CMPUT 651 (Fall 2019)

NLP Tasks and Linear Classification

Lili Mou
lmou@ualberta.ca
lili-mou.github.io

A Landscape of AI

Artificial Intelligence
- AAAI, IJCAI

...

Logical reasoning

Linguistics

Machine Learning
- ICML, NeurIPS, ICLR

Machine Learning Theory (e.g.,COLT)

(Applications)

Natural Language Processing
- ACL, EMNLP, NAACL, COLING

Data Mining
- KDD, ICDM

Computer Vision
- CVPR, ICCV, ECCV

Speech Processing
- INTERSPEECH, ICASSP

What do NLPers actually do?

- The best resource to be acquainted with NLP research is conference proceedings

SUBMISSIONS

ACL 2019 has the goal of a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Applications
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Document Analysis
- Generation
- Information Extraction and Text Mining
- Linguistic Theories, Cognitive Modeling and Psycholinguistics
- Machine Learning
- Machine Translation
- Multidisciplinary
- Multilinguality
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Sentence-level semantics
- Sentiment Analysis and Argument Mining
- Social Media
- Summarization
- Tagging, Chunking, Syntax and Parsing
- Textual Inference and Other Areas of Semantics
- Vision, Robotics, Multimodal, Grounding and Speech
- Word-level Semantics

Call for papers (ACL 2019)

But they're usually out-of-date

What do NLPers actually do?

- The best resource to be acquainted with NLP research is conference proceedings

Table of Contents of ACL'19

But there're simply too many

A Few Applications of NLP

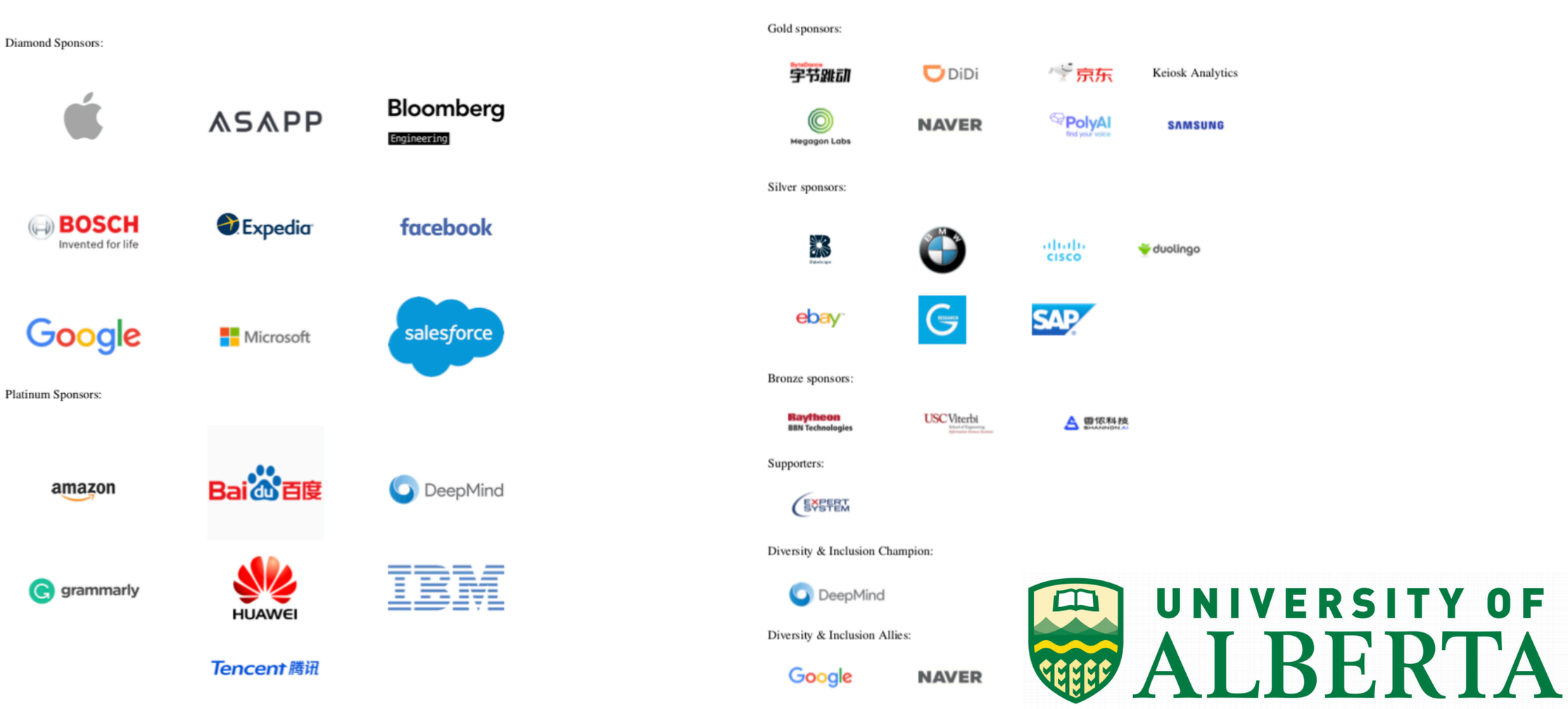
- Linguistic aspects
  - Part-of-speech tagging, parsing
- Natural language understanding
  - Sentiment classification, relation extraction
  - Question answering, machine comprehension
- Natural language generation
  - Machine translation, text summarization, dialogue systems

Researchers like to invent new tasks to publish more papers

Question: Suppose we have N models and M tasks, how many papers can we publish? $2^{N+M} - N - M + 1$

Why is NLP important?

- Related to the core of AI
  - One of the most difficult area that AI hasn't conquered
  - Many unsolved scientific problems
- Bringing revenues for companies



Why is NLP difficult?

- Ambiguity
  - Word sense disambiguation

    *I went to the **bank** to deposit some money.*

    *I got wet on the river **bank**.*

Why is NLP difficult?

- Ambiguity
  - Word sense disambiguation

    *I went to the **bank** to deposit some money.*

    *I got wet on the river **bank**.*

  - Coreference resolution

    *The book doesn't fit into the bag, because **it** is too small*

    *The book doesn't fit into the bag, because **it** is too big*

Why is NLP difficult?

- Subtlety
  - Word sense disambiguation
    - How many senses?

