CS626: Speech, Natural Language Processing and the Web

Named Entity Identification and Recognition- Classification and Sequence Labelling

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

Week 6 of 30th August, 2021

Two kinds of insights needed

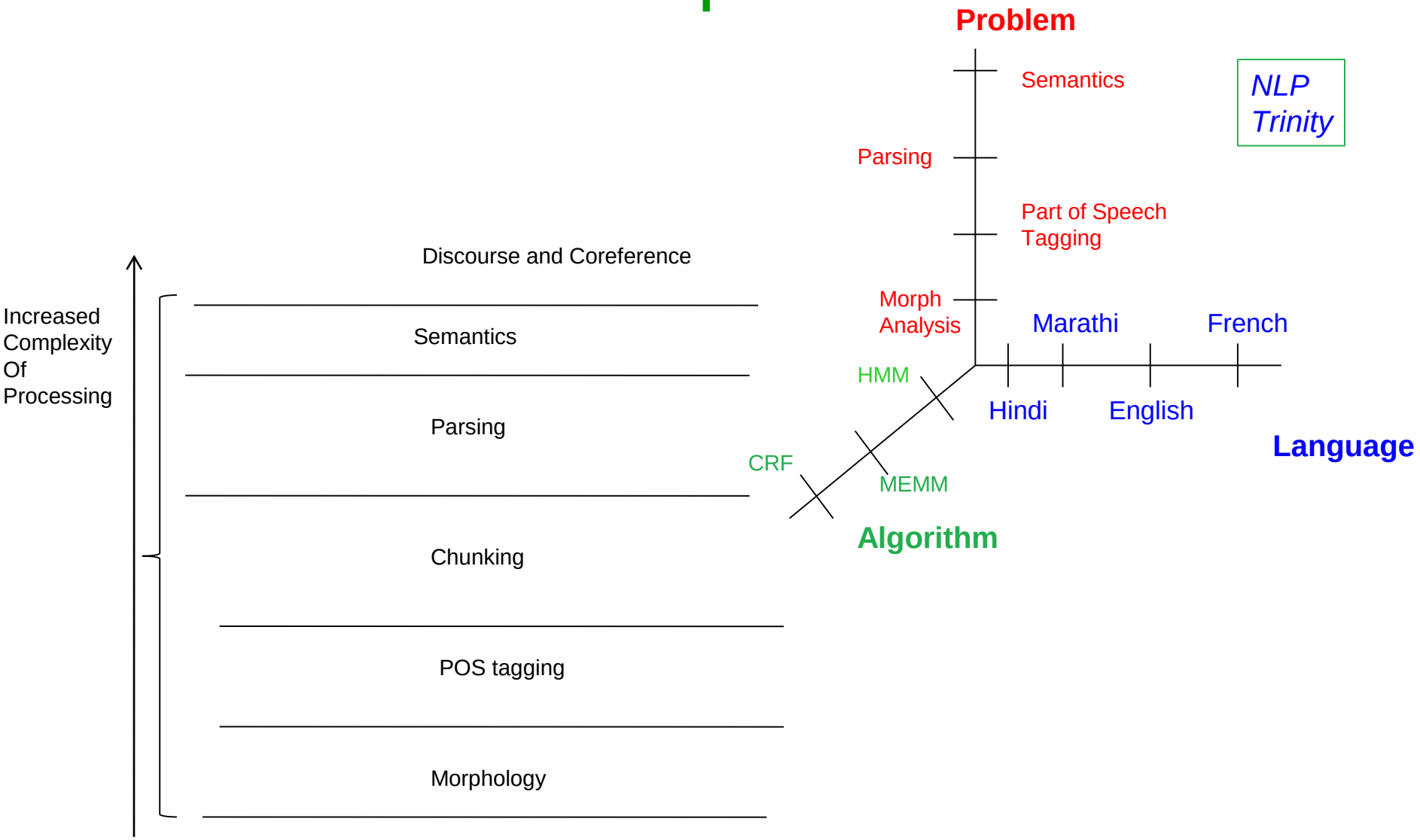
- How language operates
- Linguistics
- Properties of the PROBLEM

- How machine learning operates
- Sense of Distributions
- Properties of Data

Task *vs.* Technique Matrix

Task (row) vs. Technique (col) Matrix	**Rules Based/Knowledge-Based**	**Classical ML**				**Deep Learning**		
		Perceptron	Logistic Regression	SVM	Graphical Models (HMM, MEMM, CRF)	Dense FF with BP and softmax	RNN-LSTM	CNN
Morphology								
POS								
Chunking								
Parsing								
NER, MWE								
Coref								
WSD								
Machine Translation								
Semantic Role Labeling								
Sentiment								
Question Answering								

NLP Perspective



Inherent resilience of the structure called LANGUAGE

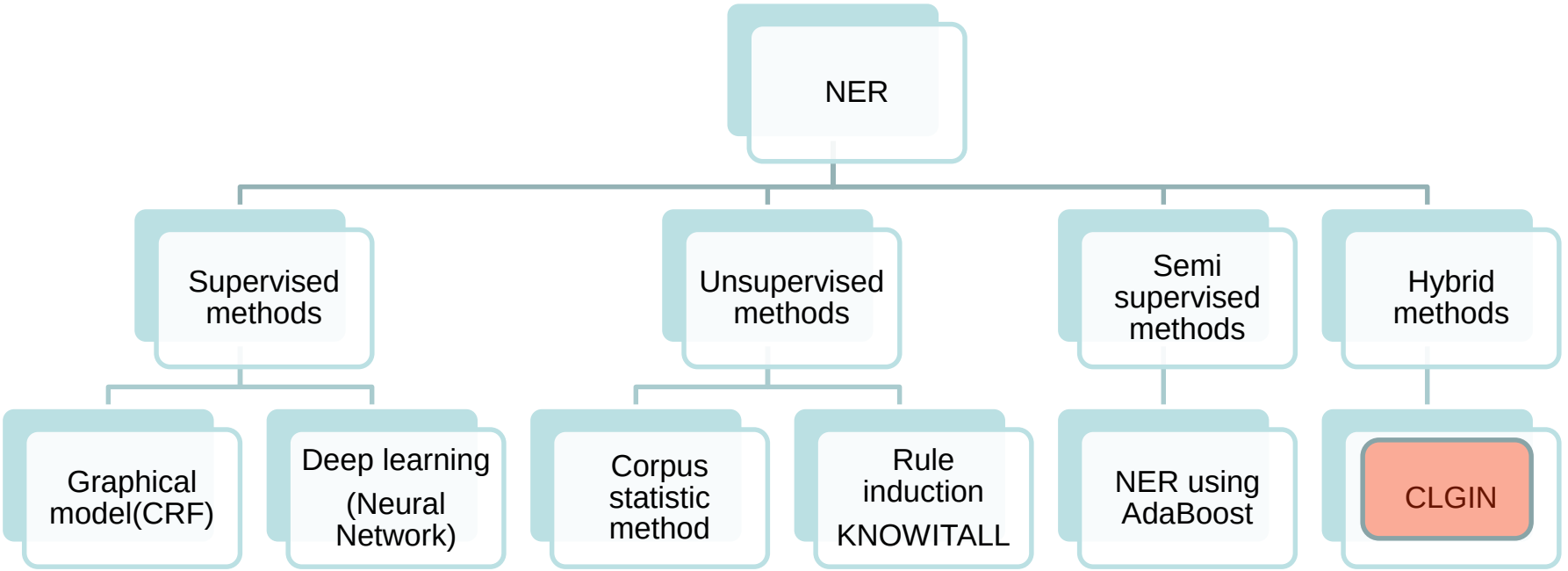
- Example: Apple increased its laptop production
- I know that Apple increased its laptop production
- I know apples are costly (fruit apple): plural ‘s’ disambiguates
- An apple a day keeps the doctor away: article disambiguates
- I bought an Apple for my accounting work: capital ‘A’ disambiguates
- Vulnerability: He has an apple: looseness in capitalization makes disambiguation impossible
- Some insight into how language operates

Multilingual Named Entity Recognition

- If a named entity is recognized in one language, it can be added to lexicon (gazetteer list) and used for processing tasks in other languages

- *Extract Once, Use Many times*

Approaches


NER
Supervised methods
Unsupervised methods
Semi supervised methods
Hybrid methods
Graphical model(CRF)
Deep learning (Neural Network)
Corpus statistic method
Rule induction KNOWITALL
NER using AdaBoost
CLGIN

Background: Information Extraction

Definition: Information Extraction

- To extract information that fits pre-defined schemas or templates
- IE Definition
  - Entity: an object of interest such as a person or organization
  - Attribute: A property of an entity such as name, type
  - Relation: A relationship that holds between two or more entities such as Position of Person in a Company

Information Extraction

- **Note** the difference between Named Entity Identification and Named Entity Recognition
- **Named Entity Identification** is a binary classification problem which classifies whether a given token is a named entity or not
- **Named Entity Recognition** involves detection and categorization of named entities

