FNLP
Classification – Supervised Models

Yansong Feng
fengyansong@pku.edu.cn

Wangxuan Institute of Computer Technology
Peking University

February 26, 2025

Recap: Machine Learning Approaches to WSD

- Supervised learning :
  - a classification task
  - sense inventories are known
  - annotated data available for training
- Unsupervised learning :
- Semi-Supervised learning :

Naïve Bayes for WSD

The classifier we are interested in:

$$s^* = \arg\max_{s^*} P(C|s_k)P(s_k)$$

We need to estimate $P(C|s_k)$ and $P(s_k)$.

Make Independent Assumption w.r.t. $v_x \in C$:

$$P(C|s_k) = P(\{v_x|v_x \in C\}|s_k) = \prod_{v_x \in C} P(v_x|s_k)$$

- context features are assumed to be independent
- one word is independent from the other
  - <u>`THIS IS NOT TRUE!`</u>
  - but we then have an easier model

Naïve Bayes for WSD

Estimations:

the conditional:

$$P(v_x|s_k) = \frac{Count(v_x, s_k)}{Count(s_k)}$$

the prior:

$$P(s_k) = \frac{Count(s_k)}{Count(w)}$$

Testing: given an unseen instance with context C'

- for all context features v_x in C', compute:

$$P(s_k|C') \propto P(C'|s_k)P(s_k) = \prod_{v_x \in C'} P(v_x|s_k)P(s_k)$$

- choose s^*:

$$s^* = \arg\max_{s_k} \prod_{v_x \in C'} P(v_x|s_k)P(s_k)$$

Supervised Learning

- Naïve Bayes for WSD

$$sense^* = \arg\max_{sense} \prod_{word \in C} P(word|sense)P(sense)$$

Supervised learning :

- the label inventory is known
- annotated data available for training: $(x^1, y^1), (x^2, y^2)...(x^m, y^m)$ where x_i is the data sample (e.g., a sentence with a specified target word), y^i the corresponding label (e.g., the sense of the target word).

Supervised Learning

- Naïve Bayes for WSD

$$sense^* = \arg\max_{sense} \prod_{word \in C} P(word|sense)P(sense)$$

Supervised learning :

- the label inventory is known
- annotated data available for training: $(x^1, y^1), (x^2, y^2)...(x^m, y^m)$ where x_i is the data sample (e.g., a sentence with a specified target word), y^i the corresponding label (e.g., the sense of the target word).
- the goal: find a function $g : y = g(x)$

Two Views

- Goal: find a function $g : y = g(x)$

Two Views

- Goal: find a function $g : y = g(x)$
- Probabilistically:
  - $p(y|x)$
  - $g(x) = \arg\max_y p(y|x)$

Two Views

- Goal: find a function $g : y = g(x)$
- Probabilistically:
  - $p(y|x)$
  - $g(x) = \arg\max_y p(y|x)$
- Two views:
  - discriminative models: learn $p(y|x)$ directly
  - generative models: learn $p(x, y)$ first

Questions

- Is Naïve Bayes a generative model or discriminative model?

Questions

- Is Naïve Bayes a generative model or discriminative model?
- what about large language models?

Generative Models

Look at Joint Probabilities over both observed data x and classes y:

- Learn a conditional distribution $p(y|x)$, but start from $p(x, y)$

Generative Models

Look at Joint Probabilities over both observed data x and classes y:

- Learn a conditional distribution $p(y|x)$, but start from $p(x,y)$
- Bayes rule: $p(x,y) = p(y)p(x|y)$

Generative Models

Look at Joint Probabilities over both observed data x and classes y:

- Learn a conditional distribution $p(y|x)$, but start from $p(x, y)$
- Bayes rule: $p(x, y) = p(y)p(x|y)$
- Then:

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'} p(y')p(x|y')}$$

Generative Models

Look at Joint Probabilities over both observed data x and classes y:

- Learn a conditional distribution $p(y|x)$, but start from $p(x, y)$
- Bayes rule: $p(x, y) = p(y)p(x|y)$
- Then:

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'} p(y')p(x|y')}$$

- We get:

$$g(x) = \arg\max_y p(y|x) \tag{1}$$

$$= \arg\max_y \frac{p(y)p(x|y)}{\sum_{y'} p(y')p(x|y')} \tag{2}$$

$$= \arg\max_y p(y)p(x|y) \tag{3}$$

Generative Models

Look at Joint Probabilities over both observed data x and classes y:

- Learn a conditional distribution $p(y|x)$, but start from $p(x, y)$
- Bayes rule: $p(x, y) = p(y)p(x|y)$
- Then:

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'} p(y')p(x|y')}$$

- We get:

$$\begin{aligned} g(x) &= \arg\max_y p(y|x) && (1) \\ &= \arg\max_y \frac{p(y)p(x|y)}{\sum_{y'} p(y')p(x|y')} && (2) \\ &= \arg\max_y p(y)p(x|y) && (3) \end{aligned}$$

- Finally: $g(x) = \arg\max_y p(y)p(x|y) = \arg\max_y p(x, y)$

Example: Bayesian Modeling

Bayesian Modeling

Applying Bayes rule to the unknown variables of a data modeling problem

Usually, we are interested in two aspects:

- Data generation distribution: $x \sim p(x|y)$
- Model prior: $y \sim p(y)$

Example: Bayesian Modeling

Bayesian Modeling

Applying Bayes rule to the unknown variables of a data modeling problem

Usually, we are interested in two aspects:

- Data generation distribution: $x \sim p(x|y)$
- Model prior: $y \sim p(y)$

The goal is to learn y:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$$

Example: Bayesian Modeling

Bayesian Modeling

Applying Bayes rule to the unknown variables of a data modeling problem

Usually, we are interested in two aspects:

- Data generation distribution: $x \sim p(x|y)$
- Model prior: $y \sim p(y)$

The goal is to learn y:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$$

but, usually:

- Data generation distribution, $p(x|y)$, is **complicated**
- Analytically solving $p(y|x)$ can be **intractable**

Example: Bayesian Modeling

So, we have to **make assumptions**.

Example (Book-Category)

- Y: book category
- X: words in a book
- $P(X|Y)$: what kind of words a detective story will use
- $P(Y|X)$: what we are interested in finally

Example: Bayesian Modeling

So, we have to **make assumptions**.

Example (Book-Category)

- Y: book category
- X: words in a book
- $P(X|Y)$: what kind of words a detective story will use
- $P(Y|X)$: what we are interested in finally

Example: Bayesian Modeling

So, we have to **make assumptions**.

Example (Book-Category)

- Y: book category
- X: words in a book
- $P(X|Y)$: what kind of words a detective story will use
- $P(Y|X)$: what we are interested in finally

We need a story:

Book-Category

- $Y \sim Multinomial(\gamma)$
- $X|Y \sim p(X|\theta_Y)$

Example: Bayesian Modeling

So, we have to **make assumptions**.

Example (Book-Category)

- Y: book category
- X: words in a book
- $P(X|Y)$: what kind of words a detective story will use
- $P(Y|X)$: what we are interested in finally

We need a story:

Book-Category

- $Y \sim Multinomial(\gamma)$
- $X|Y \sim p(X|\theta_Y)$

That will give us: $p(X, Y|\gamma, \Theta) = p(X|Y, \Theta)p(Y|\gamma)$

Example: Bayesian Modeling

So, we have to **make assumptions**.