*The **bank** was on fire*

*The **bank** blocked my credit card*

WordNet
http://wordnetweb.princeton.edu/perl/webwn

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: bank Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

Verb

- S: (v) **bank** (tip laterally) *"the pilot had to bank the aircraft"*
- S: (v) **bank** (enclose with a bank) *"bank roads"*
- S: (v) **bank** (do business with a bank or keep an account at a bank) *"Where do you bank in this town?"*
- S: (v) **bank** (act as the banker in a game or in gambling)
- S: (v) **bank** (be in the banking business)
- S: (v) deposit, **bank** (put into a bank account) *"She deposits her paycheck every month"*
- S: (v) **bank** (cover with ashes so to control the rate of burning) *"bank a fire"*
- S: (v) count, bet, depend, swear, rely, **bank**, look, calculate, reckon (have faith or confidence in) *"you can count on me to help you any time"; "Look to your friends for support"; "You can bet on that!"; "Depend on your family in times of crisis"*

Rules-Based Methods

- Rules for sentiment analysis
  - If #(positive words) > #(negative words), then predict positive
  - How about this sentence:

    Although CMPUT 651 might be interesting, it's less likely that I'll like it.

- There is an exception to every (NLP) rule!

CMPUT 651 (Fall 2019)

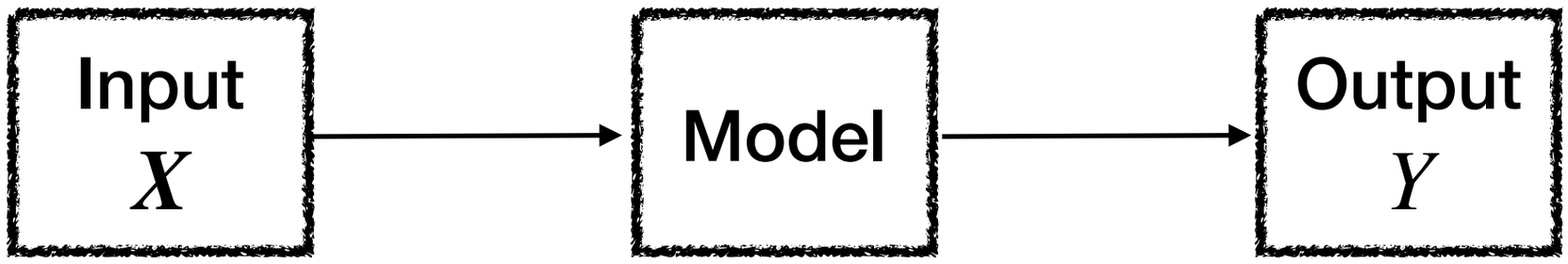
Machine Learning

- Supervised learning
  - Regression, classification
- Unsupervised learning
  - Clustering, PCA, etc.
  - Representation learning

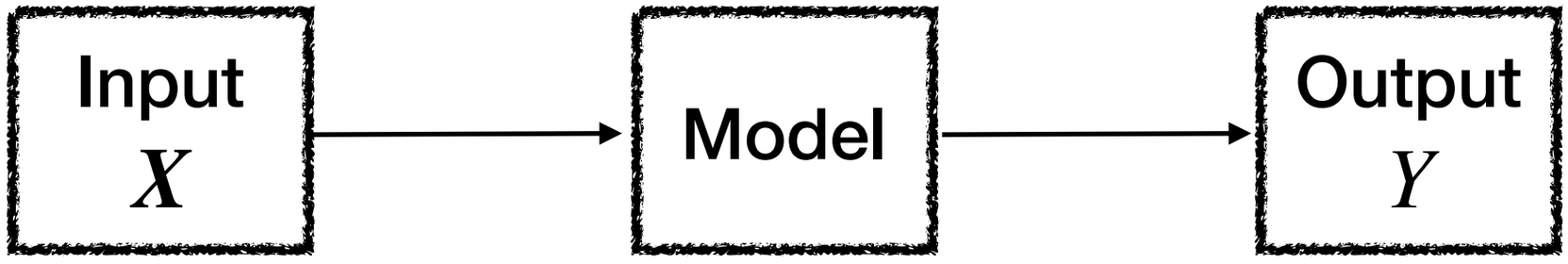
UNIVERSITY OF ALBERTA

CMPUT 651 (Fall 2019)

Supervised Learning



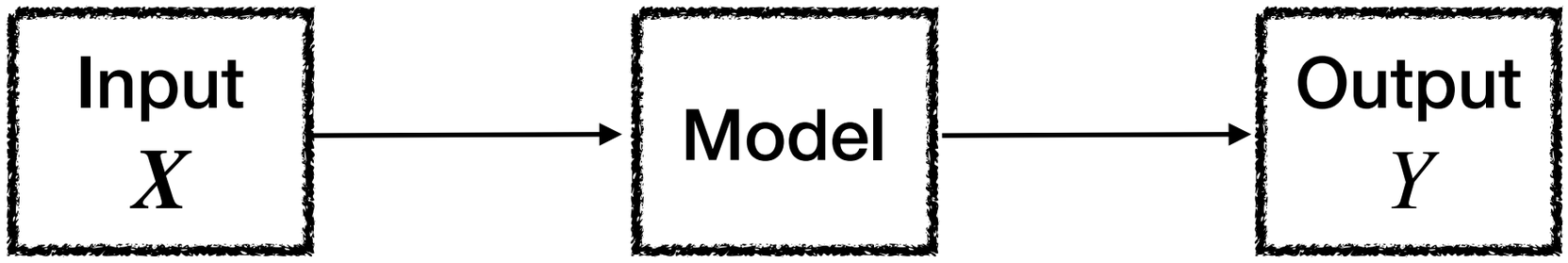
Supervised Learning



- **Training:** <X, Y> known for many samples
- **Validation:** <X, Y> also known
  - Pretend Y unknown and see model performance
  - Hyperparameter tuning, model selection
- **Test:** Given new X*, predict Y*
  - Industrial application: Deploy your system
  - Research: Y* in fact also known, report test performance

