Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

4.Learning under Bias

4.1 Semi-Supervised Learning as Transfer Learning

4.2 Transferring Data

理工学研究科情報工学専攻
國井慎也

Learning under Bias

- Three types of Bias
  - Sampling bias
    - A corpus of sentences of less than 40 words.
      - Overrepresent short sentences
      - Contain less connectives or complementizers
  - Distribution bias
    - Parsing dialect using annotated standard corpus.
  - Problem bias
    - Concept drift
      - the predictions become less accurate as time passes

$P(x), P(y), P(y|\boldsymbol{x})$

- $P(y|\boldsymbol{x})$
  - The source data may reflect a different conditional distribution $P(y|\boldsymbol{x})$ than the target data.
    - would mean that the bias is not just sampling bias
- $P(x)$
  - The bias may be a difference in the marginal distribution $P(\boldsymbol{x})$.

The covariate shift assumption

- $P(y)$
  - The data bias may be a bias in $P(y)$.

Class imbalance assumption

$$P(x), P(y), P(y|\boldsymbol{x})$$

	P(**x**)	P(y)	P(y|**x**)
sampling bias	MAYBE	MAYBE	NO
distribution bias	MAYBE	MAYBE	MAYBE
problem bias	MAYBE	MAYBE	YES

- In NLP domain adaptation
  - Without assuming a bias in $P(y|\boldsymbol{x})$
  - Most studies assume a bias in $P(\boldsymbol{x})$
  - In WSD, the class imbalance problem P(y) had attracted.

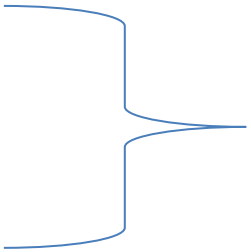
Semi-supervised Learning as Transfer Learning

- Semi-supervised learning may correct data bias,
- but its successfulness depends on the target distribution.
- Semi-supervised learning only works if the bias is relatively small.

- Labeled data is scarce and biased.
- Semi-supervised learning can correct bias when e.g. KL-divergence is relatively small.

Transferring data

- We want to built a model which beyond Semi-supervised learning.
  - Data point
  - Features
  - Parameters



transfer to target domain

We use some instance, some features and some parameters when learning a model for the target.

Transferring data

- The maximum likelihood
  - Do not necessary provide a good influence.

- Shimodaira proposes

$$-\sum_{i=1}^{n} -w_i(\boldsymbol{x}_i)\log(y_i|x_i,\theta)$$

$$w(\cdot) = \frac{P_T(\boldsymbol{x})}{P_S(\boldsymbol{x})}$$

but we can't compute density functions.

Importance weighting

- You can obtain an estimated importance weight function by:
  - kernel mean matching (Huang et al., 2007),
  - minimizing KL-divergence (Sugiyama et al., 2007).

$$\boldsymbol{KL}(\boldsymbol{w}(\boldsymbol{x})\widehat{\boldsymbol{P}}_{\boldsymbol{s}}(\boldsymbol{x}), \widehat{\boldsymbol{P}}_{\boldsymbol{T}}(\boldsymbol{x}))$$

experiment

- data: 20newsgroup
- 20 passes
- The weight function is estimated by logistic regression.

Source	Target	NB	W-NB	Perc.	W-Perc.
Hockey-IBM	Baseball-Mac	94.76	95.14	86.32	90.28
Hockey-Crypt	Motorcycles-Electronics	88.99	90.63	76.58	77.22
Motorcycles-Electronics	MidEast-Medicine	72.93	65.16	69.69	71.24
Graphics-Misc(Politics)	Windows-Misc(Religion)	94.58	95.36	89.16	89.94