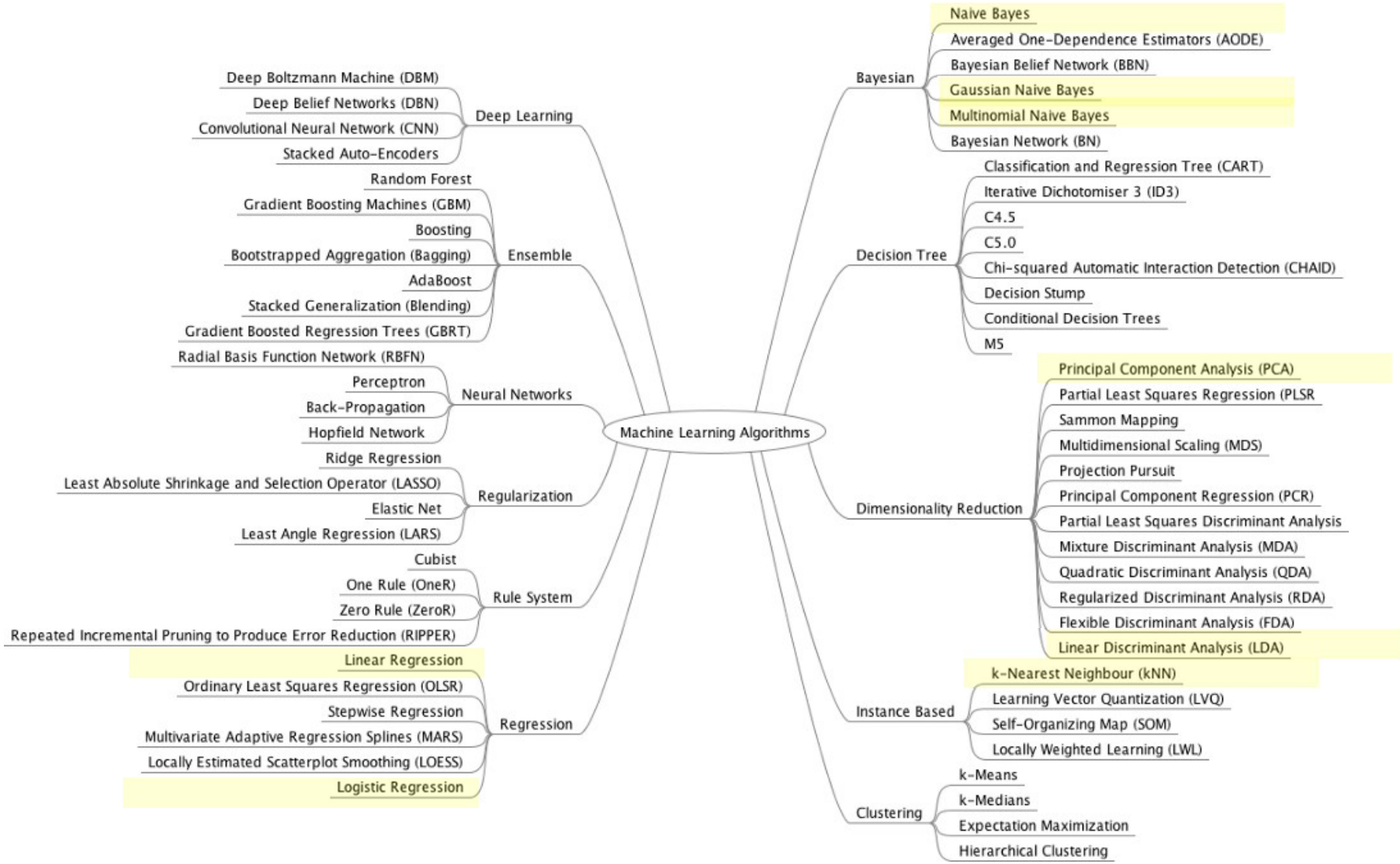
Textbooks

- We are introducing 2 new texbooks to supplement our primary textbook (PMLiR)
- The Elements of Statistical Learning
  - In the notes, "ESL" refers to the book "The Elements of Statistical Learning“
  - You should have from the prerequisite courses.
  - You can buy it here
  - Book is available for free as a PDF here (author's site here)
- An Introduction to Statistical Learning
  - In the notes, "ISLR" refers to the book "An Introduction to Statistical Learning“
  - You should have from the prerequisite courses.
  - You can buy it here
  - Book is available for free as a PDF here (author's site here)