Goal of IE

As a task: **Filling slots from text- unstructured→structured**

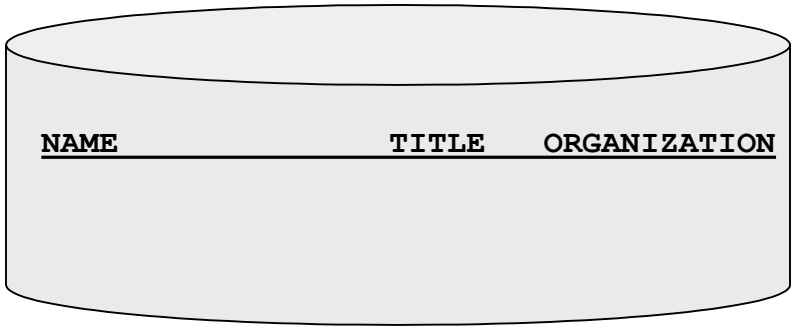
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



Courtesy of William W. Cohen

Unstructured→structured

As a task:

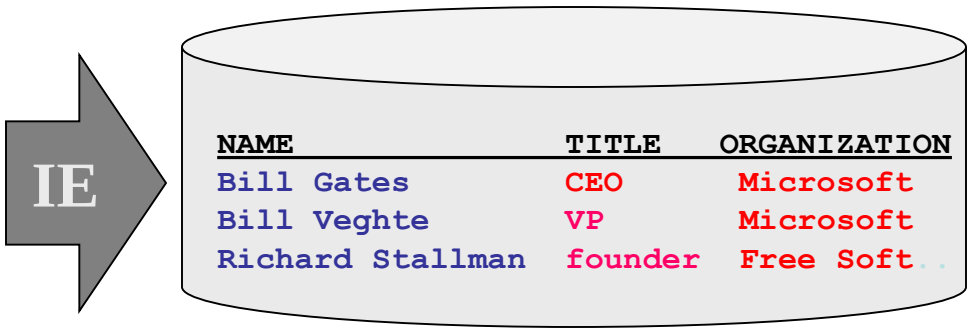
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



Courtesy of William W. Cohen

Named Entity Identification (NEI) and Named Entity Recognition (NER)

Definition

NERC – Named Entity Recognition and Classification (NERC) involves identification of proper names in texts, and classification into a set of pre-defined categories of interest as:

- Person names (names of people)
- Organization names (companies, government organizations, committees, etc.)
- Location names (cities, countries etc)
- Miscellaneous names (Date, time, number, percentage, monetary expressions, number expressions and measurement expressions)

NEI and NER

- **Note** the difference between Named Entity Identification and Named Entity Recognition
- **Named Entity Identification** is a binary classification problem which classifies whether a given token is a named entity or not
- **Named Entity Recognition** involves detection and categorization of named entities

Challenge of NEI and NER

- Variation of NEs – e.g. Prof. Manning, Chris Manning, Dr Chris Manning
- Ambiguity of NE types:
  - Washington (location vs. person)
  - May (person vs. month)
  - Ford (person vs organization)
  - 1945 (date vs. time)
- Ambiguity with common words, e.g. “Kabita“

  - Name of person vs. poem

More complex problems in NER

- Issues of style, structure, domain, genre etc.
- Punctuation, spelling, spacing, formatting, ... all have an impact:

Dept. of Computing and Maths

Manchester Metropolitan University

Manchester

United Kingdom

Many to Many Relationship

- The many-many relationship between entities and their names makes the task of NER more complex

	NAMES			
ENTITIES	E353	Manning	Prof_Manning	Chris_Manning
	E201	Oxygen	O_2	
	E356	Kolkata	Calcullta	West Bengal Capital
	E404	IIT	Indian_Institute_of_ Technology	Indian_Institute_of_ Tech.

- Rows are labelled with **entity ids** and columns contain **names**

E.g.: E353 to represent the specific person entity ‘Chris Manning'

What would be skyline NER performance

- Human performance is considered to be the ultimate goal to be reached by the m/c.
- Human performance is measured by IAA (Inter Annotator Agreement)
- The connection between human performance and IAA is a subtle one
- Still IAA can indirectly provide a skyline
- E.g., WSD IAA is around 85% which translates to skyline performance of percentage in the vicinity of 80s
- Multiplication: humans are no match for machines
- Similarly Question Answering looms as a falling frontier for humans (e.g. Jeopardy contest → WATSON)

Applications

- Intelligent document access
  - Browse document collections by the entities that occur in them
  - Application domains:
    - News
    - Scientific articles, e.g, MEDLINE abstracts
- Information retrieval and extraction
  - Augmenting a query given to a retrieval system with NE information, more refined information extraction is possible
  - For example, if a person wants to search for document containing ‘kabiTA’ as a proper noun, adding the NE information will eliminate irrelevant documents with only ‘kabiTA’ as a common noun

Applications

- Machine translation
  - NER plays an important role in translating documents from one language to other
  - Often the NEs are transliterated rather than translated
  - For example, ‘yAdabpur bishvabidyAlaYa’ → ‘Jadavpur University’
- Automatic Summarization
  - NEs given more priorities in deciding the summary of a text
  - Paragraphs containing more NEs are most likely to be included into the summary

Applications

- Question-Answering Systems
  - NEs are important to retrieve the answers of particular questions (Who is the PM of India/Which country is Modi PM of?)
- Speech Related Tasks
  - NER is important for identifying the number format, telephone number and date format
  - In speech rhythm- necessary to provide a short break after the name of person
  - Solving Out Of Vocabulary words is important in speech recognition

Corpora, Annotation

- MUC-6 and MUC-7 corpora - English
- CONLL shared task corpora
  - http://cnts.uia.ac.be/conll2003/ner/ : NEs in English and German
  - http://cnts.uia.ac.be/conll2002/ner/ : NEs in Spanish and Dutch
- ACE – English - http://www.ldc.upenn.edu/Projects/ACE/
- TIDES surprise language exercise (NEs in Hindi)
- NERSSEAL shared task- NEs in Bengali, Hindi, Telugu, Oriya and Urdu (http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5)

Corpora, Annotation

- Biomedical and Biochemical corpora
  - BioNLP-04 shared task
  - BioCreative shared tasks
  - AiMed

Tag set

Text is tagged (1/2)

Identifying and classifying elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Text is tagged (2/2)

Example:

Jim bought 300 shares of ABC Corp. in 2006.

<ENAMEX TYPE="PERSON"> Jim </ENAMEX> bought <NUMEX

TYPE = "QUANTITY"> 300 </NUMEX> shares of <ENAMEX TYPE =

"ORGANIZATION"> ABC Corp. </ENAMEX> in <TIMEX TYPE = "

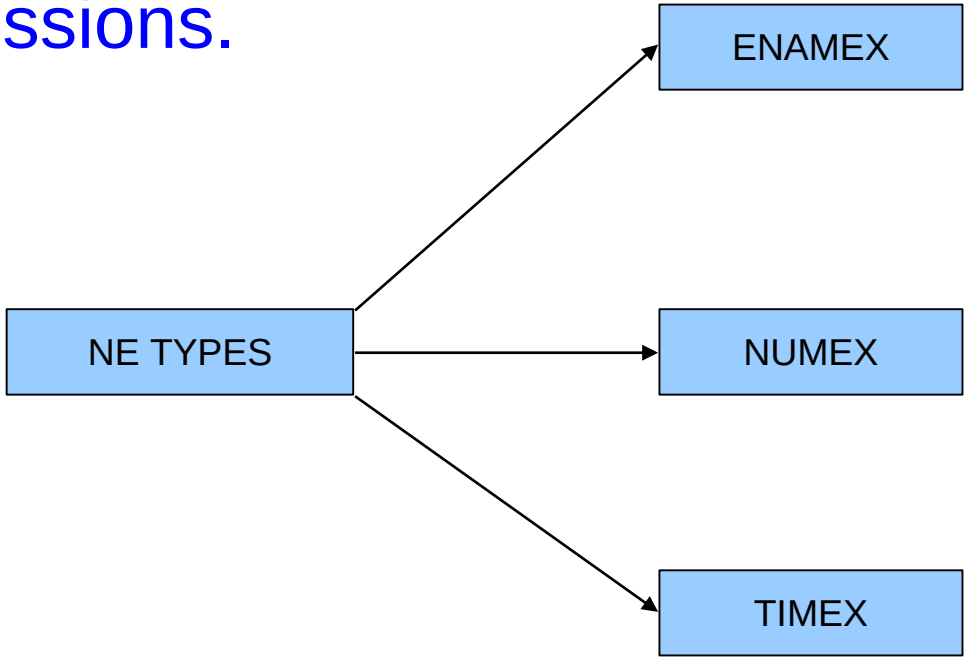
DATE "> 2006 </TIMEX>.

