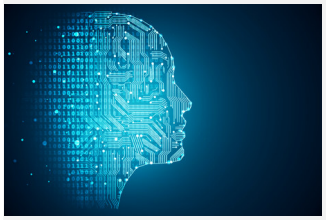
PennState Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

PennState Clinical and Translational Science Institute



Data Science for Researchers and Scholars

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Objectives of the Predictive Modeling Module

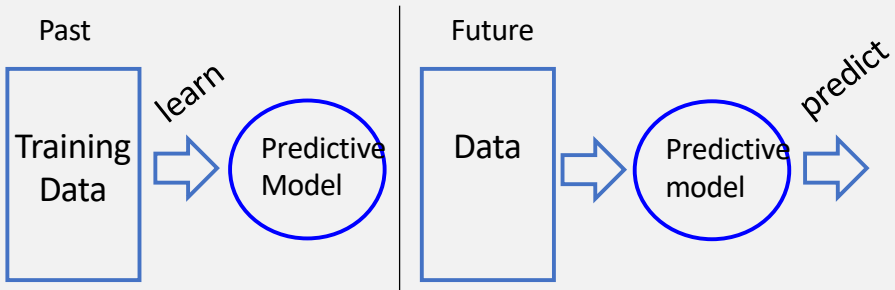
- The predictive modeling module will equip you to:
  - Decide if a data science problem is amenable to an ML solution
  - If so, identify what ML methods might be applicable
  - Understand how ML methods work and when they might fail
  - Use ML to extract knowledge from data
  - Rigorously evaluate ML algorithms
  - Interpret ML trained models
  - Communicate results and any caveats
  - Practice ML responsibly

On a lighter note..

Upon completion of the ML module, you will be able to laugh at these signs, or at least know why one might…

Predictive Modeling Using Machine Learning

Machine learning is about (computationally) predicting the future based on the past



Machine Learning is…

- About methods for detecting patterns in data, and using the uncovered patterns to predict the future
- Concerned with methods for extracting actionable knowledge from data

Why should machines learn?

Practical

- Explicitly programming machines to perform many tasks, e.g., recognizing images, classifying documents, diagnosing diseases, predicting human behavior, etc. is hard, and often infeasible
- If we can get machines to acquire the knowledge needed for a particular task from observations (data) or interactions (experiments), we can
  - Dramatically reduce the cost of developing AI systems
  - Dramatically accelerate knowledge acquisition from data
  - Dramatically accelerate scientific discovery
  - Dramatically improve healthcare, education, public policy, manufacturing, ….
- Machine learning offers some of the most powerful tools for predictive modeling from data

Why should machines learn? – Science of learning

Machine learning offers algorithmic models of learning that can provide useful insights into

- How humans and animals learn
- Information requirements of learning tasks
- The precise conditions under which learning is possible
- Inherent difficulty of learning tasks
- How to improve learning – e.g. value of active versus passive learning
- Computational architectures for learning

Representative applications of machine learning

Scientific applications

- Predicting protein structure and function from sequence
- Constructing gene networks from gene expression time series
- Constructing brain networks from fMRI data
- Elucidating risk factors for cardiovascular disease from EHR data
- Predicting material properties from material composition
- Understanding the formation of social ties
- …

Representative applications of machine learning

Humanities or scholarly applications

- Modeling literary styles
- Modeling musical styles
- Generating music
- Generating text
- Generating art
- Translating language
- Detecting linguistic structure
- Learning grammar

Representative applications of machine learning

Practical applications

- Detecting SPAM
- Determining credit-worthiness
- Recommending products, movies, web pages..
- Targeting advertisements
- Predicting stock prices
- Detecting malware
- Driving cars
- Detecting fraud
- Predicting power consumption

Machine Learning – related disciplines

- Applied Statistics
  - Emphasizes statistical models of data
  - Methods typically applied to small data sets
  - Often done by a statistician increasingly assisted by a computer
  - Often assumes simple model structure
- Machine learning
  - Relies on (often, but not always statistical) inference from data
  - Emphasizes efficient data structures and algorithms
  - Supports use of knowledge to constrain models
  - Assumes flexible model structure
  - Obtaining guarantees regarding the quality of learned models
  - Scalability to large, complex data sets (big data)
- Data Mining – roots in databases
- Pattern recognition – roots in signal and image processing

What is Machine Learning?

- A program *M* is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance as measured by *P* on tasks in *T* in an environment *Z* improves with experience *E*.

Example 1

T – cancer diagnosis

E – a set of diagnosed cases

P – accuracy of diagnosis on new cases

Z – noisy measurements, occasionally misdiagnosed training cases

M – a program that runs on a general purpose computer

What is Machine Learning?

Example 2

T – personalized movie recommendation, e.g., on Netflix

E – movie ratings data from individuals

P – accuracy of predicted movie ratings

10% improvement in prediction accuracy – $1 million prize

What is Machine Learning?

Example 3

T – Predicting protein-RNA interactions

E – A data set of known interactions

P – accuracy of predicted interactions

What is Machine Learning?

Example 4

T – Reconstructing functional connectivity of brains from brain activity (e.g., fMRI) data

E – fMRI data

P – accuracy of the reconstructed network

What is Machine Learning?

Example 5

T – solving integral calculus problems, given rules of integral calculus

E – a set of solved problems

P – score on test consisting of problems not in *E*

What is Machine Learning?

Example 6

T – predicting the risk of a disease before the onset of clinical symptoms

E – longitudinal gut microbiome data coupled with diagnostic tests

P – accuracy of predictions

What is Machine Learning?

Example 7

T – predicting sleep quality from actigraphy data

E – actigraphy data with sleep stage labels

P – accuracy of predictions

What is Machine Learning?