Landscape of algorithms

We will cover many of the algorithms listed

CUNY School of Professional Studies

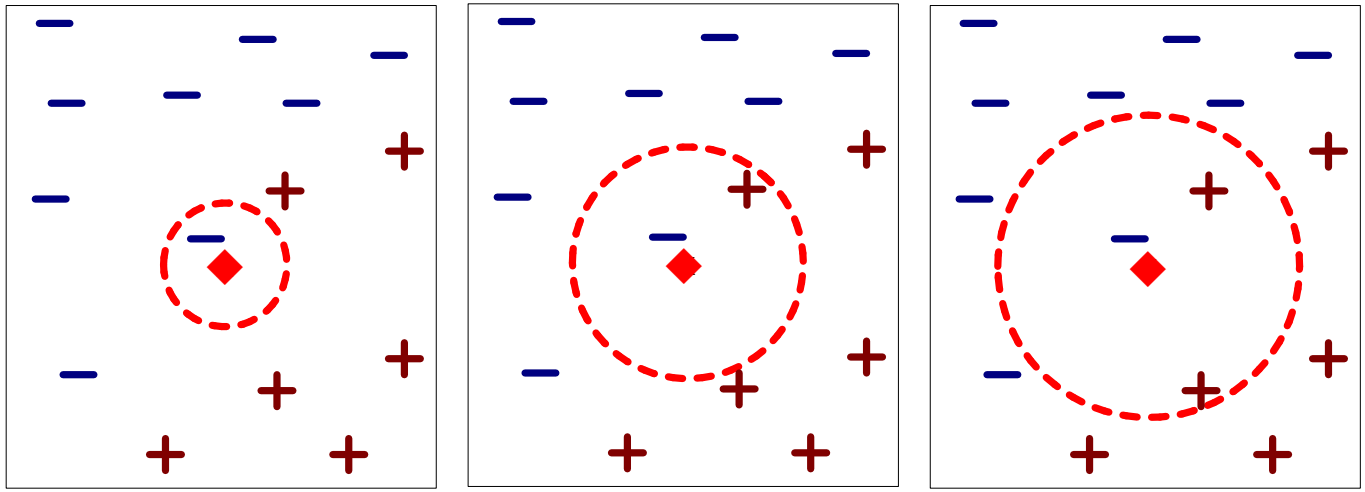

Machine Learning Algorithms
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

K-Nearest Neighbor

Classify data according to its k-closest neighbors

CUNY School of Professional Studies

k-Nearest Neighbor (KNN)



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

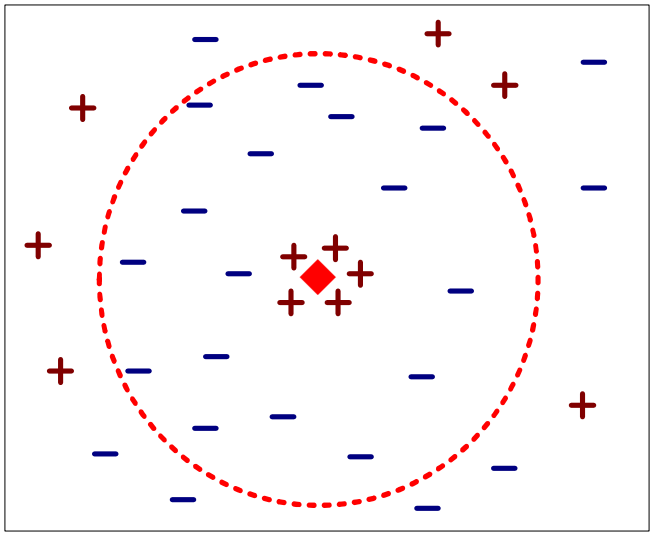
Choice of k

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other

Rule of thumb:

$$k = \sqrt{N}$$

N: number of training points

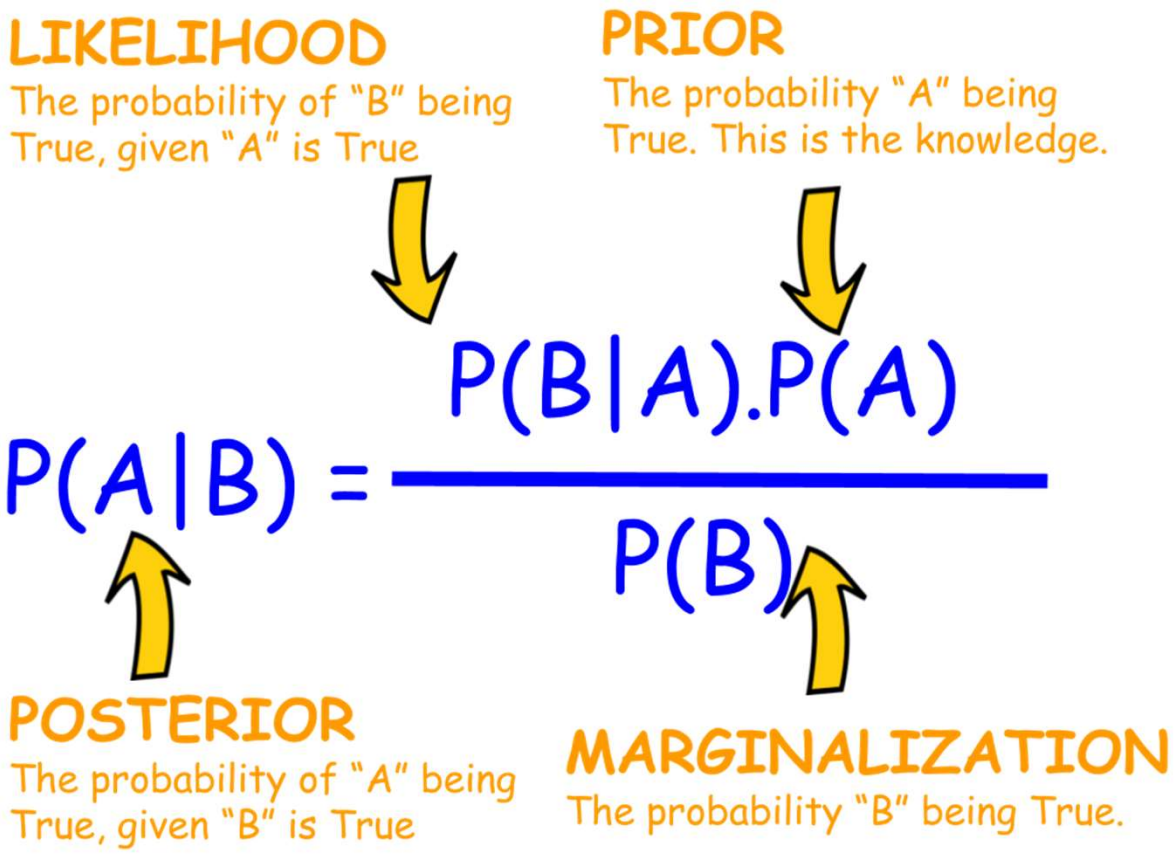


Naïve Bayes

Classification using Bayes Theorem.