MUC TAGSET

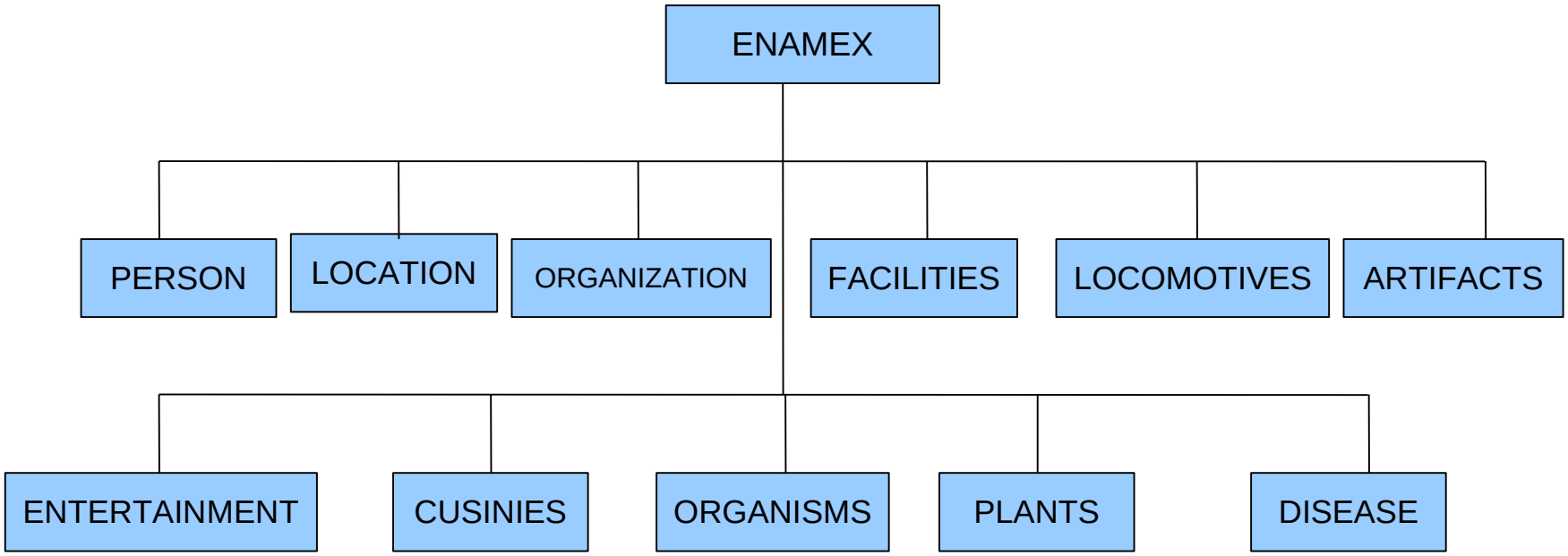
The Named entity hierarchy contains 106 tags

It is divided into three major classes Entity Name, Time and Numerical

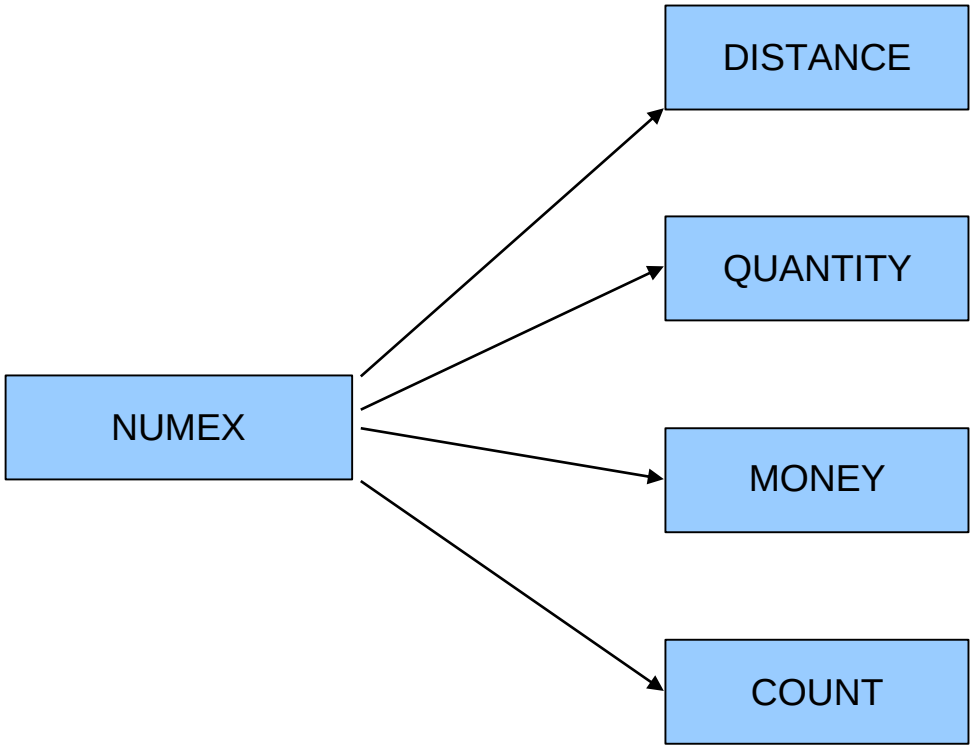
expressions.



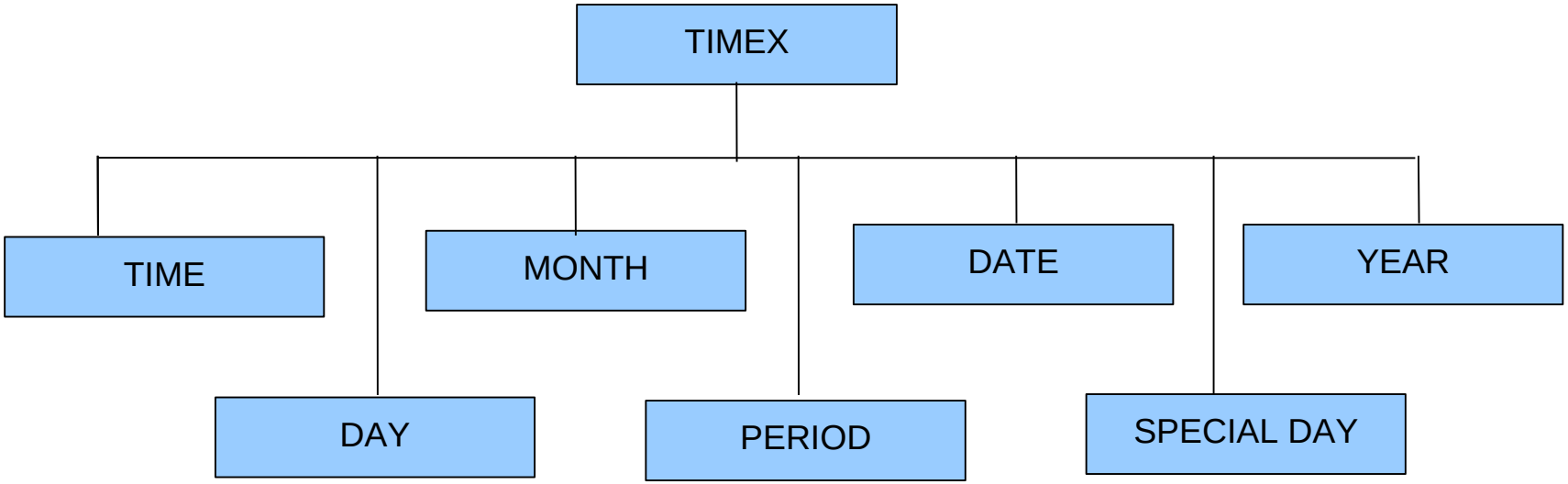
ENAMEX TYPES



NUMEX TYPES



TIMEX Types



PERSON

- Person
  - Individual
    - Family name
    - Title
  - Group

- Persons are entities limited to humans. A person may be a single individual or a group.

- Individual refer to names of each individual person, also includes names of fictional characters found in stories/novels etc.