Supervised Learning

- E.g., sentiment analysis

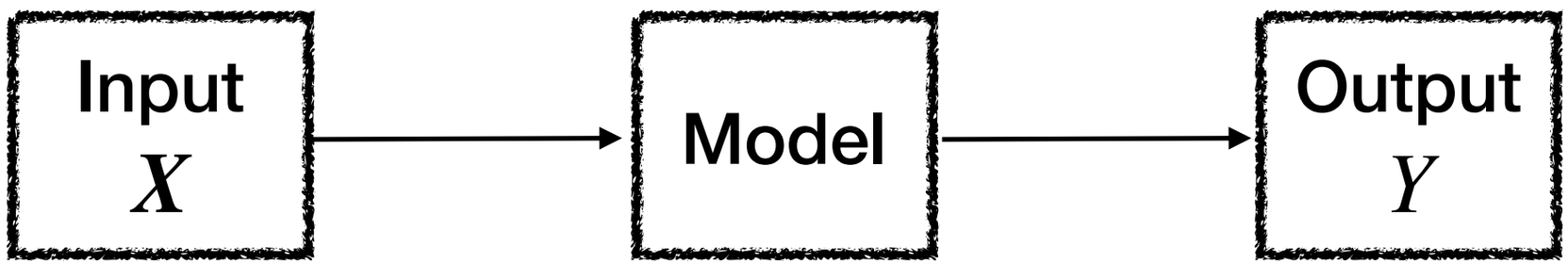


  - BoW (Indicator, tf·idf)
  - N-gram features
  - Sentiment lexicon

- NLP challenges:
  - Varying length
  - Unseen words

Supervised Learning

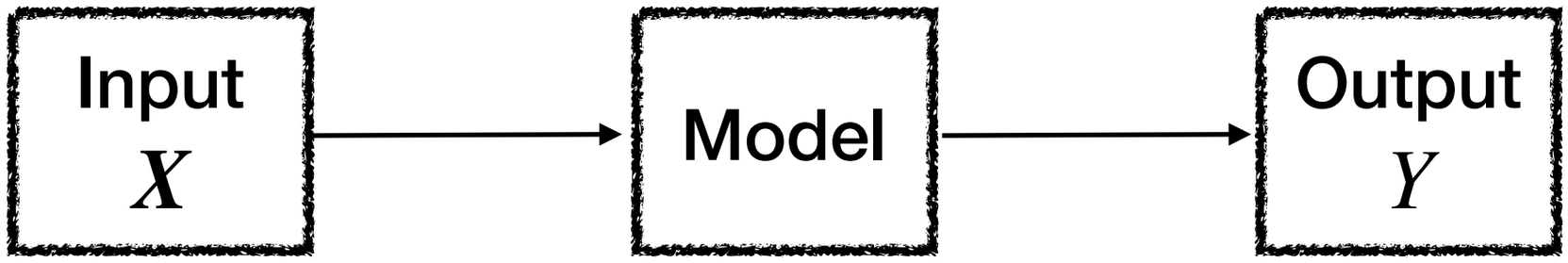
- E.g., sentiment analysis



- Regression?
- Classification?
  - Binary
  - Multi-class

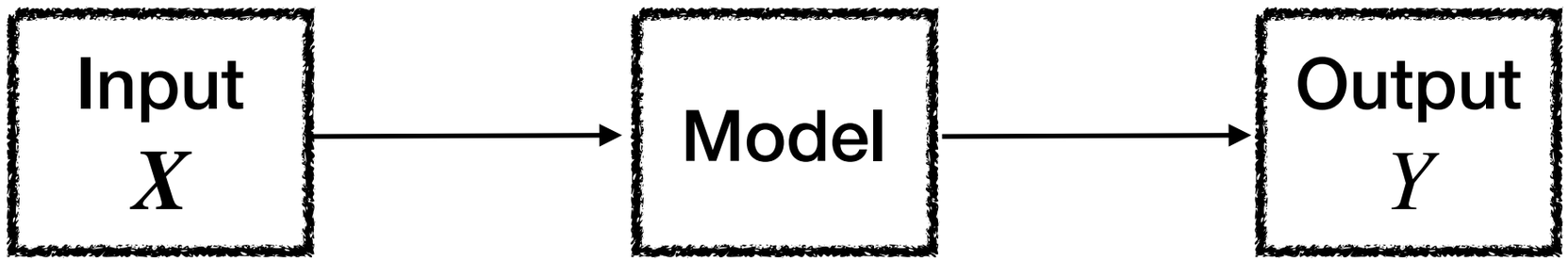
Supervised Learning

- E.g., sentiment analysis



  - Non-probabilistic
    - SVM, Fisher's discriminant
  - Generative
    - Naive Bayes
  - Discriminative
    - Logistic regression, softmax

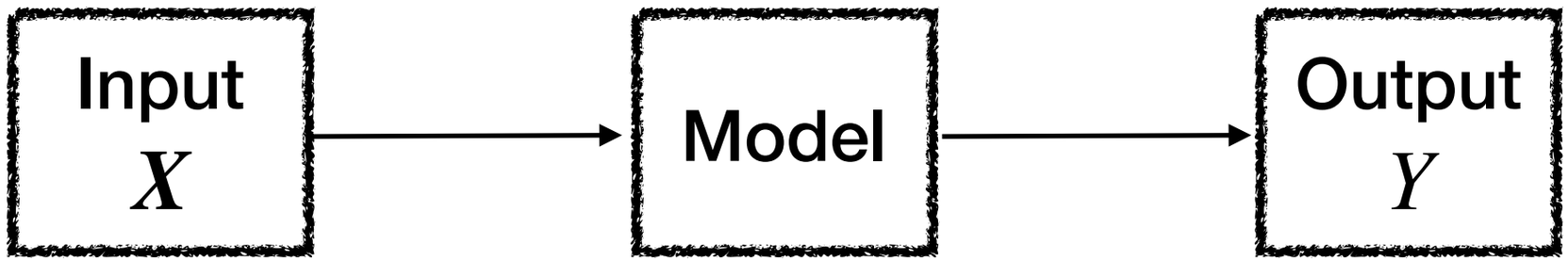
Probabilistic Classification



- Probabilistic modeling of $\boldsymbol{X}, Y$
  - We can say something about $p(\boldsymbol{X}), p(Y), p(\boldsymbol{X}, Y), p(\boldsymbol{X} \mid Y)$, or $p(Y \mid \boldsymbol{X})$
  - Max a posterior inference <=> Minimal empirical loss