CUNY School of Professional Studies

Bayes Theorem



Source: Guide to Intelligent Data Science, Berthold et al

10

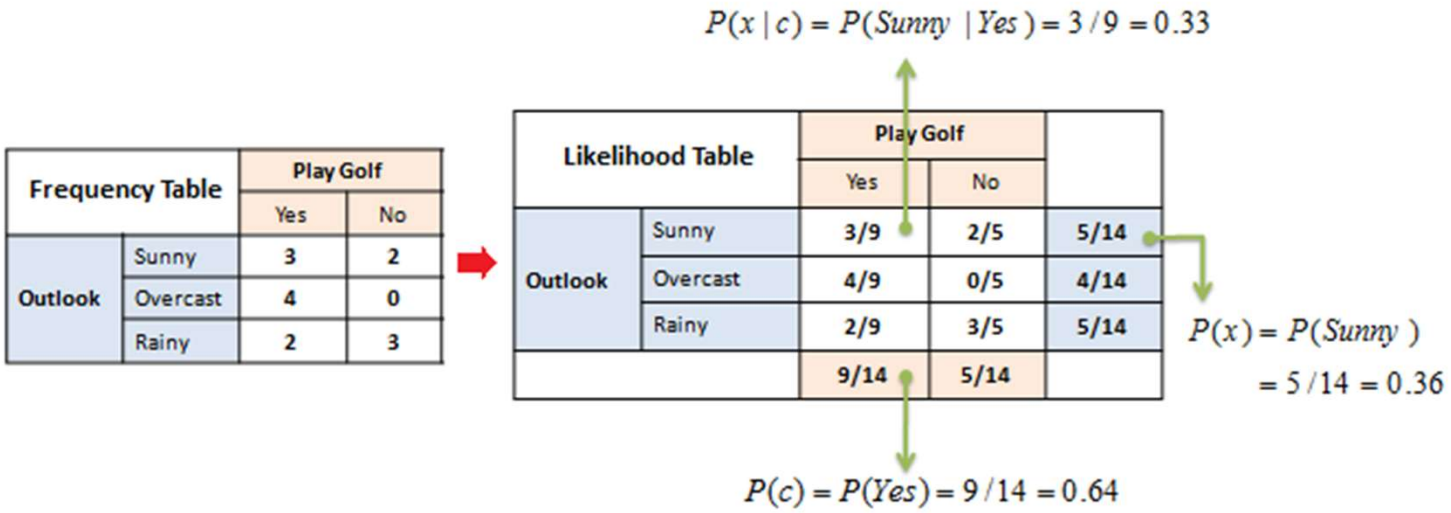
Naïve Bayes: Example

Predicting whether you should play golf

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Naïve Bayes: Example

Let's look at the data:



Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

Source: Saed Sayad

Naïve Bayes: Example

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

		Play Golf	
		Yes	No
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5

Source: Saed Sayad

Naïve Bayes: Example

Will I play golf in the following example?

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes \mid X) = P(Rainy \mid Yes) \times P(Cool \mid Yes) \times P(High \mid Yes) \times P(True \mid Yes) \times P(Yes)$$

$$P(Yes \mid X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No \mid X) = P(Rainy \mid No) \times P(Cool \mid No) \times P(High \mid No) \times P(True \mid No) \times P(No)$$

$$P(No \mid X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

Source: Saed Sayad

14

Naïve Bayes: Strengths and Weaknesses

- Strengths:
  - Simplicity and computational efficiency.
  - It does a great job handling categorical features directly, without any preprocessing.
  - Outperforms more sophisticated classifiers when working with a large number of predictors
  - It handles noisy and missing data pretty well.
- Weaknesses:
  - Needs a sizable amount of data
  - It is naïve: assumption of independence between inputs & classes
  - Doesn't work well for datasets with a large number of continuous features
  - It assumes that all features within a class are not only independent but are equally important

PC1

Source: Guide to Intelligent Data Science, Berthold et al

Naïve Bayes: Use-cases

- Spam detection
- Sentiment analysis (news articles)
- Document classification
- Many classification problems…

16

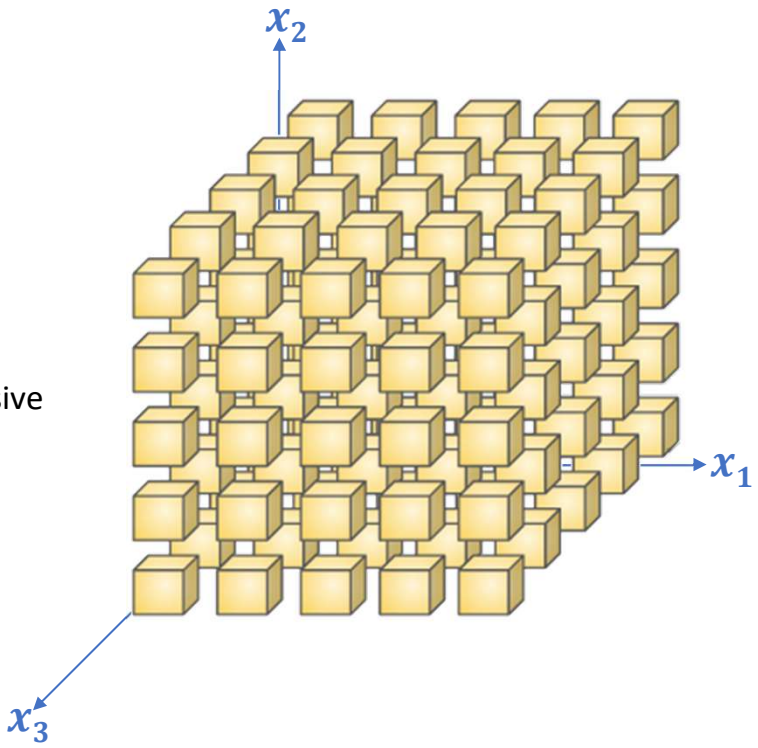
Curse of Dimensionality

As dimensions increase, the data we need to generalize grows exponentially

CUNY School of Professional Studies

Curse of dimensionality

- The Iris data set has 150 instances in 4-dimensions: that is ~3.5 values per dimension!
- Labeled data is hard to get and expensive (about $2/instance on average for outsourced labeling services)



Source: therbootcamp.github.io