- Individual name occurs with family name
- an individual is often referred with a title such as Mr., Mrs., Ms., Dr., etc along with their name

- GROUP refers to set of individual

PERSON: Example

Mr. Chandrababu Naidu is the President of Telugu Desam Party

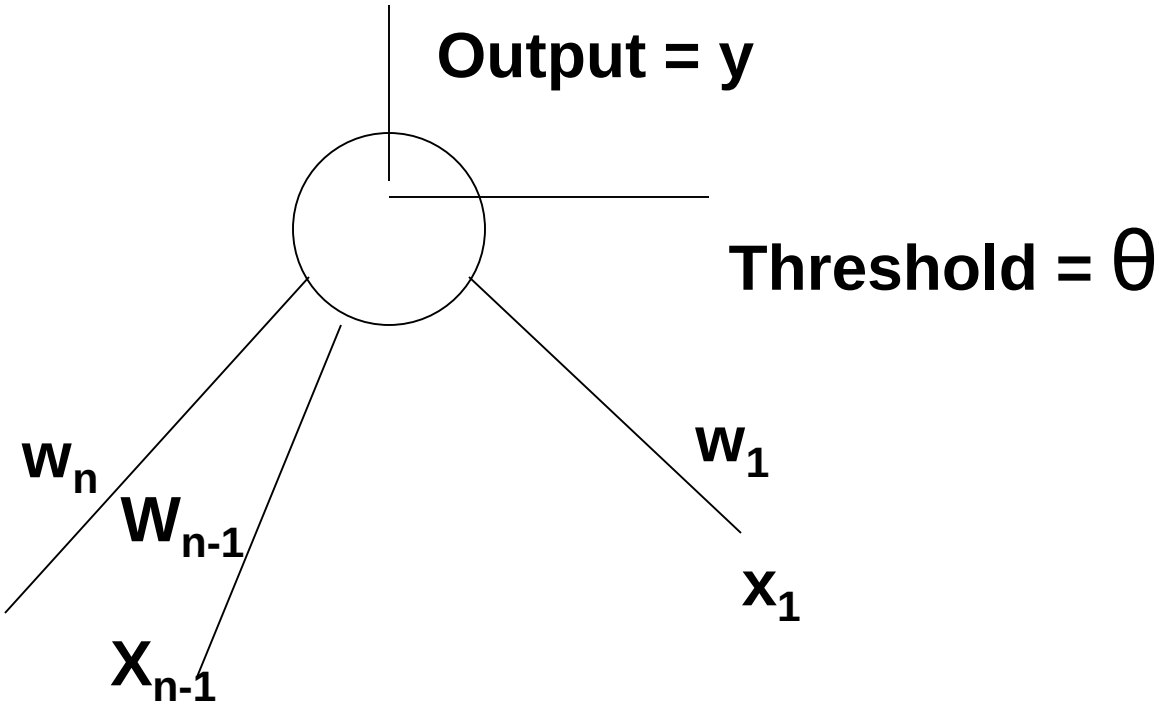
Chandrababu : Person => Individual
Naidu : Person => Family Name
Mr. : Person => Title

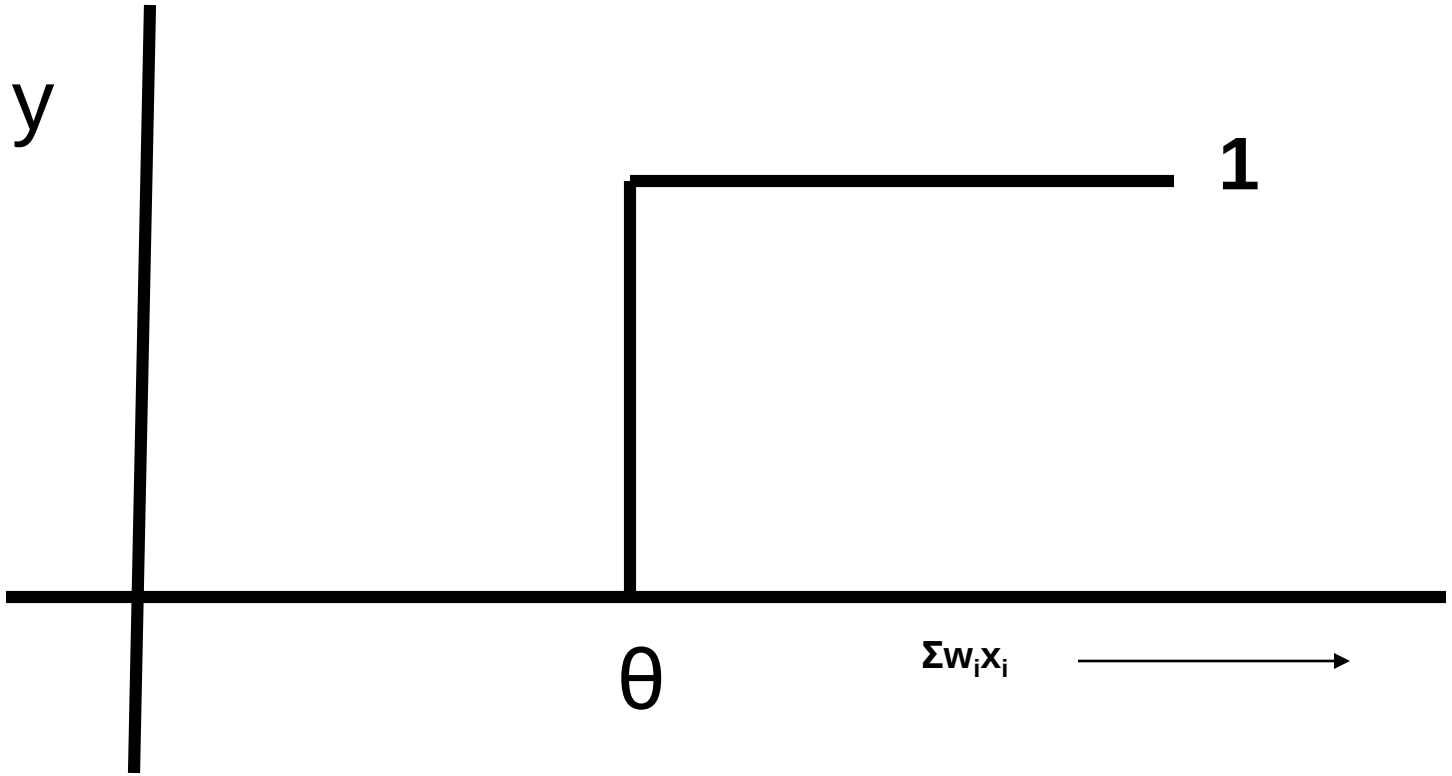
2)Apolo Hospital doctors : Person => GROUP

First example of Linear Separator: Perceptron

The Perceptron Model

A perceptron is a computing element with input lines having associated weights and the cell having a threshold value. The perceptron model is motivated by the biological neuron.





Step function / Threshold function

y = 1 for Σwixi >=θ

=0 otherwise

Features of Perceptron

- **Input output behavior is discontinuous and the derivative does not exist at $\Sigma w_i x_i = \theta$**

- **$\Sigma w_i x_i - \theta$ is the net input denoted as net**

- **Referred to as a linear threshold element - linearity because of x appearing with power 1**

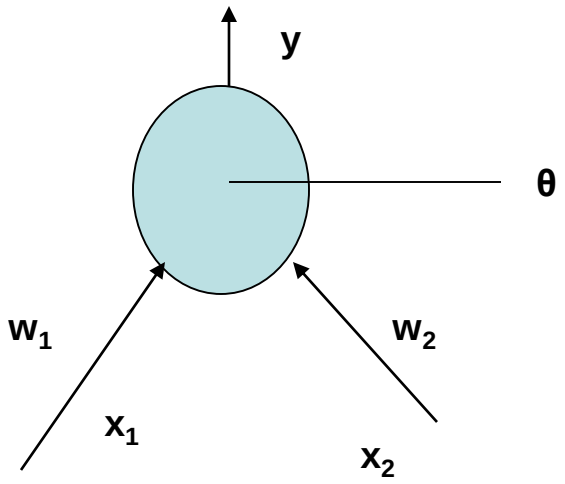
- **y= f(net): Relation between y and net is non-linear**

Computation of Boolean functions

AND of 2 inputs

X1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1

The parameter values (weights & thresholds) need to be found.



Computing parameter values

w1 * 0 + w2 * 0 <= θ ➔ θ >= 0; since y=0

w1 * 0 + w2 * 1 <= θ ➔ w2 <= θ; since y=0

w1 * 1 + w2 * 0 <= θ ➔ w1 <= θ; since y=0

w1 * 1 + w2 *1 > θ ➔ w1 + w2 > θ; since y=1

w1 = w2 = = 0.5

satisfy these inequalities and find parameters to be used for computing AND function.

Other Boolean functions

- **OR can be computed using values of w1 = w2 = 1 and = 0.5**

- **XOR function gives rise to the following inequalities:**

w1 * 0 + w2 * 0 <= θ ➔ θ >= 0

w1 * 0 + w2 * 1 > θ ➔ w2 > θ

w1 * 1 + w2 * 0 > θ ➔ w1 > θ

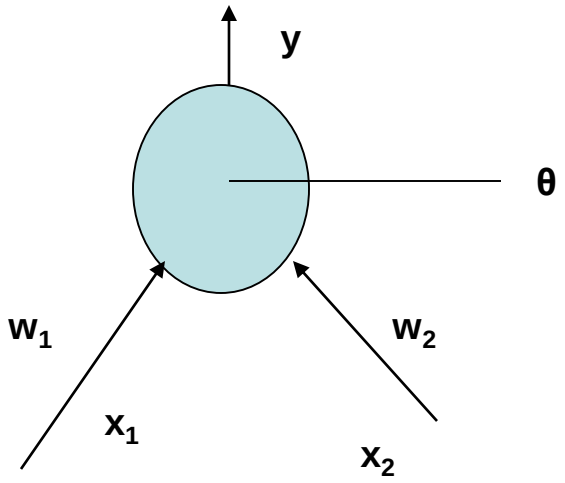
w1 * 1 + w2 *1 <= θ ➔ w1 + w2 <= θ

No set of parameter values satisfy these inequalities.