$$y = \operatorname{argmax} p(Y \mid \boldsymbol{x})$$

Probabilistic Classification



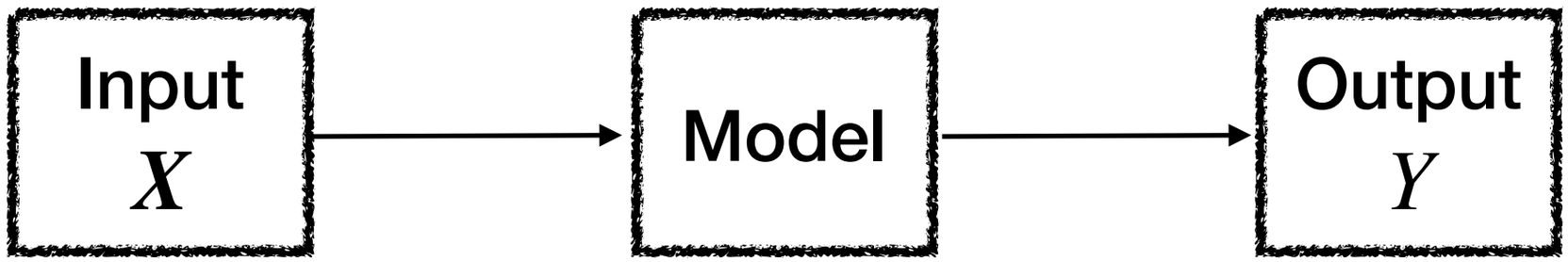
- Probabilistic modeling of $\boldsymbol{X}, Y$
  - We can say something about $p(\boldsymbol{X}), p(Y), p(\boldsymbol{X}, Y), p(\boldsymbol{X} \mid Y)$, or $p(Y \mid \boldsymbol{X})$
  - Max a posterior inference <=> Minimal empirical loss

$$y = \operatorname{argmax} p(Y \mid \boldsymbol{x})$$

  - **Generative model:** Model $p(\boldsymbol{X}, Y)$

$$p(Y \mid x) \propto_Y p(x, Y) = p(Y)p(x \mid Y)$$

Probabilistic Classification

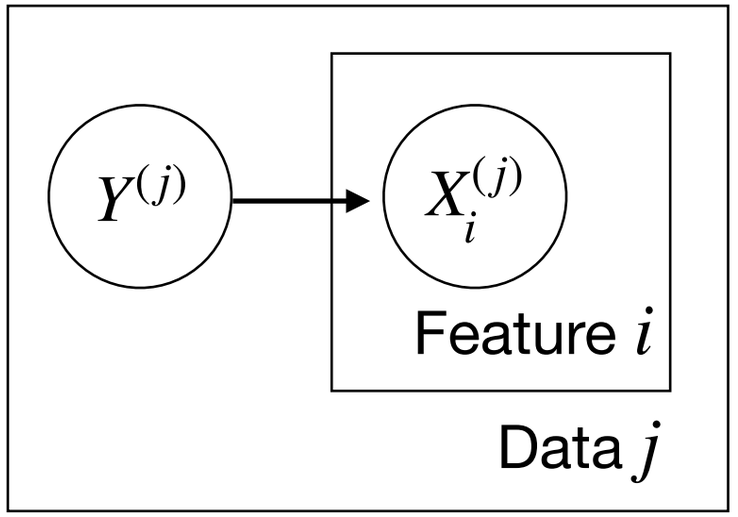


- Probabilistic modeling of $\boldsymbol{X}, Y$
  - We can say something about $p(\boldsymbol{X}), p(Y), p(\boldsymbol{X}, Y), p(\boldsymbol{X} \mid Y)$, or $p(Y \mid \boldsymbol{X})$
  - Max a posterior inference <=> Minimal empirical loss

$$y = \operatorname{argmax} p(Y \mid \boldsymbol{x})$$

  - **Discriminative model:** Directly modeling $p(Y \mid \boldsymbol{X})$

Naïve Bayes



- Inference objective

  $y = \operatorname{argmax} p(Y \mid \boldsymbol{x})$

- Training objective: Maximize

$$p(\mathcal{D}) = p(\{\boldsymbol{X}^{(j)}, Y^{(j)}\}_{j=1}^{m})$$

max. joint prob.

$$= \prod_{j=1}^{m} p(Y^{(j)}) p(\boldsymbol{X}^{(j)} \mid Y^{(j)})$$

data iid

$$= \prod_{j=1}^{m} \left[p(Y^{(j)}) \prod_{i=1}^{n} p(\boldsymbol{X}_i^{(j)} \mid Y^{(j)}) \right]$$

Naive Bayes assumpt.

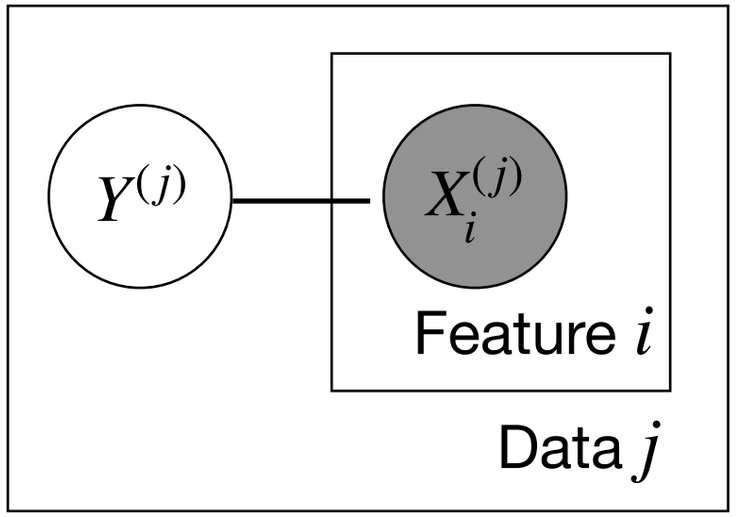
$$= \left[\prod_{j=1}^{m} p(Y^{(j)}) \right] \left[\prod_{y} \prod_{j:Y^{(j)}=y} p(X_i^{(j)} \mid y) \right]$$

massaging

- Each factor is directly parametrized
- Maximum likelihood estimation for multinomial distributions is simply counting!

Logistic Regression

[classification task]



- Inference objective

  $y = \operatorname{argmax} p(Y|\boldsymbol{x})$

- Training objective: Maximize

$$\text{maximize} \quad \prod_{j=1}^{n} p(Y^{(j)}|\boldsymbol{X}^{(j)})$$

$$\Leftrightarrow \quad \text{maximize} \quad \log \prod_{j=1}^{n} p(Y^{(j)}|\boldsymbol{X}^{(j)})$$

(if optimum is achieved)

$$= \sum_{j=1}^{n} \log p(Y^{(j)}|\boldsymbol{X}^{(j)})$$

$$\Leftrightarrow \quad \text{minimize} \quad \sum_{j=1}^{n} \left[-t^{(j)} \log y^{(i)} - (1 - t^{(j)}) \log(1 - y^{(j)})\right]$$

Abusing some notations

$t^{(j)}$: The ground truth label of $Y^{(j)}$, $t^{(j)} = 0$ or $t^{(j)} = 1$ for binary classification

$y^{(j)}$: The predicted probability that $Y^{(j)} = 1$, i.e., $y^{(j)} \stackrel{\text{def}}{=} \Pr\{Y^{(j)} = 1 | \boldsymbol{x}^{(j)}; \mathscr{M}\}$