- Example 8
- *T* – Predicting material properties from material composition or material structure
- *E* – Databases of materials – composition, structure, properties
- *P* – accuracy of material property predictions

What is Machine Learning?

Example 9

- *T* – driving a car
- *E* – Observations of driver actions under a broad range of conditions
- *P* – suitable measure of good driving – safety, efficiency, ...

Key requirements of ML

- There are patterns to be learned
- There are data to learn from

Applicant information:

age	23 years
gender	male
annual salary	$30,000
years in residence	1 year
years in job	1 year
current debt	$15,000
...	...

Approve credit?

Learning to approve credit

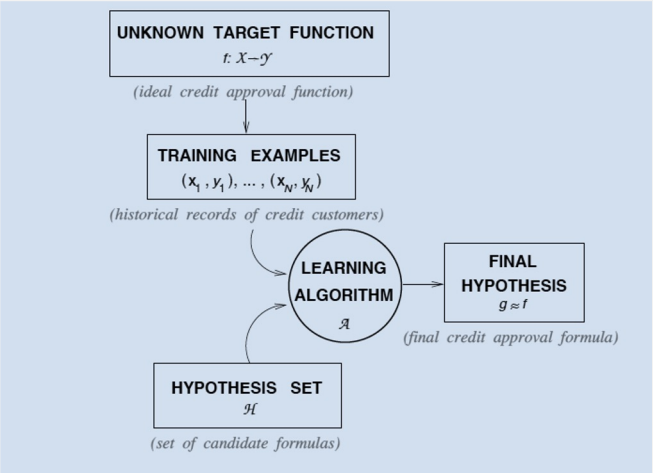
Formalization:

- Input: $\mathbf{x}$ (*customer application*)
- Output: y (*good/bad customer?*)
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ (*ideal credit approval formula*)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ (*historical records*)

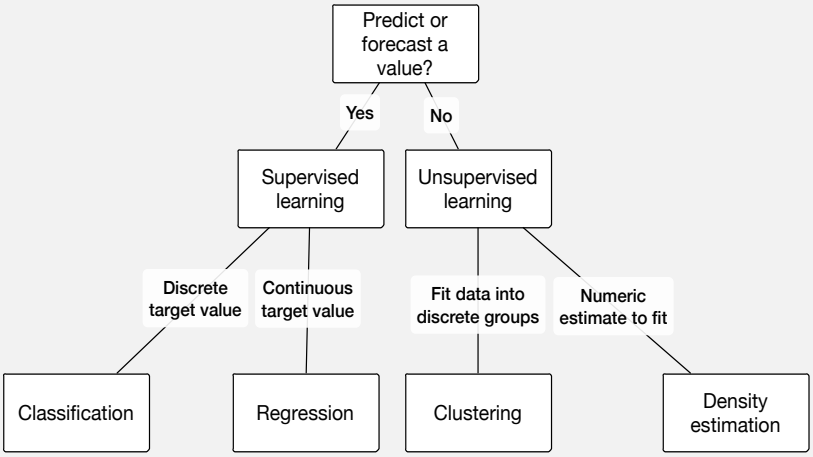
$$\downarrow \quad \downarrow \quad \downarrow$$

- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ (*formula to be used*)

Learning to approve to credit



Flavors of machine learning



Classification

- How would you write a program to distinguish a picture of you from a picture of someone else?
  - Provide examples pictures of you and pictures of other people and let a classifier learn to distinguish the two.
- How would you write a program to determine whether a sentence is grammatical or not?
  - Provide examples of grammatical and ungrammatical sentences and let a classifier learn to distinguish the two.
- How would you write a program to distinguish cancerous cells from normal cells?
  - Provide examples of cancerous and normal cells and let a classifier learn to distinguish the two.

Example: To play or not to play tennis

- **Example dataset**

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

- **Three key elements**
  - Class label ("label", denoted by *y*)
  - Features ("attributes")
  - Feature values ("attribute values", denoted by *x*)

  Feature values can be binary, nominal or continuous
- **A** *labeled* dataset **is a collection of** (*x, y*) **pairs**

Example: To play or not to play tennis?

- **Example dataset**

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

- **Task:**

Class	Outlook	Temperature	Windy?
???	Sunny	Low	No

- **Predict the class of this "test" sample**
- Requires us to **generalize** from the training data

Machine Learning for Classification

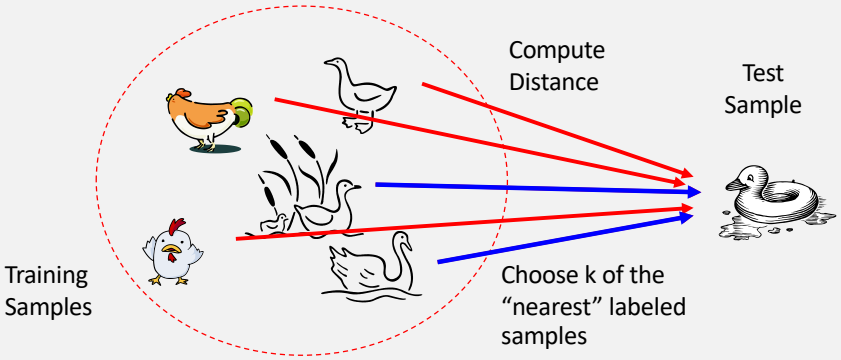
Ingredients for classification

- Idea: Incorporate your knowledge of the problem into a learning system
- Sources of knowledge:
  - Feature representation
    - Crucial for the success of machine learning
    - Can be problem-specific
    - A good representation takes you half way
  - Training data
    - High quality labeled data can be hard to get
    - We may have to get by with the available data
    - Data may be biased for various reasons
  - Model training
    - No single learning algorithm outperforms all others on every task ("no free lunch")
    - Different algorithms have different inductive biases
    - Different algorithms make different assumptions

Can you give me an example of a representation that is trivially bad?
Can you give me an example where data is available for free?

