CS412 office hour

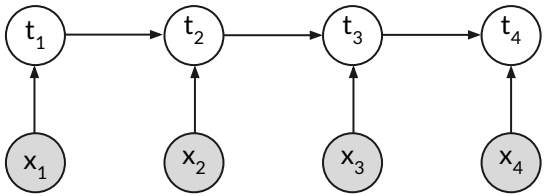
Apr 26, 2019

Today’s Office Hour

- QA

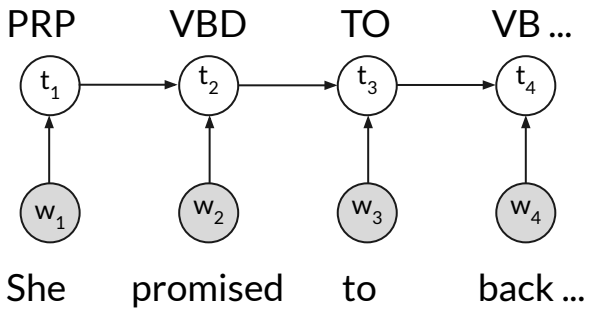
Plate Notation

- Handy for repetitions
  - E.g. you have a set of documents, each has a few R.V.
- Not great for complex models. Cannot handle interdependent repetitions.
  - E.g . dependencies between time steps



Observed vs hidden variables

- Observed: What you see in your input
- Hidden: What you don't see in your input
  - E.g . POS tagging
  - Observed: words
  - Hidden: POS tag for each word



Imbalanced Classes in Binary Classification

- Rare positive examples but numerous negative ones.
  - Cancer Diagnosis
  - Credit Card Fraud Detection
  - Fire alarm
- Objectives:
  - Maximize true positive.
  - False positive > False Negative.
  - Creating a fraud alert at a correct transaction is preferred than missing fraudulent transactions.

Training paradigms on Imbalanced Data

- OverSampling - Re-sampling of data from positive class
  - Total Observations = 1000 , Fraudulent Observations =20, Non Fraudulent Observations = 980
  - Event Rate= 2 %
  - In this case we are replicating 20 fraud observations 20 times.
  - Fraudulent Observations after replicating the minority class observations= 400
  - Total Observations in the new data set after oversampling=1380
  - Event Rate for the new data set after over sampling= 400/1380 = 29 %

https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

Sampling Paradigms

- UnderSampling - Randomly eliminate tuples from negative class
  - Total Observations = 1000, Fraudulent Observations =20, Non Fraudulent Observations = 980
  - Event Rate= 2 %
  - Take 10% samples without replacement from Non Fraud instances. And combining them with Fraud instances.
  - Non Fraudulent Observations after random under sampling = 10% of 980 =98
  - Total Observations after combining them with Fraudulent observations = 20+98=118
  - Event Rate for the new dataset after under sampling = 20/118 = 17%

Comparison between Sampling Techniques

	OverSampling	UnderSampling
Advantages	no information loss	● Improve run time ● Reduce storage if training data is huge.
Disadvantages	Increases the likelihood of overfitting as replicate positive class.	● Can discard potentially useful information ● Chosen random sample maybe biased. Not accurate representation of training data.

Other Techniques

- Threshold-moving
  - Move the decision threshold so that the positive class samples are easier to classify.
  - Less chance of costly false negative errors, can increase false positive error.
- Ensemble techniques
  - Ensemble multiple classifiers to take majority decision.
  - Each classifier has all positive class samples and sample of negative class samples.