Logistic Regression

[classification task]

- Inference objective

  $y = \operatorname{argmax} p(Y|\boldsymbol{x})$

- Training objective: Maximize

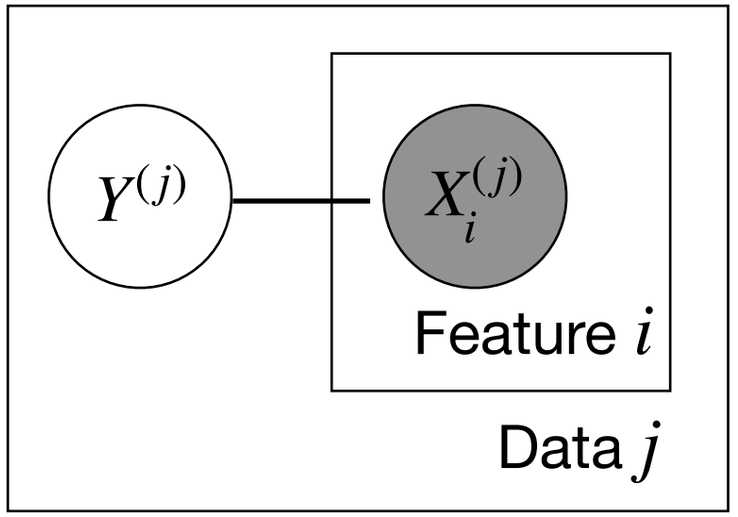
$$\text{maximize} \quad \prod_{j=1}^{n} p(Y^{(j)}|\boldsymbol{X}^{(j)})$$

$$\Leftrightarrow \quad \text{maximize} \quad \log \prod_{j=1}^{n} p(Y^{(j)}|\boldsymbol{X}^{(j)})$$

(if optimum is achieved)

$$= \sum_{j=1}^{n} \log p(Y^{(j)}|\boldsymbol{X}^{(j)})$$

$$\Leftrightarrow \quad \text{minimize} \quad \sum_{j=1}^{n} \left[-t^{(j)} \log y^{(i)} - (1-t^{(j)}) \log(1-y^{(j)})\right]$$



Statisticians seem to be pessimistic creatures who think in terms of losses.
Decision theorists in economics and business talk instead in terms of gains (utility).

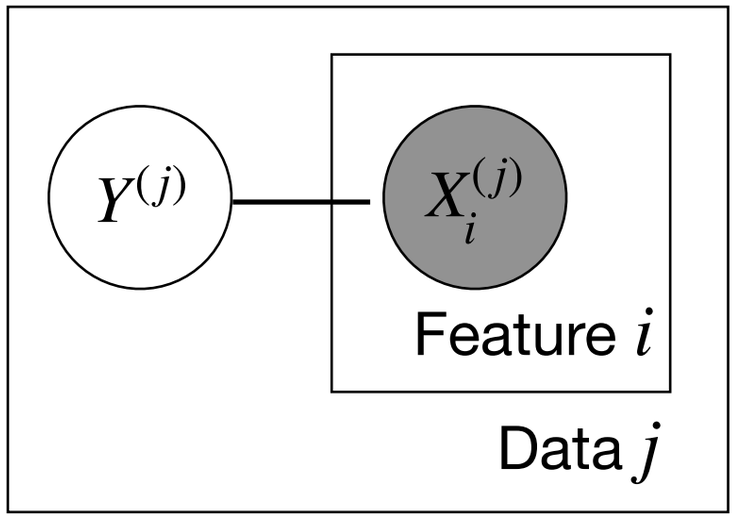
James O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*.

Abusing some notations

$t^{(j)}$: The ground truth label of $Y^{(j)}$, $t^{(j)} = 0$ or $t^{(j)} = 1$ for binary classification

$y^{(j)}$: The predicted probability that $Y^{(j)} = 1$, i.e., $y^{(j)} \stackrel{\text{def}}{=} \Pr\{Y^{(j)} = 1 | \boldsymbol{x}^{(j)}; \mathscr{M}\}$

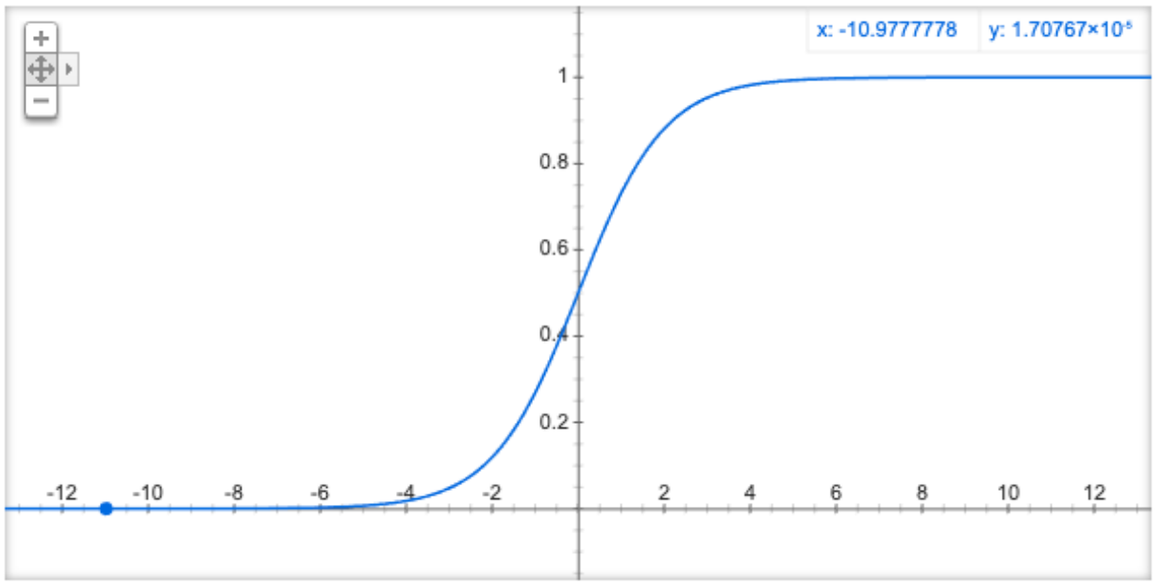
Logistic Regression



- How can we model $p(Y \mid \boldsymbol{X})$?
  - Simplest case: linear classification
  - Obtain a score by $s = \boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0$
  - Squash it as a probability