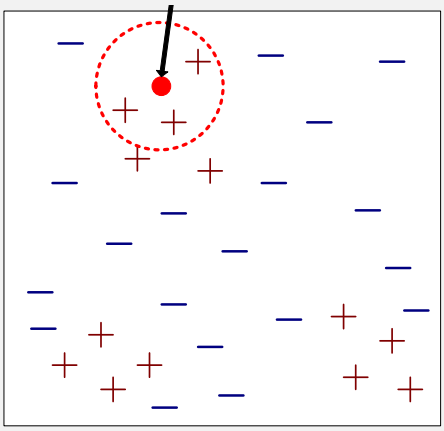
Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



Nearest-Neighbor Classifiers

Query sample

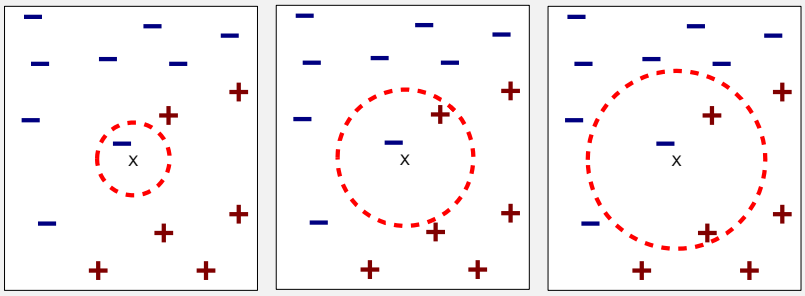


Require three things

- The set of stored training samples and their labels
- Distance Metric to compute distance between samples
- The value of *K*, the number of nearest neighbors to retrieve

- To classify a query sample:
  - Compute distance to training samples
  - Identify *K* nearest neighbors
  - Use class labels of the K nearest neighbors to determine the class label of the query sample (e.g., by taking majority vote)

Definition of Nearest Neighbor



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

K-nearest neighbors of a sample x are data points that have the k smallest distance to x

K nearest neighbor classifier

Data samples are assumed to lie in an n-dimensional space – e.g., the Euclidean space

An instance X is described by a feature vector

$$X_p = [x_{1p} \cdots x_{Np}]$$

Where x_{ip} denotes the value of the ith feature in X_p

$$d(X_p, X_r) = \left(\sum_{i=1}^{N} (x_{ip} - x_{ir})^2 \right)$$

Defines the Euclidean distance between two points in the Euclidean space

Standardization

Standardization can be important when the variables are not all measured on the same scale

- 0-1 scaling

  4, 3, 1 2

  e.g. 3 → (3-min)/(max-min)=(3-1)/(4-1)=2/3
- Z-score scaling: subtract out the mean, divide by std. deviation

K nearest neighbor Classifier

<u>Learning Phase</u>

For each training example $(X_i, f(X_i))$, store the example in memory

<u>Classification phase</u>

Given a query instance X_q, identify the k nearest neighbors $X_1 \ldots X_k$ of X_q

Assign X_q the label of the majority class

$$g(X_q) = \underset{\omega}{argmax} \sum_{i=1}^{K} \delta(\omega, f(X_i))$$ where

$\delta(a,b) = 1$ iff $a = b$ and $\delta(a,b) = 1$.

Distance weighted K nearest neighbor Classifier

<u>Learning Phase</u>

For each training example (X_i, $f(X_i)$), store the example in memory

<u>Classification phase</u>

Given a query instance X_q, identify the k nearest neighbors $of\ X_q$ - $KNN(X_q) = \{X_1 \ldots X_k\}$

And obtain a weighted vote, with each nearest neighbor getting a vote in favor of its class label that is weighted by the distance to the query

$$w_i = \frac{1}{d(X_i, X_q)^2}$$

Distance Measures

- Distance
  - Depends on the data representation
  - Distance measure chosen

An Employee DB

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

Word Frequencies for Documents

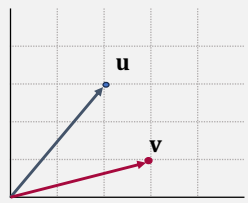
	w1	w2	w3	w4	w5	w6
Doc1	0	4	0	0	0	2
Doc2	3	1	4	3	1	2
Doc3	3	0	0	0	3	0
Doc4	0	1	0	3	0	0
Doc5	2	2	2	3	1	4

Representation has to be chosen with some care
Distance measure should be chosen to work with the representation

A little mathematical detour – Vector Spaces

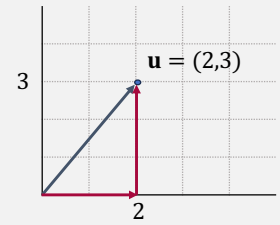
Vectors and vector spaces

- A n-dimensional vector is an ordered n-tuple, e.g., of real numbers $(x_1 \cdots x_n)$
- The set of all such vectors is called a vector space, e.g., $\mathbb{R}^n$
- $(x_1 \cdots x_n)$ can be viewed as a point in an n-dimensional vector space with x_i's as coordinates
- Example:
  - $n = 2$, vector space is $\mathbb{R}^2$
  - $\mathbf{u} = (2,3)$ and $\mathbf{v} = (3,1)$ are vectors
- $\mathbf{u} = 2(1,0) + 3(0,1)$
  - $\mathbf{v} = 3(1,0) + 1(0,1)$