AND of 2 inputs

X1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1

The parameter values (weights & thresholds) need to be found.



Constraints on w1, w2 and θ

w1 * 0 + w2 * 0 <= θ ➔ θ >= 0; since y=0

w1 * 0 + w2 * 1 <= θ ➔ w2 <= θ; since y=0

w1 * 1 + w2 * 0 <= θ ➔ w1 <= θ; since y=0

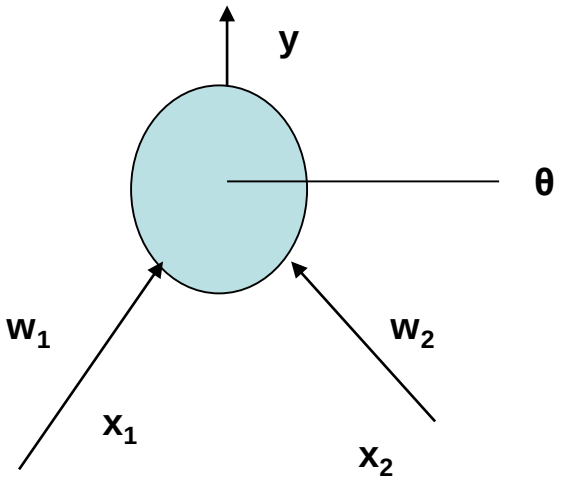
w1 * 1 + w2 *1 > θ ➔ w1 + w2 > θ; since y=1

w1 = w2 = = 0.5, θ=0.9

These inequalities are satisfied by ONE particular region

Non Linearly Separable: XOR

X1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0



The parameter values (weights & thresholds) need to be found.

Constraints on w1, w2 and θ

w1 * 0 + w2 * 0 <= θ ➔ θ >= 0; since y=0

w1 * 0 + w2 * 1 < θ ➔ w2 < θ; since y=1

w1 * 1 + w2 * 0 < θ ➔ w1 < θ; since y=1

w1 * 1 + w2 *1 <= θ ➔ w1 + w2 <= θ; since y=0

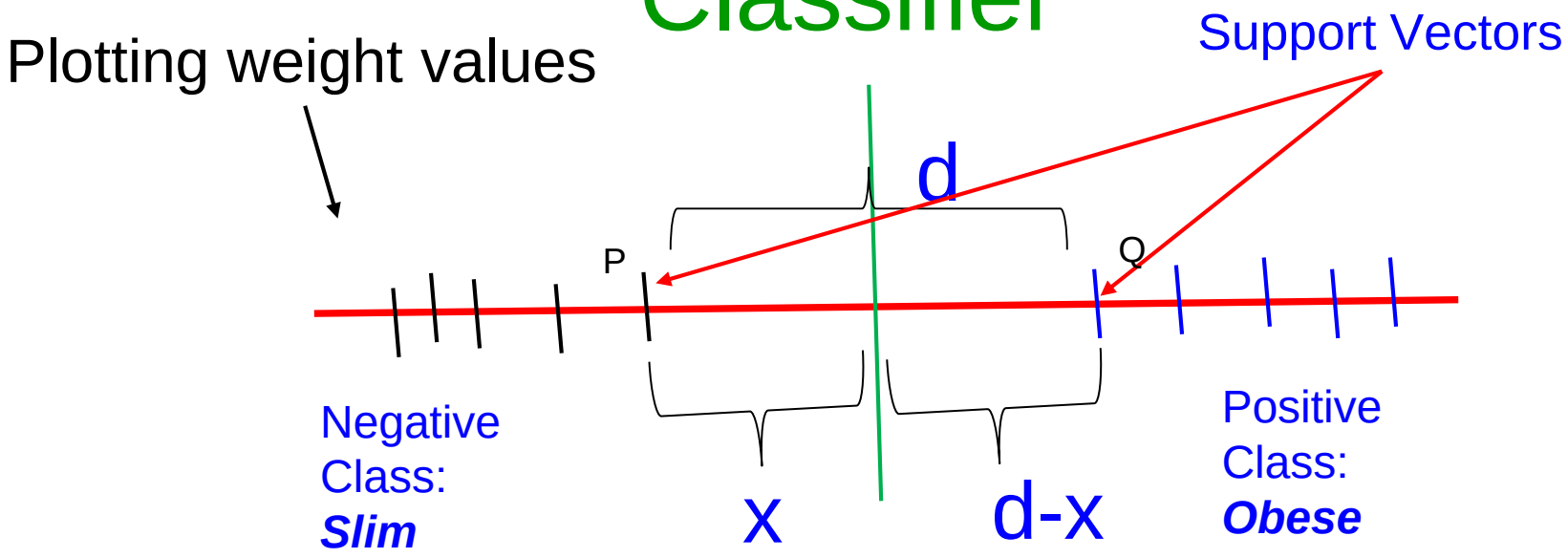
These inequalities are unsatisfiable.

XOR is NOT linearly separable

Support Vector Machine (SVM)

Slide courtesy: Dr. Sachin Pawar, CFILT and TCS

SVM is a Maximum Margin Classifier



P and Q are “support” points

$$margin,\ m = x^2 + (d-x)^2$$

$$\frac{dm}{dx} = 2x - 2(d-x)$$

$$equate\ to\ 0$$

$$x = \frac{d}{2}$$

What if the threshold is not in the middle

- Threshold in the middle maximizes margin
- This DECREASES the probability of misclassification
- Intuitively, if the threshold abuts too closely one of the supports- say negative class support P- then in future any negative point away from P will be wrongly labelled as positive

- Maximum margin minimizes the probability of misclassification

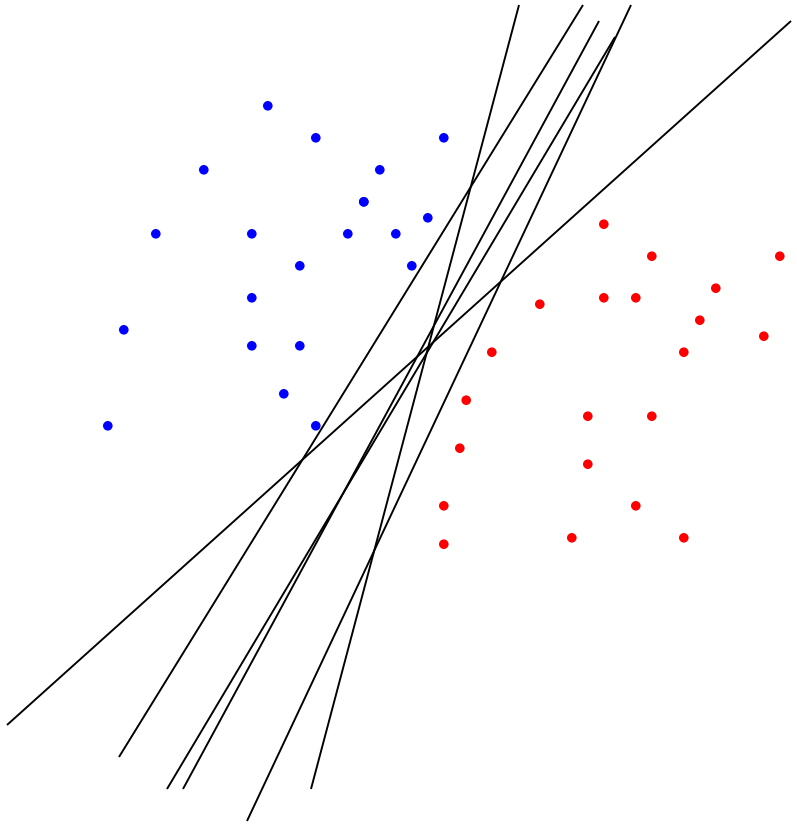
- Refer to Bias Variance trade off

Introduction

- Support Vector Machine (SVM): Learns a linear separator for separating instances belonging to two different classes
  - In case of 1 dimensional instances, the separator is a point
  - In case of 2 dimensional instances, the separator is a line
  - In case of 3 dimensional instances, the separator is a plane
  - In case of instances in more than 3 dimensional space, the separator is a hyperplane