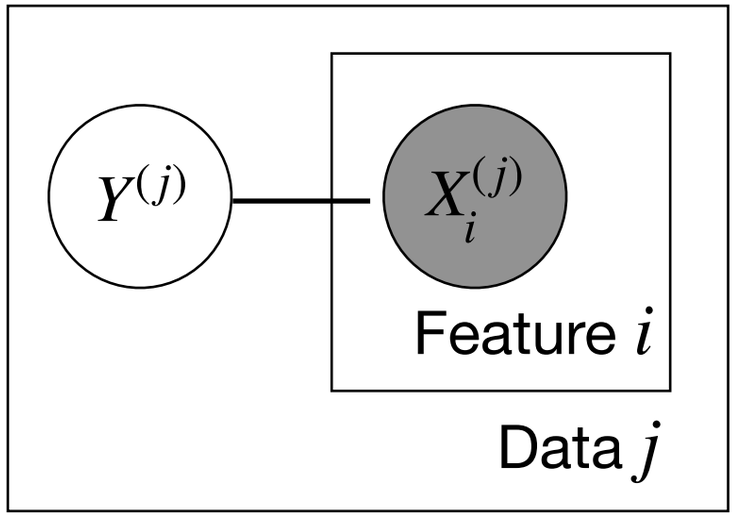
$$y = p(Y = 1 \mid \boldsymbol{x}) \overset{\text{def}}{=} \text{sigmoid}(s) \overset{\text{def}}{=} \frac{1}{1 + e^{-s}}$$

Graph for 1/(1+e^-x)



UNIVERSITY OF ALBERTA

Logistic Regression



- Why sigmoid?
  - Generalized linear model
    - $\boldsymbol{x}$ enters the model by linear transformation
    - y responses by exponential family
      - Why exponential family?
        - Most distributions we use are EXP
- Is it possible to use other functions to squash the distribution?
  - Yes. E.g., Probit regression (cdf of normal)

Logistic vs. Probit Regression

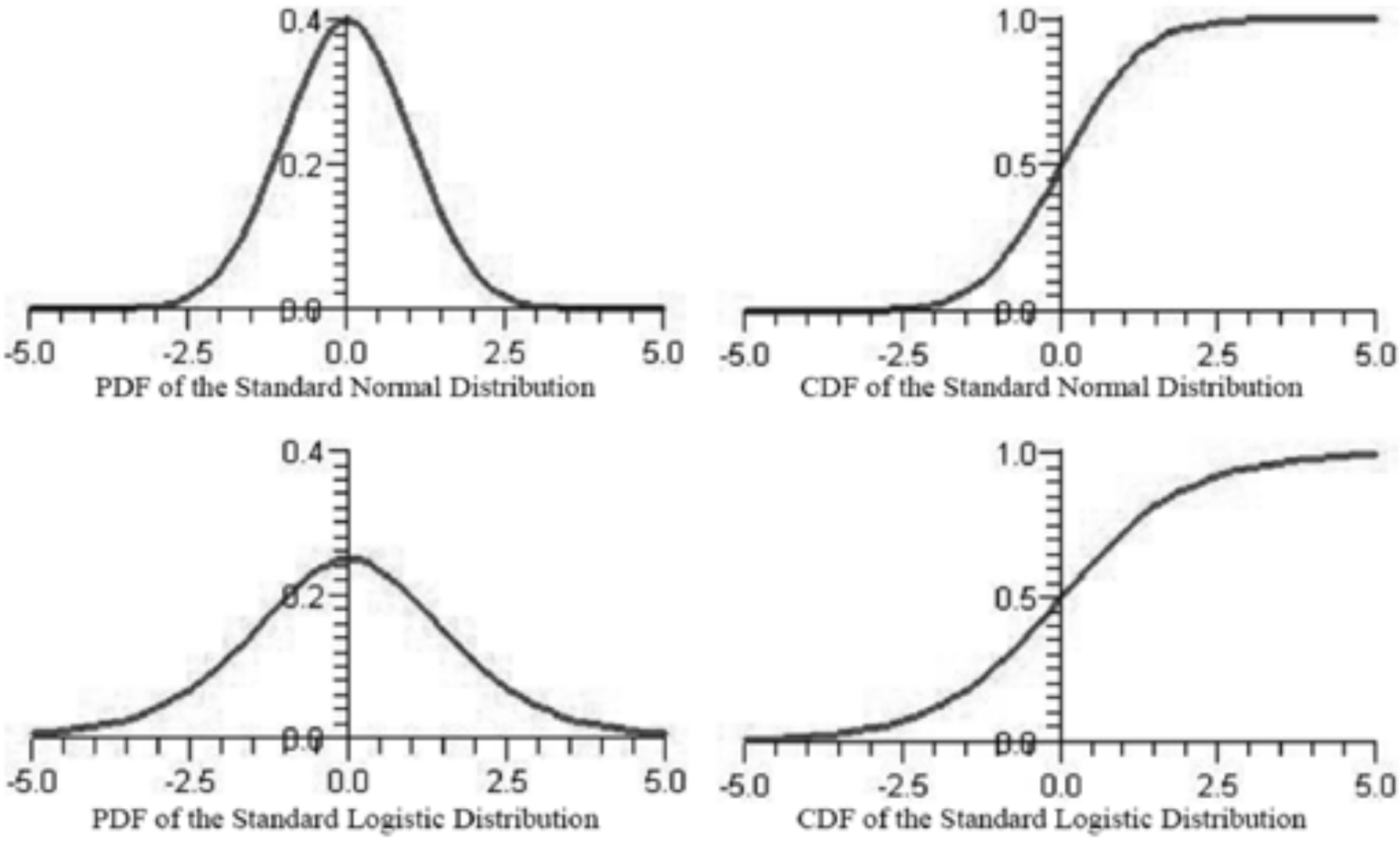


Figure 1. The Standard Normal and Standard Logistic Probability Distributions

Source : Park (2010)

Logistic Regression

- Training data $\mathscr{D} = \{(\boldsymbol{x}^{(j)}, y^{(j)})\}_{j=1}^{n}$

- Model $$y^{(j)} = \frac{1}{1 + e^{-(\theta_0 + \boldsymbol{\theta}^\top \boldsymbol{x}^{(j)})}}$$

  Model parameters: $\Theta = (\theta_0, \boldsymbol{\theta})$

- Loss $$J = \sum_{j=1}^{n} \left[-t^{(j)} \log y^{(i)} - (1 - t^{(j)}) \log(1 - y^{(j)}) \right]$$