Vectors and vector spaces

- Equality of vectors
  - $\mathbf{u} = \mathbf{v}$ iff $\forall i \quad u_i = v_i$
- Vector negation
  - $-\mathbf{u} = (-u_1, \cdots, -u_n)$
- Sum/difference of two vectors is their coordinate-wise sum/difference
  - $\mathbf{u} + \mathbf{v} = (u_1 + v_1, \cdots, u_n + v_n)$
  - $\mathbf{u} - \mathbf{v} = (u_1 - v_1, \cdots, u_n - v_n)$
- Scalar multiple of a vector $\mathbf{u}$
  - $c\mathbf{u} = (cu_1, \cdots, cu_n)$
- Zero vector
  - $\mathbf{0} = (0, 0, \cdots, 0, 0)$



$\mathbf{u} = 2(1,0) + 3(0,1)$

Properties of vector addition and scalar multiplication

- Let $\mathbf{u}, \mathbf{v}$, and $\mathbf{w}$ be vectors in $\mathbb{R}^n$, and let c and d be scalars.
  - $\mathbf{u} + \mathbf{v}$ is a vector in $\mathbb{R}^n$
  - $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
  - $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$
  - $\mathbf{u} + \mathbf{0} = \mathbf{u}$
  - $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$
  - $c\mathbf{u}$ is a vector in $\mathbb{R}^n$
  - $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$
  - $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$
  - $c(d\mathbf{u}) = (cd)\mathbf{u}$
  - $1(\mathbf{u}) = \mathbf{u}$

Example

Let $\mathbf{u} = (2, -1, 5, 0)$, $\mathbf{v} = (4, 3, 1, -1)$, and $\mathbf{w} = (-6, 2, 0, 3)$ be vectors in $\mathbb{R}^4$. Solve for $\mathbf{x}$ in each of the following.

(a) $\mathbf{x} = 2\mathbf{u} - (\mathbf{v} + 3\mathbf{w})$

(b) $3(\mathbf{x} + \mathbf{w}) = 2\mathbf{u} - \mathbf{v} + \mathbf{x}$

(a)
$$\begin{aligned}
\mathbf{x} &= 2\mathbf{u} - (\mathbf{v} + 3\mathbf{w}) \\
&= 2\mathbf{u} - \mathbf{v} - 3\mathbf{w} \\
&= (4, -2, 10, 0) - (4, 3, 1, -1) - (-18, 6, 0, 9) \\
&= (4 - 4 + 18, -2 - 3 - 6, 10 - 1 - 0, 0 + 1 - 9) \\
&= (18, -11, 9, -8).
\end{aligned}$$

(b) $3(\mathbf{x}+\mathbf{w}) = 2\mathbf{u} - \mathbf{v} + \mathbf{x}$

$$3\mathbf{x} + 3\mathbf{w} = 2\mathbf{u} - \mathbf{v} + \mathbf{x}$$

$$3\mathbf{x} - \mathbf{x} = 2\mathbf{u} - \mathbf{v} - 3\mathbf{w}$$

$$2\mathbf{x} = 2\mathbf{u} - \mathbf{v} - 3\mathbf{w}$$

$$\mathbf{x} = \mathbf{u} - \tfrac{1}{2}\mathbf{v} - \tfrac{3}{2}\mathbf{w}$$

$$= (2,1,5,0) + \left(-2, \tfrac{-3}{2}, \tfrac{-1}{2}, \tfrac{1}{2}\right) + \left(9,-3,0, \tfrac{-9}{2}\right)$$

$$= \left(9, \tfrac{-11}{2}, \tfrac{9}{2}, -4\right)$$

Linear combination of vectors

- The vector $\mathbf{x}$ is called a linear combination of $\mathbf{v}_1, \cdots, \mathbf{v}_m$ if it can be expressed in the form

$$\mathbf{x} = c_1\mathbf{v}_1 + \cdots c_1\mathbf{v}_m$$

Example: Given $\mathbf{x} = (-1, -2, -2), \mathbf{u} = (0,1,4), \mathbf{v} = (-1,1,2)$, and $\mathbf{w} = (3,1,2)$ in $\mathbb{R}^3$, find a, b, and c such that $\mathbf{x} = a\mathbf{u} + b\mathbf{v} + c\mathbf{w}$

$$\begin{aligned} -b + 3c &= -1 \\ a + b + c &= -2 \\ 4a + 2b + 2c &= -2 \end{aligned}$$

$$\Rightarrow a = 1,\ b = -2,\ c = -1$$

Thus $\mathbf{x} = \mathbf{u} - 2\mathbf{v} - \mathbf{w}$

Hence, $\mathbf{x} = \mathbf{u} - 2\mathbf{v} - 1$

Spanning set of vector space

- If every vector in a given vector space can be written as a linear combination of vectors in a given set S, then S is called **a spanning set** of the vector space.

Example

$(1,0,0)$, $(0,1,0)$ and$(0,0,1)$ form a spanning set of vector space $\mathbb{R}^3$

Linear independence

$S = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k\}$ be a set of vectors in a vector space $\mathbf{V}$

Let $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0}$

(1) If the equation has only the trivial solution $(c_1 = c_2 = \cdots = c_k = 0)$ then S is called linearly independent.

(2) If the equation has a nontrivial solution (i.e., not all zeros), then S is called linearly dependent.