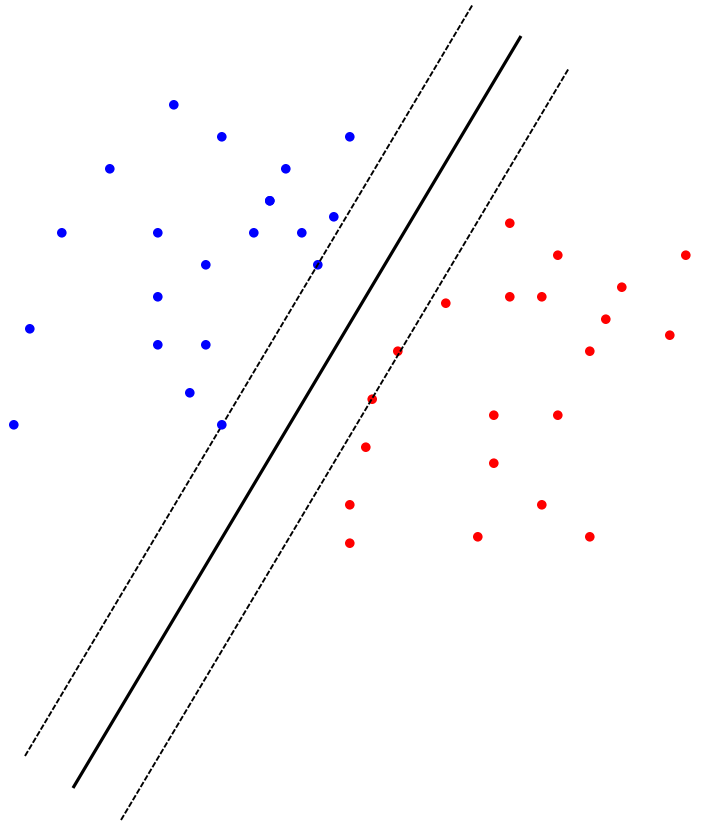
- Given a set of linearly separable instances, there exist infinite number of linear separators which can separate the instances into 2 classes
  - SVM chooses that linear separator which has the maximum “margin”
  - Intuition: A linear separator with the maximum margin will generalize better for new unseen instances in test data
  - So called, BIAS VARIANCE TRADE OFF

SVM: Linear Separator



- An example of two dimensional instances
- Two classes:
  - Positive and Negative
- Linearly separable data
- Infinite number of linear separators are possible

SVM: Linear Separator



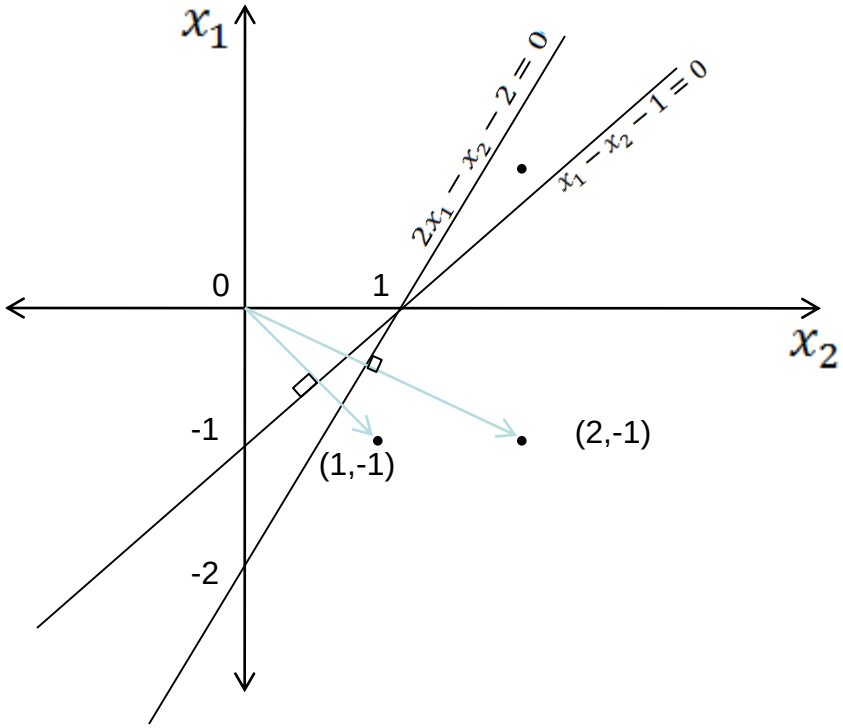
- **Goal**: To find the linear separator which provides the maximum margin of separation between two classes
- Support vectors: Instances which lie on the margin boundaries

Representation of a linear separator

- A linear separator in n-dimensional space is represented using an equation of the form:
- Can be expressed in a vector form as:

$$w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_0 = 0$$

$$\mathbf{w}^T \cdot \mathbf{x} + w_0 = 0$$

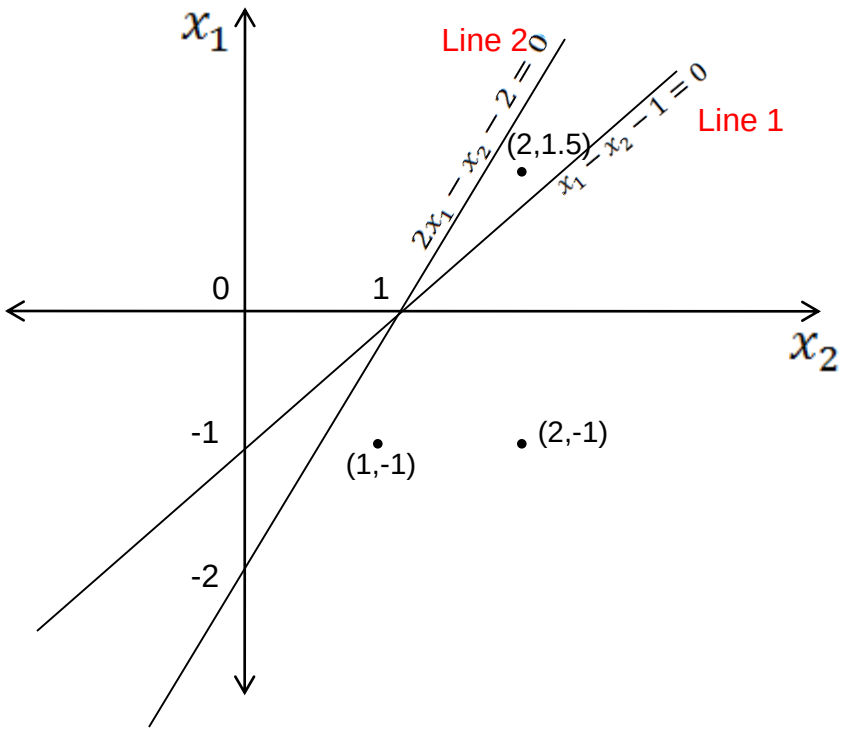


- The vector w is perpendicular to the lines of the form $w^T.x+w_0=0$

- The learning task in SVM is to determine the optimal values of the parameters representing the linear separator with the maximum margin, i.e., w and w_0

Representation of a linear separator

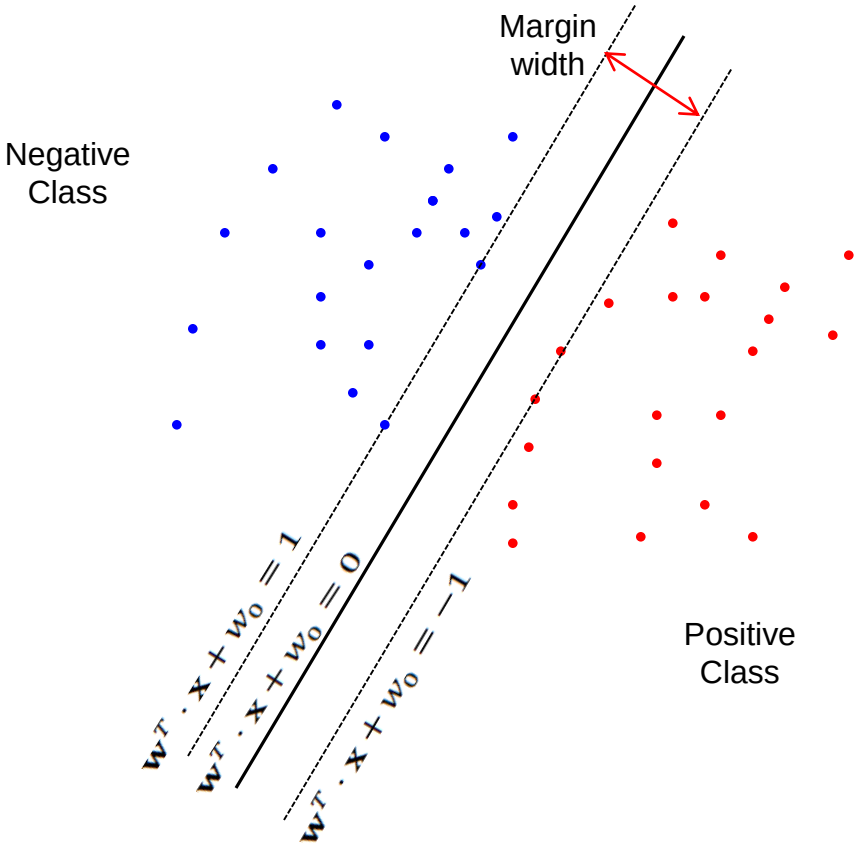
- A linear separator in n-dimensional space is represented using an equation of the form: $w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_0 = 0$
- Can be expressed in a vector form as: $\mathbf{w}^T \cdot \mathbf{x} + w_0 = 0$



For a particular linear separator, any point x^i lying on the "positive" side will have positive value of $w^T.x^i+w_0$

- Also, any point x^i lying on the "negative" side will have negative value of $w^T.x^i+x_0$
- E.g., the point (2,1.5) is on positive side of Line 2 (gives value 0.5) and on negative side of Line 1 (-0.5)