- Optimization: gradient descent

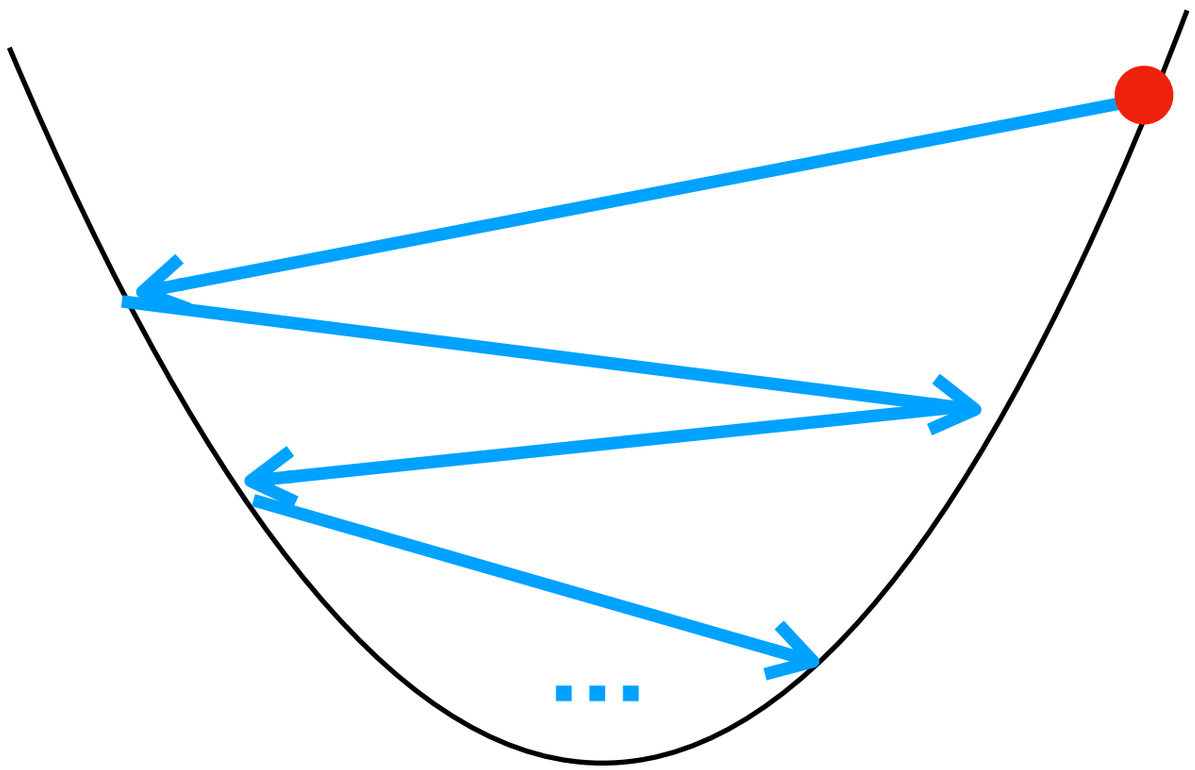
$$\theta_i = \theta_i - \alpha \frac{\partial J}{\partial \theta_i}$$

Gradient Descent

- Start from a (possibly) arbitrary position
- Take a mini-step against the gradient

$$\frac{\partial J}{\partial \theta_i} \overset{\text{def}}{=} \lim_{\delta \to 0} \frac{J(\theta_i + \delta) - J(\theta_i)}{\delta}$$

$$\theta_i = \theta_i - \alpha \frac{\partial J}{\partial \theta_i}$$



Gradient of LR

- For simplicity, we consider one sample

$$J = -t \log y - (1-t)\log(1-y)$$

- Steepest gradient descent

$$\frac{\partial J}{\partial \theta_i} = (y-t)x_i$$

Mini-Batch

- Each data point i, we have defined loss $J^{(i)}$
- We would like to optimize total loss $J = \sum_{i=1}^{n} J^{(i)}$
- We approximate J by a set of samples

$$J_{batch} = \sum_{i \in batch} J^{(i)}$$

Softmax

Multi-class classification

- Suppose target label $t \in \{0,1,\cdots,n\}$
- For each target i, we compute the **logit** $o_i = \boldsymbol{w}_i^\top \boldsymbol{x}$
- Normalize logits are probabilities

$$p(Y=i) \triangleq y_i \propto_i \exp\{\boldsymbol{w}_i^\top \boldsymbol{x}\} \qquad \text{positive}$$

$$y_i = \frac{\exp\{w_i^\top \boldsymbol{x}\}}{\sum_i^{'} \exp\{w_{i'}^\top \boldsymbol{x}\}} \qquad \text{normalizing}$$

- Cross-entropy loss

$$J = -t_i \log y_i \qquad \boldsymbol{t}\text{: one-hot}$$

$$= -\log \boldsymbol{y}[t] \qquad t\text{: ID representation}$$

Applying the single classifier to nearly all NLP tasks

- Part-of-speech (POS) tagging

This	lecture	is	really	boring
`article`	`noun`	`verb`	`adverb`	`adjective`

[Predict the tag among all candidate tags for a word]

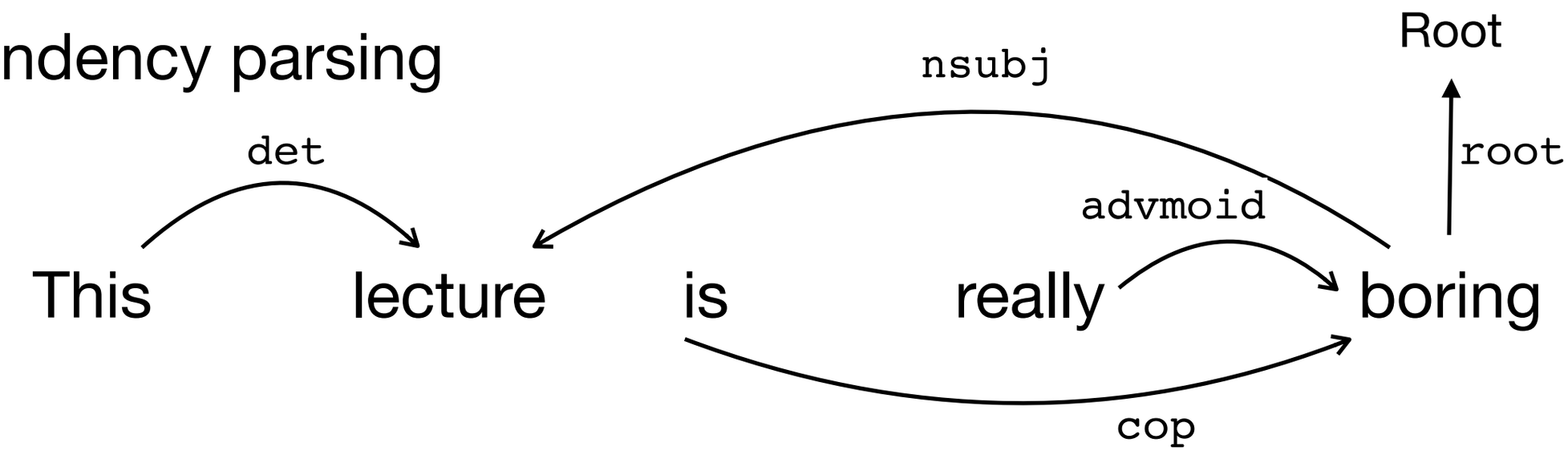
- Chunking

This	lecture	is	really	boring
–	|	|	–	

[Predict - or | for each two consecutive words]