Testing for linear independence

Determine whether the following set of vectors in $\mathbb{R}^3$ linearly independent

$$S = \{\underset{\mathbf{v}_1}{(1,2,3)}, \underset{\mathbf{v}_2}{(0,1,2)}, \underset{\mathbf{v}_3}{(-2,0,1)}\}$$

Solution

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0} \quad \Rightarrow \quad \begin{aligned} c_1 \qquad\quad - 2c_3 &= 0 \\ 2c_1 + \ c_2 + \qquad &= 0 \\ 3c_1 + 2c_2 + \ c_3 &= 0 \end{aligned}$$

$$\Rightarrow c_1 = c_2 = c_3 = 0 \quad (\text{only the trivial solution})$$

$$\Rightarrow S \text{ is linearly independent}$$

Length of a vector in $\mathbb{R}^n$

The length of a vector $\mathbf{v} = (v_1, \cdots, v_n)$ in $\mathbb{R}^n$ is given by

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_1^2 \cdots v_n^2}$$

- The length of a vector is also called its norm.
- Notes: Properties of length

(1) $\|\mathbf{v}\| \geq 0$

(2) $\|\mathbf{v}\| = 1 \Rightarrow \mathbf{v}$ is a unit vector.

(3) $\|\mathbf{v}\| = 0$ iff $\mathbf{v} = 0$

(4) $\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$

Euclidian length - Examples

- In $\mathbb{R}^5$, the length of $\mathbf{v} = (0, -2, 1, 4, -2)$ is given by

$$\|\mathbf{v}\| = \sqrt{0^2 + (-2)^2 + 1^2 + 4^2 + (-2)^2} = \sqrt{25} = 5$$

- In R^3 the length of $\mathbf{v} = (\frac{2}{\sqrt{17}}, \frac{-2}{\sqrt{17}}, \frac{3}{\sqrt{17}})$ is given by

$$\|\mathbf{v}\| = \sqrt{\left(\frac{2}{\sqrt{17}}\right)^2 + \left(\frac{-2}{\sqrt{17}}\right)^2 + \left(\frac{3}{\sqrt{17}}\right)^2} = \sqrt{\frac{17}{17}} = 1$$

($\mathbf{v}$ is a unit vector)

p-norm of a vector

- The p-norm of a vector $\mathbf{v}$ (also called the L_p norm) is given by

$$\|\mathbf{v}\|_p = \sqrt[p]{v_1^2 + v_1^2 \cdots v_n^2}$$

Length of a scalar multiple of a vector

Let **v** be a vector in $\mathbb{R}^n$ and c be a scalar. Then

$$\|c\mathbf{v}\| = |c|\,\|\mathbf{v}\|$$

$$\mathbf{v} = (v_1, v_2, \cdots, v_n)$$

$$\Rightarrow c\mathbf{v} = (cv_1, cv_2, \cdots, cv_n)$$

$$\|c\mathbf{v}\| = \|(cv_1, cv_2, \cdots, cv_n)\|$$

$$= \sqrt{(cv_1)^2 + (cv_2)^2 + \cdots + (cv_n)^2}$$

$$= \sqrt{c^2(v_1^2 + v_2^2 + \cdots + v_n^2)}$$

$$= |c|\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

$$= |c|\,\|\mathbf{v}\|$$

Normalizing vector $\mathbf{v}$

If $\mathbf{v}$ is a nonzero vector in $\mathbb{R}^n$, then the vector $\mathbf{u} = \dfrac{\mathbf{v}}{\|\mathbf{v}\|}$

has length 1 and has the same direction as $\mathbf{v}$.

This vector $\mathbf{u}$ is called the unit vector in the direction of $\mathbf{v}$.

The process of finding such a vector is called normalization

$\mathbf{v}$ is nonzero $\Rightarrow \|\mathbf{v}\| \neq 0 \Rightarrow \dfrac{1}{\|\mathbf{v}\|} > 0$

$\Rightarrow \mathbf{u} = \dfrac{1}{\|\mathbf{v}\|}\mathbf{v}$ ($\mathbf{u}$ has the same direction as $\mathbf{v}$)

$\|\mathbf{u}\| = \left\| \dfrac{\mathbf{v}}{\|\mathbf{v}\|} \right\| = \dfrac{1}{\|\mathbf{v}\|}\|\mathbf{v}\| = 1$ ($\mathbf{u}$ has length 1)

Normalizing a vector

Find the unit vector in the direction of $\mathbf{v} = (3, -1, 2)$ and verify that this vector has length 1.

Sol:

$$\mathbf{v} = (3, -1, 2) \quad \Rightarrow \|\mathbf{v}\| = \sqrt{3^2 + (-1)^2 + 2^2} = \sqrt{14}$$

$$\Rightarrow \frac{\mathbf{v}}{\|\mathbf{v}\|} = \frac{(3, -1, 2)}{\sqrt{3^2 + (-1)^2 + 2^2}} = \frac{1}{\sqrt{14}}(3, -1, 2) = \left(\frac{3}{\sqrt{14}}, \frac{-1}{\sqrt{14}}, \frac{2}{\sqrt{14}}\right)$$

$$\because \quad \sqrt{\left(\frac{3}{\sqrt{14}}\right)^2 + \left(\frac{-1}{\sqrt{14}}\right)^2 + \left(\frac{2}{\sqrt{14}}\right)^2} = \sqrt{\frac{14}{14}} = 1$$

$$\therefore \quad \frac{\mathbf{v}}{\|\mathbf{v}\|} \quad \text{is a unit vector.}$$

Distance measures

The distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ is given by $d(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|$

1. Positive definiteness:

   $d(\mathbf{u},\mathbf{v}) \geq 0$ for all $\mathbf{u}$ and $\mathbf{v}$ and

   $d(\mathbf{u},\mathbf{v}) = 0$ only if $\mathbf{u} = \mathbf{v}$

2. Symmetry: $d(\mathbf{u},\mathbf{v}) = d(\mathbf{v},\mathbf{u})$ for all $\mathbf{u}$ and $\mathbf{v}$.

3. Triangle Inequality:

   $d(\mathbf{u},\mathbf{w}) \leq d(\mathbf{u},\mathbf{v}) + d(\mathbf{v},\mathbf{w})$ for all vectors $\mathbf{u}, \mathbf{v}$, and $\mathbf{w}$.