SVM: Training

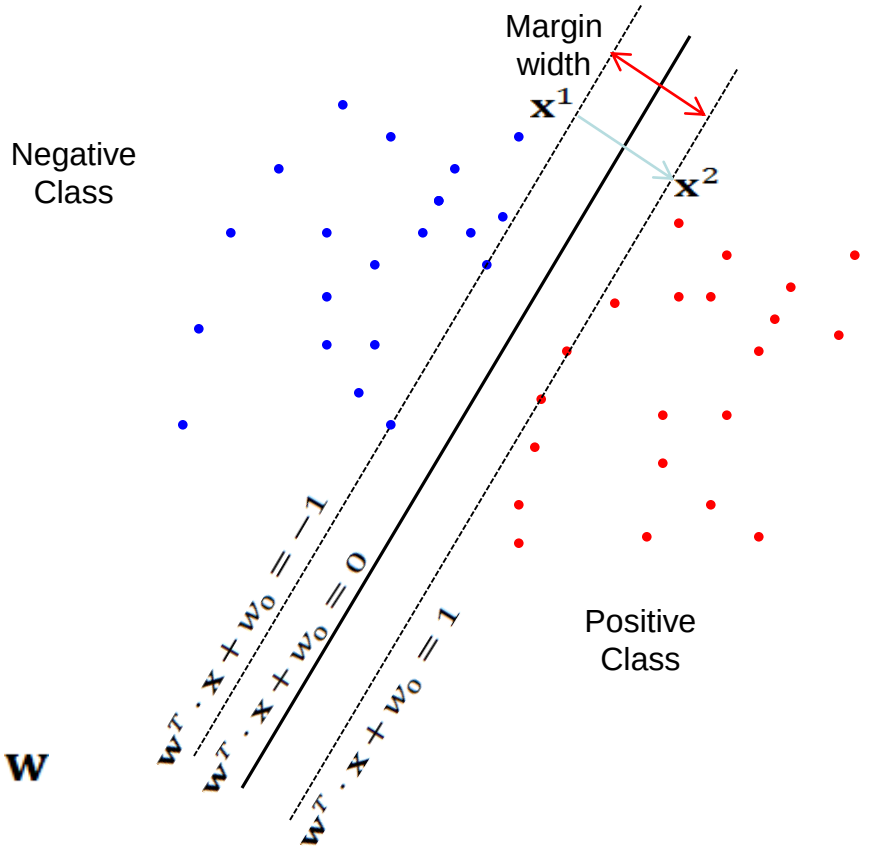


- **Training instances**:

$$\{\langle \mathbf{x}^1, y^1 \rangle, \langle \mathbf{x}^2, y^2 \rangle, \cdots \langle \mathbf{x}^N, y^N \rangle\}$$

- x^i is a point in n-dimensional space
- $y^i \in \{+1, -1\}$ is its corresponding true class label
- **Goal**: To find optimal linear separator which maximizes the margin

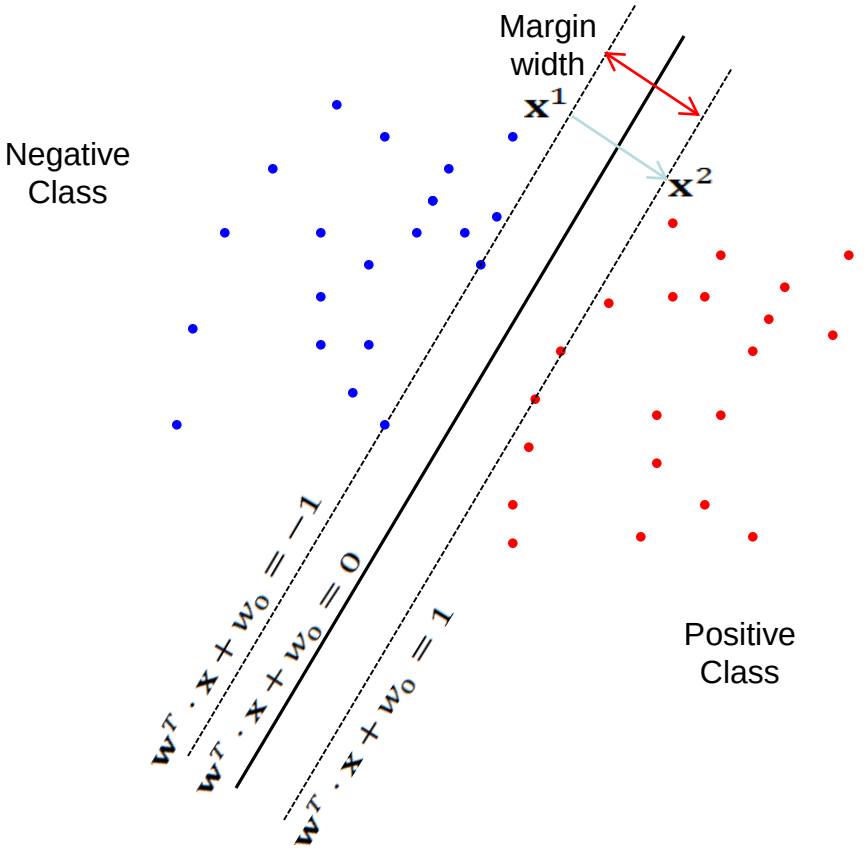
Computing Margin Width



- Consider two points x^1 and x^2 such that they lie on the opposite margins and the vector x^2-x^1 is perpendicular to the linear separator.
- The vector *w* is also perpendicular to the linear separator.
- Therefore, (x2-x1)=λ.w
- By definition,

$$\mathbf{w}^T \cdot \mathbf{x}^2 + w_0 = 1$$
$$\mathbf{w}^T \cdot \mathbf{x}^1 + w_0 = -1$$
$$\mathbf{w}^T \cdot (\mathbf{x}^2 - \mathbf{x}^1) = 2$$

Computing Margin Width



- Substituting

$$(\mathbf{x}^2 - \mathbf{x}^1) = \lambda \cdot \mathbf{w}$$

$$\mathbf{w}^T \cdot (\mathbf{x}^2 - \mathbf{x}^1) = 2$$

- Margin width:

$$\lambda \mathbf{w}^T \cdot \mathbf{w} = 2 \Rightarrow \lambda = \frac{2}{\mathbf{w}^T \cdot \mathbf{w}}$$

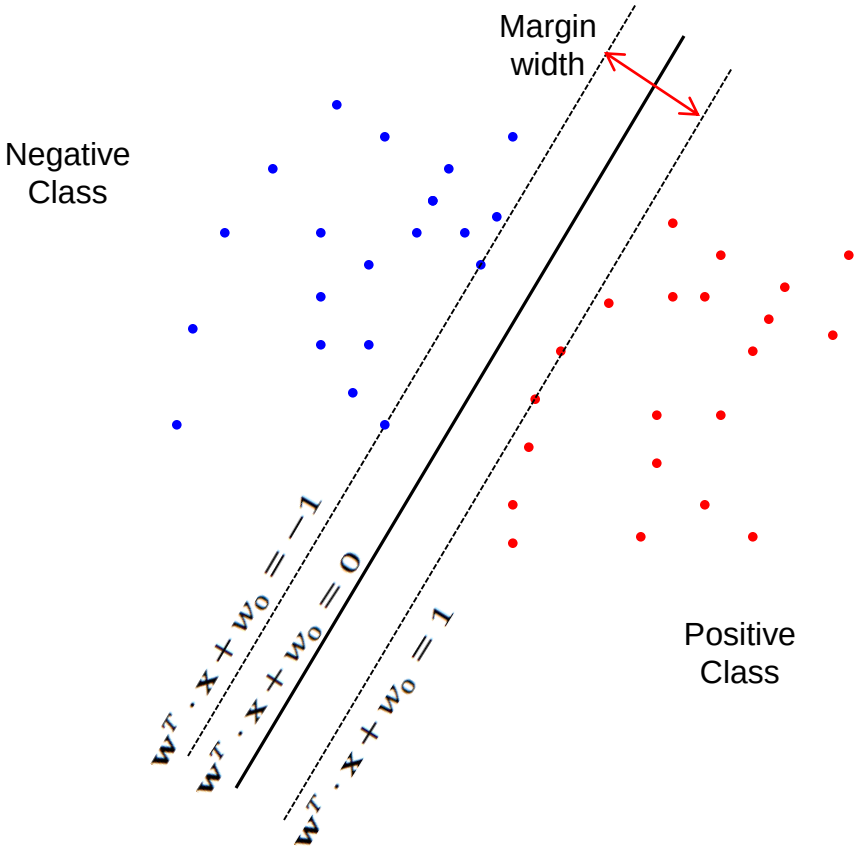
$$\|\mathbf{x}^2 - \mathbf{x}^1\| = \sqrt{(\mathbf{x}^2 - \mathbf{x}^1)^T \cdot (\mathbf{x}^2 - \mathbf{x}^1)}$$

$$\|\mathbf{x}^2 - \mathbf{x}^1\|^2 = \lambda^2 (\mathbf{w}^T \cdot \mathbf{w})$$

$$\|\mathbf{x}^2 - \mathbf{x}^1\|^2 = \frac{4}{(\mathbf{w}^T \cdot \mathbf{w})^2} (\mathbf{w}^T \cdot \mathbf{w})$$

$$\|\mathbf{x}^2 - \mathbf{x}^1\|^2 = \frac{4}{\mathbf{w}^T \cdot \mathbf{w}} \propto \frac{1}{\mathbf{w}^T \cdot \mathbf{w}}$$