Example (Book-Category)

- Y: book category
- X: words in a book
- $P(X|Y)$: what kind of words a detective story will use
- $P(Y|X)$: what we are interested in finally

We need a story:

Book-Category

- $Y \sim Multinomial(\gamma)$
- $X|Y \sim p(X|\theta_Y)$
- Prior: $\gamma \sim Dirichlet(\beta)$
- $\theta_1, \theta_2, ..., \stackrel{iid}{\sim} p(\theta)$

That will give us: $p(X, Y|\gamma, \Theta) = p(X|Y, \Theta)p(Y|\gamma)$

Bonus: Bayesian Modeling

Here is an typical example of Bayesian Modeling for a corpus

- This is an **unsupervised learning model**
- This is an **unsupervised learning model**
- This is an **unsupervised learning model**

Bonus: Bayesian Modeling

Here is an typical example of Bayesian Modeling for a corpus

Latent Dirichlet Allocation

- This is an **unsupervised learning model**
- This is an **unsupervised learning model**
- This is an **unsupervised learning model**

Bonus: Bayesian Modeling

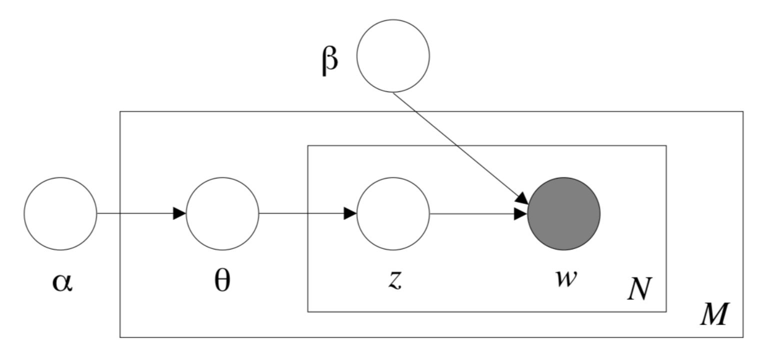
Here is an typical example of Bayesian Modeling for a corpus

Latent Dirichlet Allocation

for each document in the corpus

- choose N randomly or from a distribution
- choose $\theta \sim Dirichlet(\alpha)$
- for each of the N words:
  - choose a topic $z \sim Multinomial(\theta)$
  - choose a word w from $p(w|z, \beta)$

Bonus: Bayesian Modeling



Bonus: Bayesian Modeling

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions
- easier to incorporate various features

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions
- easier to incorporate various features
- usually achieve higher accuracy

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

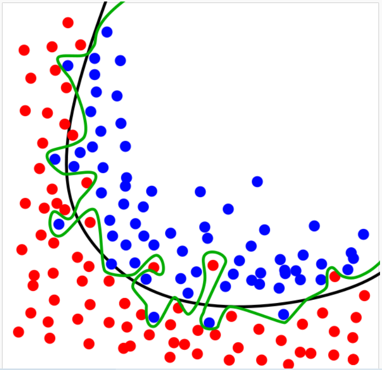
- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions
- easier to incorporate various features
- usually achieve higher accuracy
- may suffer from **overfitting**

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions
- easier to incorporate various features
- usually achieve higher accuracy
- may suffer from **overfitting**

the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.



Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions
- easier to incorporate various features
- usually achieve higher accuracy
- may suffer from **overfitting**

A lot of successful models:

- Maximum Entropy, Conditional Random Field, Logistic Regression, ...

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions
- easier to incorporate various features
- usually achieve higher accuracy
- may suffer from **overfitting**

A lot of successful models:

- Maximum Entropy, Conditional Random Field, Logistic Regression, ...
- more but not essentially probabilistic ones: SVM, perceptron, ...

Discriminative Models

Look at Conditional probabilities of unseen class labels y given observed data x only:

- model the conditional probabilities $p(y|x)$ directly
- no complusory independent assumptions
- easier to incorporate various features
- usually achieve higher accuracy
- may suffer from **overfitting**

A lot of successful models:

- Maximum Entropy, Conditional Random Field, Logistic Regression, ...
- more but not essentially probabilistic ones: SVM, perceptron, ...

But, how to make it happen?