Euclidian distance between two vectors

The distance between $\boldsymbol{u} = (0, 2, 2)$ and $\mathbf{v} = (2, 0, 1)$ is

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| = \|(0-2, 2-0, 2-1)\|$$
$$= \sqrt{(-2)^2 + 2^2 + 1^2} = 3$$

Dot Product of two vectors

The dot product of $\mathbf{u} = (u_1, u_2, \cdots, u_n)$ and $\mathbf{v} = (v_1, v_2, \cdots, v_n)$ is the scalar quantity

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

- Example

  The dot product of **u**=(1, 2, 0, -3) and **v**=(3, -2, 4, 2) is

$$\mathbf{u} \cdot \mathbf{v} = (1)(3) + (2)(-2) + (0)(4) + (-3)(2) = -7$$

Properties of the dot product

- If $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are vectors in R^n and c is a scalar, we have:

(1) $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$

(2) $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$

(3) $c(\mathbf{u} \cdot \mathbf{v}) = (c\mathbf{u}) \cdot \mathbf{v} = \mathbf{u} \cdot (c\mathbf{v})$

(4) $\mathbf{v} \cdot \mathbf{v} = \|\mathbf{v}\|^2$

(5) $\mathbf{v} \cdot \mathbf{v} \geq 0$,and $\mathbf{v} \cdot \mathbf{v} = 0$ if and only if $\mathbf{v} = 0$

Euclidean space

$\mathbb{R}^n$ was defined to be the *set* of all order n-tuples of real numbers. When $\mathbb{R}^n$ is combined with the standard operations of vector addition, scalar multiplication, vector length, and the dot product, the resulting vector space is called Euclidean n-space.

Angle between two vectors

$$\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}, 0 \le \theta \le \pi$$

Opposite direction	$\mathbf{u}\cdot\mathbf{v} < \mathbf{0}$	$\mathbf{u}\cdot\mathbf{v} = \mathbf{0}$	$\mathbf{u}\cdot\mathbf{v} > \mathbf{0}$	Same direction
				
$\theta = \pi$	$\frac{\pi}{2} < \theta < \pi$	$\theta = \frac{\pi}{2}$	$0 < \theta < \frac{\pi}{2}$	$\theta = 0$
$\cos = -1$	$\cos < 0$	$\cos = 0$	$\cos > 0$	$\cos = 1$

- The angle between the zero vector and any other vector is not defined.

Example: Finding the angle between two vectors

$$\mathbf{u} = (-4, 0, 2, -2) \quad \mathbf{v} = (2, 0, -1, 1)$$

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}} = \sqrt{(-4)^2 + 0^2 + 2^2 + (-2)^2} = \sqrt{24}$$

$$\|\mathbf{v}\| = \sqrt{\mathbf{v}\cdot\mathbf{v}} = \sqrt{2^2 + (0)^2 + (-1)^2 + 1^2} = \sqrt{6}$$

$$\mathbf{u}\cdot\mathbf{v} = (-4)(2) + (0)(0) + (2)(-1) + (-2)(1) = -12$$

$$\Rightarrow \cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{-12}{\sqrt{24}\sqrt{6}} = -\frac{12}{\sqrt{144}} = -1$$

$\Rightarrow \theta = \pi \quad \therefore$ $\mathbf{u}$ and $\mathbf{v}$ have opposite directions. $(\mathbf{u} = -2\mathbf{v})$

Orthogonal vectors

Two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ are orthogonal if

$$\mathbf{u} \cdot \mathbf{v} = 0$$

The vector $\mathbf{0}$ is said to be orthogonal to every vector.

Finding orthogonal vectors

- Determine all vectors in $\mathbb{R}^n$ that are orthogonal to $\mathbf{u} = (4, 2)$.

$$\mathbf{u} = (4\,,2) \quad \text{Let} \quad \mathbf{v} = (v_1\,, v_2)$$

$$\Rightarrow \quad \mathbf{u}\cdot\mathbf{v} = (4\,,2)\cdot(v_1\,, v_2)$$

$$= 4v_1 + 2v_2$$

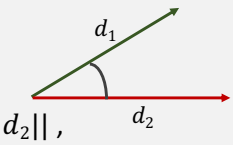
$$= 0$$

$$\begin{bmatrix} 4 & 2 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \frac{1}{2} & 0 \end{bmatrix}$$

$$\Rightarrow \quad v_1 = \frac{-t}{2} \quad , \quad v_2 = t$$

$$\therefore \quad \mathbf{v} = \left(\frac{-t}{2}, t\right), \qquad t \in R$$

Cosine Distance



If d_1 and d_2 are two document vectors, then

$$1 - \cos(\, d_1, d_2\,) \;=\; 1 - (d_1 \bullet d_2) \,/\, ||d_1||\; ||d_2||\,,$$

where $\bullet$ indicates vector dot product and $||\, d\, ||$ is the length of vector d.