SVM: Optimization Problem



- Objective:
  - Maximize the margin

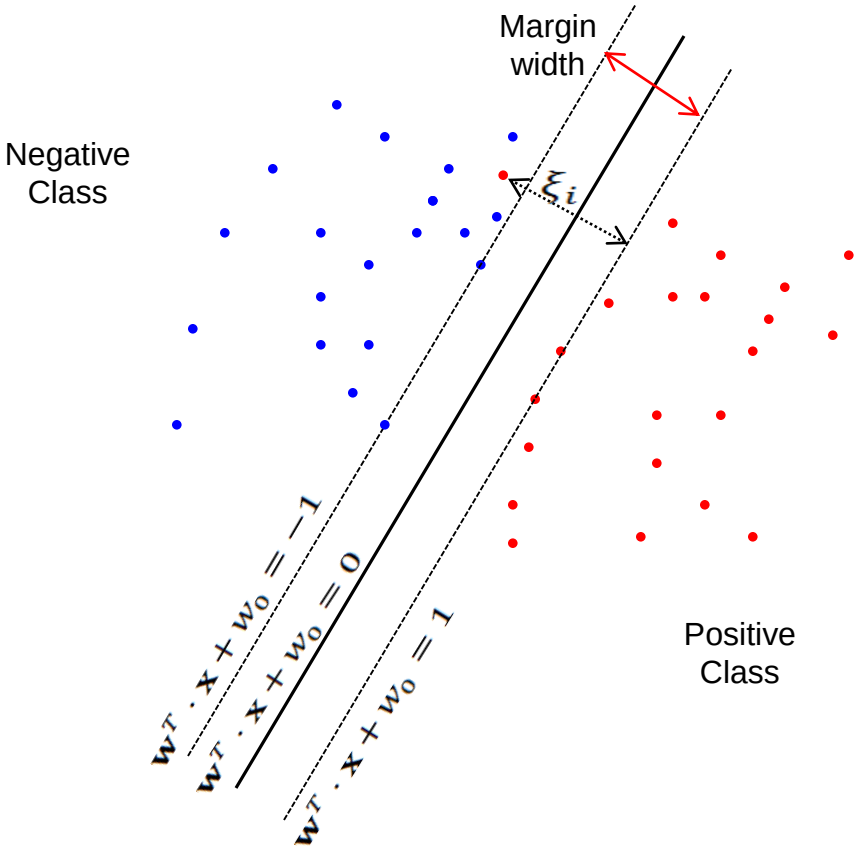
$$\min_{\mathbf{w}, w_0} \left(\frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} \right)$$

- Subject to the following constraints:
  - Every training instance should lie on the appropriate (positive / negative) side of the linear separator

$$y^i \left(\mathbf{w}^T \cdot \mathbf{x}^i + w_0 \right) \geq 1, \forall_{1 \leq i \leq N}$$

$$-y^i \left(\mathbf{w}^T \cdot \mathbf{x}^i + w_0 \right) + 1 \leq 0, \forall_{1 \leq i \leq N}$$

SVM: Soft-margin Formulation



- Objective:
  - Maximize the margin and minimize the training error

$$\min_{\mathbf{w}, w_0, \xi} \left(\frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} \right) + C \cdot \sum_{i=1}^{N} \xi_i$$

- Subject to the following constraints:
  - Introducing slack variables so that the constraint is satisfied for training instances lying on incorrect side

$$-y^i \left(\mathbf{w}^T \cdot \mathbf{x}^i + w_0 \right) + 1 - \xi_i \leq 0, \forall_{1 \leq i \leq N}$$
$$-\xi_i \leq 0, \forall_{1 \leq i \leq N}$$

Optimization using Lagrange Multipliers

- One Lagrange multiplier is associated with each distinct constraint

$$L(\alpha_1, \cdots, \alpha_N, \mu_1, \cdots, \mu_N)$$

$$= \min_{\mathbf{w}, w_0, \xi} \left(\frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} \right) + C \cdot \sum_{i=1}^{N} \xi_i$$

$$+ \sum_{i=1}^{N} \alpha_i \cdot \left(-y^i \left(\mathbf{w}^T \cdot \mathbf{x}^i + w_0 \right) + 1 - \xi_i \right)$$

$$+ \sum_{i=1}^{N} \mu_i \cdot (-\xi_i)$$

$$s.t. \quad \alpha_i \geq 0, \mu_i \geq 0, \forall_{1 \leq i \leq N}$$

Optimization using Lagrange Multipliers

- Differentiating w.r.t. $\mathbf{w}$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y^i \mathbf{x}^i = 0 \quad \Rightarrow \boxed{\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i y^i \mathbf{x}^i}$$

- Differentiating w.r.t. w_0

$$\frac{\partial L}{\partial w_0} = \sum_{i=1}^{N} -\alpha_i y^i = 0 \quad \Rightarrow \boxed{\sum_{i=1}^{N} \alpha_i y^i = 0}$$

- Differentiating w.r.t. ξ_i

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \quad \Rightarrow \boxed{\mu_i + \alpha_i = C}$$

- Substituting optimal values:

$$L(\alpha_1, \cdots, \alpha_N, \mu_1, \cdots, \mu_N)$$

$$= \left(\frac{1}{2}\mathbf{w}^{*T} \cdot \mathbf{w}^*\right) + C \cdot \sum_{i=1}^{N} \xi_i^*$$

$$+ \sum_{i=1}^{N} \alpha_i \cdot \left(-y^i\left(\mathbf{w}^{*T} \cdot \mathbf{x}^i + w_0^*\right) + 1 - \xi_i^*\right) + \sum_{i=1}^{N} \mu_i \cdot (-\xi_i^*)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N} \alpha_i y^i \mathbf{x}^i\right)^T \cdot \left(\sum_{j=1}^{N} \alpha_j y^j \mathbf{x}^j\right) + C \cdot \sum_{i=1}^{N} \xi_i^*$$

$$+ \sum_{i=1}^{N} -\alpha_i y^i \left(\sum_{j=1}^{N} \alpha_j y^j \mathbf{x}^j\right)^T \mathbf{x}^i - w_0^* \sum_{i=1}^{N} \alpha_i y^i$$

$$+ \sum_{i=1}^{N} \alpha_i (1 - \xi_i) + \sum_{i=1}^{N} \mu_i \cdot (-\xi_i^*)$$

Terms involving ξ_i cancel each other out because $\mu_i + \alpha_i = C$

Sum in the red circle is zero

- Finally, we get:

$$L(\alpha_1, \cdots, \alpha_N) = \frac{-1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j \left(\mathbf{x}^{i^T} \mathbf{x}^j \right) + \sum_{i=1}^{N} \alpha_i$$

- Dual optimization problem:
  - Objective function:

$$\max_{\alpha_1, \cdots, \alpha_N} \frac{-1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j \left(\mathbf{x}^{i^T} \mathbf{x}^j \right) + \sum_{i=1}^{N} \alpha_i$$

  - Subject to the following constraints:

$$\alpha_i \geq 0, \mu_i \geq 0, \mu_i + \alpha_i = C, \forall_i \ \ and \ \sum_{i=1}^{N} \alpha_i y^i = 0$$

$$\Rightarrow \ \ \alpha_i \geq 0, \alpha_i \leq C, \forall_i \ \ and \ \sum_{i=1}^{N} \alpha_i y^i = 0$$

Any Quadratic Programming Solver can be used for solving this

Using SVM for Predictions

- How to predict the class label for a new instance $\mathbf{x}$ given a trained SVM
- Primal Form:
  - Compute $\mathbf{w}^T \cdot \mathbf{x} + w_0$; Positive value indicates the positive class and vice versa
  - E.g., $\mathbf{w} = [2, -1]\ and\ w_0 = -2$ be the learned parameters
  - For the new instance $\mathbf{x} = [2, 4]$ $\mathbf{w}^T \cdot \mathbf{x} + w_0 = -2$ and hence Negative class is predicted
- Dual Form:
  - Compute $$\mathbf{w}^T \cdot \mathbf{x} + w_0 = \left(\sum_{i=1}^{N} \alpha_i y^i \mathbf{x}^i \right)^T \mathbf{x} = \sum_{i=1}^{N} \alpha_i y^i \mathbf{x}^{i^T} \mathbf{x}$$
  - Positive value indicates the positive class and vice versa
  - Practically, most of the α_i values are zeros; non-zero only for support vectors