Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 9

Administrativia

- Next class will be a summary
- Please email me questions about the material by Thursday 5pm
- Minor changes to the schedule. Those who are affected know it.
- Final schedule on website, please take a quick look

Last class

Unsupervised learning:

learning only from inputs

$$L(x_1, \ldots x_n | \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \underbrace{p(x_i|\theta)}_{\sum_y p(x_i, y|\theta)} \Rightarrow$$

find local maximum say using EM. Can also use Viterbi/Hard EM

Latent-variable learning:

Observe $(x_1, y_1), \ldots (x_n, y_n)$

Model has "h" - latent state

$$L(x_1, y_1, \ldots x_n, y_n|\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \underbrace{p(x_i, y_i|\theta)}_{\sum_h p(x_i, y_i, h|\theta)}$$

Semi-supervised Learning

Main idea: use a relatively small amount of annotated data, and exploit also large amounts of unannotated data

The term itself is used in various ways with various methodologies

Example: Word Clusters and Embeddings

- Learn clusters of words or embed them in Euclidean space using large amounts of text

- Use these clusters/embeddings as features in a discriminative model

$C : V \to [200]$

learn to map a word to an integer $1 \ldots 200$

use a feature

$$f_j(x, y) = \begin{cases} C(x_i) = 5 \\ \wedge\ y_i = \text{NN} \end{cases}$$

$M : V \to \mathbb{R}^d$

$\|M(\text{dog}) - M(\text{cat})\|_2$ - small?

Semi-supervised Learning: Example 2

Combine the log-likelihood for labelled data with the log-likelihood for unlabelled data

$$L(x_1, y_1, \ldots, x_n, y_n, x'_1, \ldots, x'_m|\theta) =$$

y_1: full data; $x'_1, \ldots, x'_m$: partial data

$$L(x_1, y_1, \ldots x_n, y_n|\theta) + \lambda L(x'_1, \ldots, x'_m|\theta)$$

$$= \sum_{i=1}^{n} \log p(x_i, y_i|\theta) + \lambda \sum_{i=1}^{m} \log p(x'_i|\theta)$$

$$= \sum_{i=1}^{n} \log p(x_i, y_i|\theta) + \lambda \sum_{i=1}^{m} \log \sum_{y} p(x'_i|\theta)$$

Semi-supervised Learning: Example 3

Self-training

① learn a model from labelled data

② label unlabelled data (using inference)

③ learn ~~everything~~ a model from all "complete" data

Semi-supervised Learning: Example 3

Self-training

Step 1:

Step 2:

Step 3:

Potentially, repeat step 2

Today's class

Evaluation and experimental design

- Testing for error
- Hypothesis testing

Sample complexity

Next step after prediction

- We learned/estimated a model
- We decoded and made predictions (inference)
- What's next?

Next step after prediction

- We learned/estimated a model
- We decoded and made predictions (inference)
- What's next?

Evaluation

NLP as an empirical science

NLP is an empirical science

As such, conclusions should be carefully drawn from the conducted experiments

The most common type of empirical issue NLP research addresses: comparison of methods

Improving state of the art

We have two competing methods. How do we decide which method is better?

We have two sets of predictions from each method:

$(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$

$(x_1, y_1'), \cdots (x_n', y_n')$

We have an evaluation metric for error $\text{err}(y|\text{method}_1)$, $\text{err}(y|\text{method}_2)$

We compute

$$\text{err}_1 = \frac{1}{n} \sum_{i=1}^{n} \text{err}(y_i | \text{method}_1)$$

$$\text{err}_2 = \frac{1}{n} \sum_{i=1}^{n} \text{err}(y_i' | \text{method}_2)$$

We ask, is $\text{err}_1 < \text{err}_2$?

Validity of error comparison

The empirical error as a proxy for the "true" error

Empirical error: $\text{err}_1 = \frac{1}{n}\sum_{i=1}^{n} \text{err}(y_i^1 | \text{method}_1)$

approximation

"True" error: $\text{err}^* = E_p[\text{err}(y | \text{method}_1)]$
$E_p[\text{err}(y | \text{method}_2)]$

We assume some true distribution $p(x,y)$

$E_p[\text{err}(y|\text{method}_1)) =$

$= \sum_{x,y} p(x,y)\, \text{err}(\hat{y}|\text{method}_1)$

Validity of error comparison

The empirical error as a proxy for the “true” error

Empirical error: $\text{err}_1 = \frac{1}{n}\sum_{i=1}^{n} \text{err}(y_i^1 | \text{method}_1)$

“True” error: $\text{err}^* =$

The empirical error is a random variable, that depends on the training data we have

Error comparison

Say your system's empirical error is 25.6% and my system's error is 35.5%. Is your system better?