Example:

$d_1 \;=\; 3\;2\;0\;5\;0\;0\;0\;2\;0\;0$

$d_2 \;=\; 1\;0\;0\;0\;0\;0\;0\;1\;0\;2$

$$d_1 \bullet d_2 = \; 3\times1 \,+\, 2\times0 \,+\, 0\times0 \,+\, 5\times0 \,+\, 0\times0 \,+\, 0\times0 \,+\, 0\times0 \,+\, 2\times1 \,+\, 0\times0 \,+\, 0\times2 \;=\; 5$$

$$||d_1|| \;=\; (3\times3 + 2\times2 + 0\times0 + 5\times5 + 0\times0 + 0\times0 + 0\times0 + 2\times2 + 0\times0 + 0\times0)^{0.5}$$

$$=\; (42)^{\,0.5} \;=\; 6.481$$

$$||d_2|| \;=\; (6)^{\,0.5} \;=\; 2.245$$

$$\cos(\, d_1, d_2\,) \;= 0.3150$$

Distance Measures

Distances in vector spaces

- Euclidean distance $\sqrt{\sum_{j=1}^{n}(p_j - q_j)^2}$
- Minkowski distance
  - a generalization of Euclidean distance
  - $\sqrt[p]{\sum_{j=1}^{n}|p_j - q_j|^p}$

Distance measures in Boolean spaces

- $p = 1$ Manhattan distance
- $p = 2$ Euclidean distance

Distance measures for data with nominal attributes

- Nominal attributes can take 2 or more values, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Simple matching – distance between two objects is simply the number of mismatched attributes divided by the total number of attributes
- One hot encoding – Encode each M-valued nominal attribute an M-bit vector

  Red: 1 0 0 0, Yellow: 0 1 0 0; Blue: 0 0 1 0 …
- Use distance measures designed for vectors …

Distance measures

- *d (p, q)* between two points *p* and *q* is a proper distance measure if it satisfies:

  1. Positive definiteness:

  $d(p, q) \geq 0$ for all p and q and

  $d(p, q) = 0$ only if $p = q$.

  2. Symmetry: $d(p, q) = d(q, p)$ for all p and q.

  3. Triangle Inequality:

  $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r.

Cosine Distance

- If d_1 and d_2 are two document vectors, then

$$1\text{-}\cos(d_1, d_2) = 1-(d_1 \bullet d_2) / ||d_1||\ ||d_2||,$$

where $\bullet$ indicates vector dot product and $||d||$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3x1 + 2x0 + 0x0 + 5x0 + 0x0 + 0x0 + 0x0 + 2x1 + 0x0 + 0x2 = 5

$||d_1||$ = $(3x3+2x2+0x0+5x5+0x0+0x0+0x0+2x2+0x0+0x0)^{0.5}$ = $(42)^{0.5}$ = 6.481

$||d_2||$ = $(6)^{0.5}$ = 2.245

$\cos(d_1, d_2)$ = .3150

Distance Measures

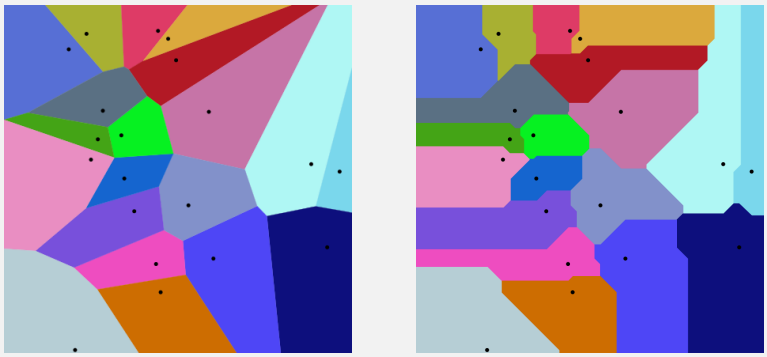
Distances in vector spaces

- Euclidean distance $\sqrt{\sum_{j=1}^{d}(p_j - q_j)^2}$
- Minkowski distance
  - a generalization of Euclidean distance
  - $\sqrt[n]{\sum_{j=1}^{d}|p_j - q_j|^n}$

Distance measures in Boolean spaces

- *n*=1 Manhattan distance
- n=2 Euclidean distance

Decision Boundary of the 1 NN classifier



Manhattan distance Euclidian distance

Query points in the polygon surrounding the training data point are closer to it than any other training data point

Image source: Wikipedia

P-spectrum similarity for sequences over an alphabet

- The p-spectrum of a string is the histogram – vector of number of occurrences of all possible contiguous substrings – of length p
- We can define a similarity function $K(s, t)$ over $\sum^* \times \sum^*$ as the inner product of the p-spectra of s and t.

$$s = statistics$$

$$t = computation$$

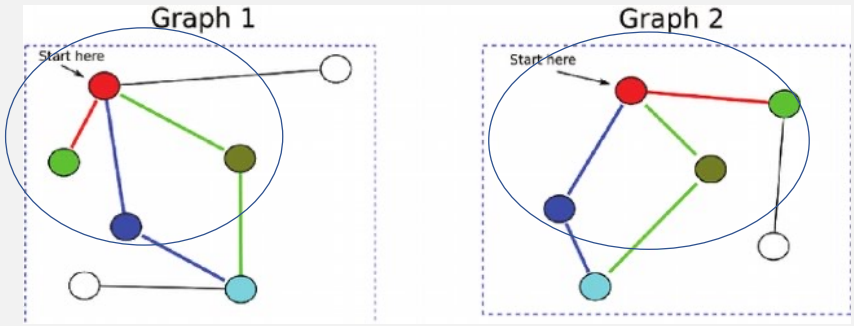
$$p = 3$$

$$\text{Common substrings}: \ tat, ati$$

$$K(s,t) = 2$$

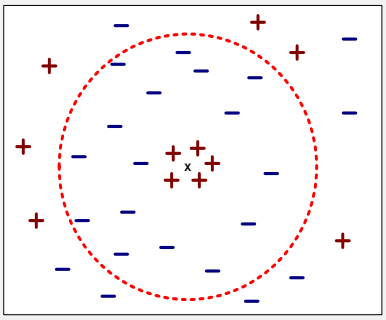
Can you think of a similarity function for graphs?