Applying the single classifier to nearly all NLP tasks

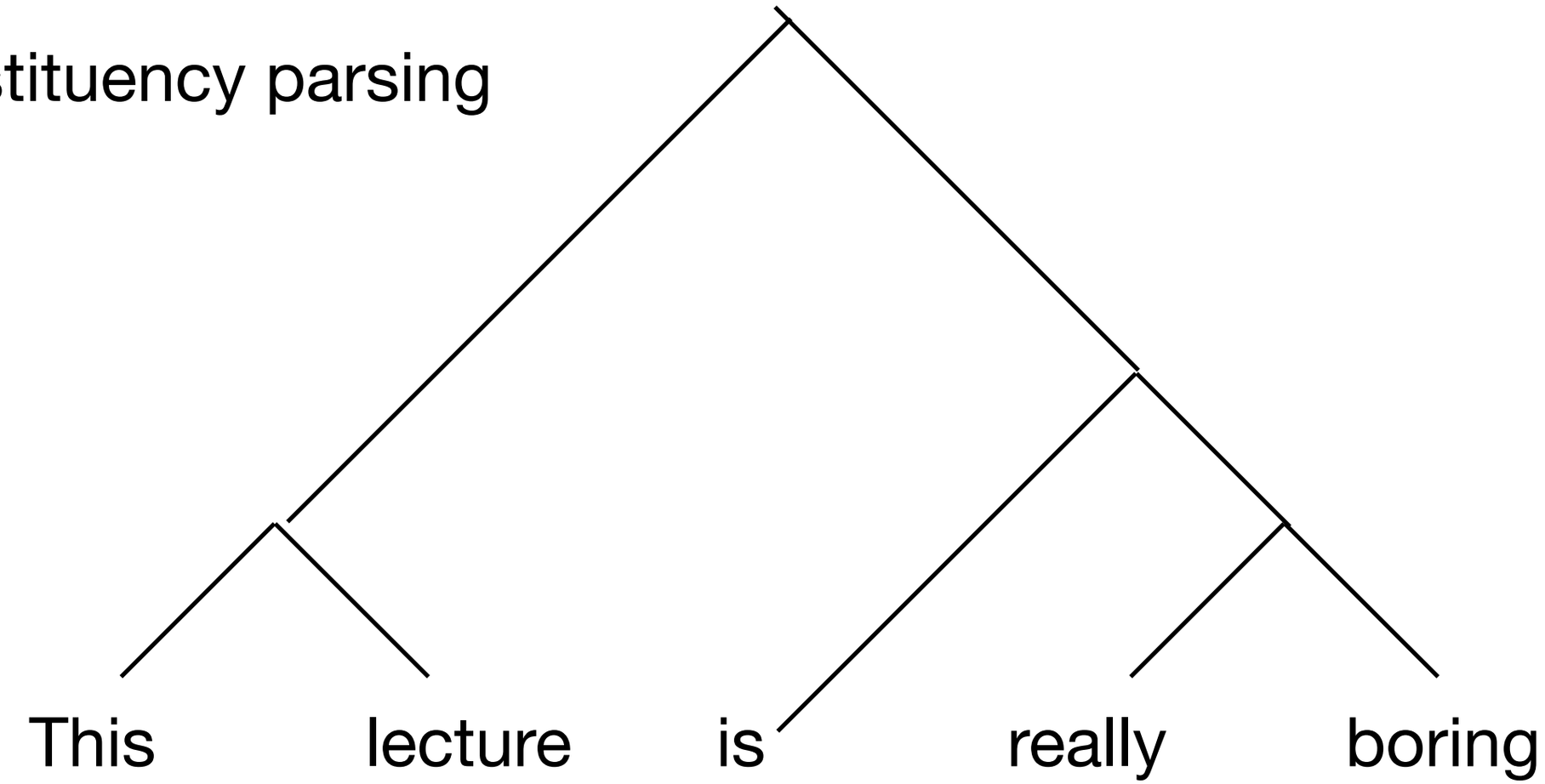
- Dependency parsing



[For each word, predict which word it depends on with what type]

Applying the single classifier to all NLP tasks

- Constituency parsing



[For each consecutive components, predict if you'd like to further combine them and possibly what type?]

Drawbacks of LR/Softmax

- Classification is non-linear
  - May not even represented as fixed-dimensional features
- Do not consider the relationship of labels within one data sample

The	lecture	is	really	boring
determiner	?	verb	adverb	adjective

Three	professors	lecture	IntroNLP
CardinalNumber	Noun	?	ProperNoune

https://www.merriam-webster.com/dictionary/lecture

lecture noun

lec·ture | \ ˈlek-chər, -shər\

Definition of *lecture* (Entry 1 of 2)

1 : a discourse given before an audience or c

2 : a formal reproof

lecture verb

lectured; lecturing \ ˈlek-chə-riŋ, ˈlek-shriŋ\

Definition of *lecture* (Entry 2 of 2)

intransitive verb

Ungraded Homework

- Derive Naïve Bayes, logistic regression again without referring to other materials
- Derive softmax derivative yourself
- Derive the decision boundary of LR/softmax
  - Decision boundary of Class i and Class j:

$$\{\boldsymbol{x} \in \mathbb{R}^n : p(Y = i \,|\, \boldsymbol{x}) = p(Y = j \,|\, \boldsymbol{x})\}$$

Do not submit ungraded homework

- Reading
  - Ch 4. Bishop, *Pattern Recognition and Machine Learning*
  - Ch 12. Jurafsky and Martin, *Speech and Language Processing*

CMPUT 651 (Fall 2019)

Thank you!

Q&A

UNIVERSITY OF ALBERTA