Statistical NLP: Lecture 4

Mathematical Foundations I:
Probability Theory

(Ch2)

1

Notions of Probability Theory

- ***Probability theory*** deals with predicting how likely it is that something will happen.
- The process by which an observation is made is called an ***experiment*** or a ***trial***.
- The collection of ***basic outcomes*** (or ***sample points***) for our experiment is called the ***sample space***.
- An ***event*** is a subset of the sample space.
- Probabilities are numbers between 0 and 1, where 0 indicates impossibility and 1, certainty.
- A ***probability function/distribution*** distributes a probability mass of 1 throughout the sample space.

2

Conditional Probability and Independence

- ***Conditional probabilities*** measure the probability of events given some knowledge.
- ***Prior probabilities*** measure the probabilities of events before we consider our additional knowledge.
- ***Posterior probabilities*** are probabilities that result from using our additional knowledge.
- The ***chain rule*** relates intersection with conditionalization (important to NLP)
- ***Independence*** and ***conditional independence*** of events are two very important notions in statistics.

3

Bayes' Theorem

- ***Bayes' Theorem*** lets us swap the order of dependence between events. This is important when the former quantity is difficult to determine.
- ***P(B|A) = P(A|B)P(B)/P(A)***
- P(A) is a ***normalization constant***.

4

Random Variables

- A ***random variable*** is a function
  X: sample space --> R^n
- A ***discrete random variable*** is a function
  X: sample space --> S
  where S is a countable subset of R.
- If X: sample space --> {0,1}, then X is called a ***Bernoulli trial***.
- The ***probability mass function*** for a random variable X gives the probability that the random variable has different numeric values.

5

Expectation and Variance

- The ***expectation*** is the ***mean*** or average of a random variable.
- The ***variance*** of a random variable is a measure of whether the values of the random variable tend to be consistent over trials or to vary a lot.

6

Joint and Conditional Distributions

- More than one random variable can be defined over a sample space. In this case, we talk about a ***joint*** or ***multivariate*** probability distribution.
- The ***joint probability mass function*** for two discrete random variables X and Y is: p(x,y)=P(X=x, Y=y)
- The ***marginal probability mass function*** totals up the probability masses for the values of each variable separately.
- Similar intersection rules hold for joint distributions as for events.

7

Estimating Probability Functions

- What is the probability that the sentence "The cow chewed its cud" will be uttered? Unknown ==> P must be ***estimated*** from a sample of data.
- An important measure for estimating P is the ***relative frequency*** of the outcome, i.e., the proportion of times a certain outcome occurs.
- Assuming that certain aspects of language can be modeled by one of the well-known distribution is called using a ***parametric*** approach.
- If no such assumption can be made, we must use a ***non-parametric*** approach.

8

Standard Distributions

- In practice, one commonly finds the same basic form of a probability mass function, but with different constants employed.
- Families of pmfs are called ***distributions*** and the constants that define the different possible pmfs in one family are called ***parameters***.
- Discrete Distributions: the ***binomial distribution***, the ***multinomial distribution***, the ***Poisson distribution***.
- Continuous Distributions: the ***normal distribution***, the ***standard normal distribution***.

9

Bayesian Statistics I: Bayesian Updating

- Assume that the data are coming in sequentially and are independent.
- Given an a-priori probability distribution, we can update our beliefs when a new datum comes in by calculating the ***Maximum A Posteriori (MAP)*** distribution.
- The MAP probability becomes the new prior and the process repeats on each new data.

10

Bayesian Statistics II: Bayesian Decision Theory

- Bayesian Statistics can be used to evaluate which model or family of models better explains some data.
- We define two different models of the event and calculate the ***likelihood ratio*** between these two models.

11