- Two graphs are similar if their *k*-hop neighborhoods are similar.



Nearest Neighbor Classification…

- Choosing the value of *k*:
  - If *k* is too small, the model can be sensitive to noise
  - If *k* is too large, neighborhood may include too many samples from other classes

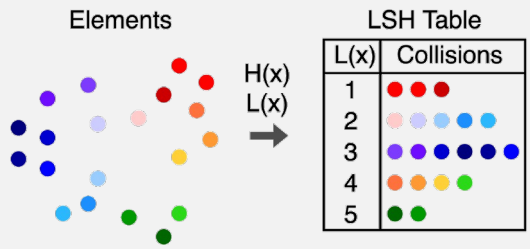


Nearest neighbor classifiers

- Nearest neighbor classifiers are conceptually simple
- Learn by simply memorizing the training data
- The computational effort of learning is low
- The storage requirements of learning is high
  - need to memorize the training data
- Cost of classifying new instances can be high
  - Use efficient data structures and algorithms for nearest neighbor search, e.g., locality sensitive hashing
- A distance measure needs to be defined over the input space
- Performance degrades when
  - Dimensionality increases
  - The number of irrelevant attributes increases
  - The attributes are highly correlated
  - Need to perform feature selection or dimensionality reduction

Hashing and nearest neighbor search

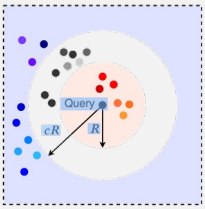
- A *locality sensitive* hash (LSH) function L(x) tries to map similar objects to the same 'hash bin' and dissimilar objects to different bins.
- A hash collision occurs when two objects x and y have the same hash value.
- Under an LSH function, the collision probability depends on how similar the objects are.



Hashing and nearest neighbor search

We say that a hash family $\mathcal{H}$ is (R, cR, p_1, p_2)-sensitive with respect to a distance function $d(x, y)$ if for any $h \in \mathcal{H}$ we have that

- If $d(x, y) \leq R$ then $\Pr_{\mathcal{H}}[h(x) = h(y)] \geq p_1$
- If $d(x, y) \geq cR$ then $\Pr_{\mathcal{H}}[h(x) = h(y)] \leq p_2$



- Select k random vectors.
- Compute the dot product of the data sample $\mathbf{x}$ with each random vector.
- Set the the k'th bit of the hash to 1 if the dot product of x with the random vector is ≥ 0 otherwise set it to 0
- The larger the k, the more accurate the similarity

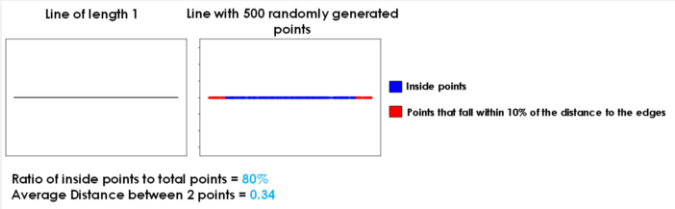
x	1	0	0	0	1	1	0	1	1	1	0	1
y	1	1	0	0	0	1	0	0	1	1	0	1

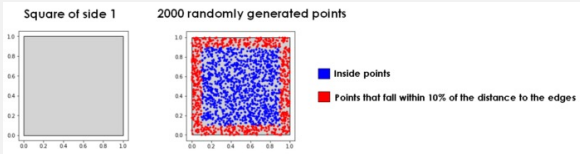
$h_2(x) = 0 \quad h_7(x) = 1$

$h_2(y) = 0 \quad h_7(y) = 0$

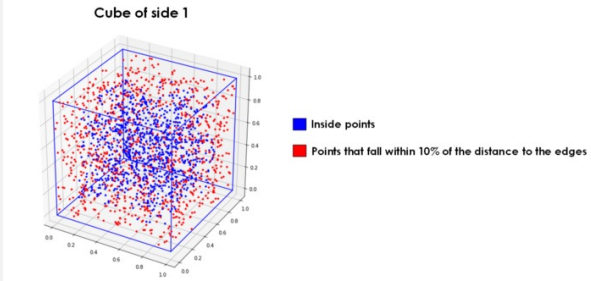
Similarity measures in high dimensions

- **As we increase the number of Dimensions, our data becomes more sparse** (the "volume" of the space increases exponentially with the number of dimensions)
- As we increase the dimensions of our data, the average similarity between pairs of data points decreases.
- In the limit, the average similarity between the closest points approaches the average similarity between the farthest points.





Ratio of inside points to total points = 63%
Average Distance between 2 points = 0.52



Ratio of inside points to total points = 51%
Average Distance between 2 points = 0.65

Function approximation (Regression)

- Function approximation is like classification except the labels are real valued

Example applications:

Predicting

- Stock value
- Income
- Power consumption

K nearest neighbor Function Approximator

Learning Phase

For each training example $(X_i, f(X_i))$, store the example in memory

Approximation phase

Given a query instance X_q, identify the k nearest neighbors $X_1 \ldots X_k$ of X_q

$$g(X_q) \leftarrow \frac{\sum_{l=1}^{K} f(X_i)}{K}$$

Value of a function (e.g., price of a product) at a query point is simply the average or inverse distance weighted average of the value of the function at the k nearest neighbors of the query point