Error comparison

Say your system's empirical error is 25.6% and my system's error is 35.5%. Is your system better?

What does it mean for the system to be better?

Error comparison

Say your system's empirical error is 25.6% and my system's error is 35.5%. Is your system better?

What does it mean for the system to be better?

Say your system's empirical error is 25.6% and my system's error is 28.5%. Is your system better?

Error comparison

Say your system's empirical error is 25.6% and my system's error is 35.5%. Is your system better?

What does it mean for the system to be better?

Say your system's empirical error is 25.6% and my system's error is 28.5%. Is your system better?

Say your system's empirical error is 25.6% and my system's error is 25.7%. Is your system better?

State of the art in NLP

Sometimes improvements reported are very small

Over time, there is perhaps a big effect

Test case: machine translation

But how do we test the validity of a single improvement?

Hypothesis testing

Null hypothesis: our systems are the same

To reject this hypothesis, we need sufficient evidence

How do we look for that evidence?

Hypothesis testing

- Set α – significance level. Usually $\alpha = 0.05$ or $\alpha = 0.01$
- Calculate t on the test data, which is an estimate of a parameter (such as the difference between errors of the two systems) – "test" statistic
- Calculate a p-value – the probability of the t getting its value if the null hypothesis is true
- See if t is "surprising" if the null hypothesis is true, i.e. $p < \alpha$. If so, reject the null hypothesis

Example: Sign test

We want to test whether two systems are the same, or one is better than the other

We list all errors: $\text{err}(y_i|\text{method}_1)$ and $\text{err}(y_i|\text{method}_2)$

method 1	method 2		
0.7	0.3	>	
0.3	0.4	<	
0.1	0.2	<	
0.2	0.2	=	
0.7	0.6	>	

> -2

< -2

≥ -5.0

< -3

Sign test: we assume in the null hypothesis that the probabilities of getting 1 or 0 are the same (0.5)

$$p(>= k, <= n-k) = \binom{n}{k}\left(\frac{1}{2}\right)^n$$

Null hypothesis is rejected if:

A theoretical way to quantify error

We learn from data

The more data we have, the better (usually)

But how better do we get with more data?

This is answered through the notion of sample complexity

Sample complexity

A measure for the number of samples that are required in order to get a desired error level

What will a sample complexity depend on?

ε - error level

d - complexity of the model

Sample complexity bounds

We usually cannot get the actual number for a desired error level

Instead we get a bound

These bounds are often loose and not so informative

They give an idea about asymptotic behavior, though

Back to pre-historic languages

We are observing $x_1, \ldots, x_n$ where $x_i \in \{\text{argh}, \text{blah}\}$

Estimation of θ:

$$\frac{\text{count}(\text{argh})}{n}$$

$$z_i = \begin{cases} 1 & \text{if } x_i = \text{argh} \\ 0 & \text{o/w} \end{cases}$$

$$\theta^* = \frac{\sum z_i}{n}$$

Back to pre-historic languages

θ_0 - "true" parameter

$\theta^* = \frac{\sum_{i=1}^{n} z_i}{n}$ where $z_i = 1$ iff $x_i = \text{argh}$.

An average of binary random variables

Hoeffding's inequality:

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \theta_0\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right) \leq \delta \quad 0.05 \quad 0.01$$

How to interpret this inequality?

$$2\exp\left(-\frac{n\varepsilon^2}{2}\right) \leq \delta$$

$$\log 2 - \frac{n\varepsilon^2}{2} \leq \log \delta$$

$$n \geq (\log \delta/2) \times 2/\varepsilon^2$$

Back to pre-historic languages

$\theta^* = \frac{\sum_{i=1}^{n} z_i}{n}$ where $z_i = 1$ iff $x_i = \text{argh}$.

An average of binary random variables

Hoeffding's inequality:

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \theta_0\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

How to interpret this inequality?

We want the right handside to be smaller than some δ:

Probably approximately correct

With probability $1 - \delta$ (want δ to be small), if

$$n \geq 2 \log \frac{\delta}{2} \times \frac{1}{\epsilon^2}$$

then $|\theta^* - \theta_0| < \epsilon$, where θ_0 is the true parameter, and θ^* is the maximum likelihood estimate.

Probably Approximately correct

Final note about evaluation

Not all metrics are the same

When designing a metric, see if:

- It correlates with human judgements? (the case of BLEU in MT)
- It correlates with the error of a downstream application? (the case of perplexity of LM and word error rate in speech)

Next class

Summary

Please send me questions by Thursday (12/2), 5pm