Mathematical Foundations

Elementary Probability Theory

Essential Information Theory

Updated 03/2008
Motivations

- Statistical NLP aims to do statistical inference for the field of NL.

- *Statistical inference* consists of taking some data (generated in accordance with some unknown probability distribution) and then making some inference about this distribution.
An example of statistical inference is the task of *language modeling* (ex how to predict the next word given the previous words)

In order to do this, we need a *model* of the language.

Probability theory helps us finding such model
Probability Theory

- How likely it is that something will happen
- Sample space $\Omega$ is listing of all possible outcomes of an experiment
- Event $A$ is a subset of $\Omega$
- Probability function (or distribution)

$$P : \Omega \rightarrow [0, 1]$$
Prior Probability

- *Prior probability*: the probability before we consider any additional knowledge

\[ P(A) \]
Conditional probability

- Sometimes we have partial knowledge about the outcome of an experiment
- Conditional (or Posterior) Probability
- Suppose we know that event B is true
- The probability that A is true given the knowledge about B is expressed by
  \[ P(A|B) \]
Conditional probability (cont)

\[ P(A \cap B) = P(A | B) \cdot P(B) = P(B | A) \cdot P(A) \]

- Joint probability of A and B.
- 2-dimensional table with a value in every cell giving the probability of that specific state occurring.
Chain Rule

\[ P(A \cap B) = P(A \mid B)P(B) \]
\[ = P(B \mid A)P(A) \]

\[ P(A \cap B \cap C \cap D \ldots) = P(A)P(B \mid A)P(C \mid A, B)P(D \mid A, B, C \ldots) \]
(Conditional) independence

- Two events $A$ and $B$ are independent of each other if
  \[ P(A) = P(A|B) \]

- Two events $A$ and $B$ are conditionally independent of each other given $C$ if
  \[ P(A|C) = P(A|B,C) \]
Bayes' Theorem

- Bayes' Theorem lets us swap the order of dependence between events
- We saw that $P(A|B) = P(A \cap B)/P(B)$
- Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
Example - find web pages about “NLP”

- T: positive test, N: page about ‘NLP’
- P(T|N) = 0.95, P(N) = 1/100,000
  P(T|~N) = 0.005

- System points a page as relevant. What is the probability it is about NLP

\[
P(N | T) = \frac{P(T | N)P(N)}{P(T)}
\]

\[
= \frac{P(T | N)P(N)}{P(T | N)P(N) + P(T | \sim N)P(\sim N)}
\]

\[
= 0.002
\]
Random Variables

- So far, event space that differs with every problem we look at
- Random variables (RV) $X$ allow us to talk about the probabilities of numerical values that are related to the event space

$$X : \Omega \rightarrow \mathbb{R}$$

$$X : \Omega \rightarrow \mathbb{S}$$
The Expectation of a RV is

\[ p(x) = p(X = x) = p(A_x) \]

\[ A_x = \{ \omega \in \Omega : X(\omega) = x \} \]

\[ \sum_x p(x) = 1 \quad 0 \leq p(x) \leq 1 \]

- The Expectation of a RV is

\[ E(x) = \sum_x xp(x) = \mu \]
Variance

- The variance of a RV is a measure of the deviation of values of the RV about its expectation.

\[
Var(X) = E((X - E(X))^2)
\]

\[
= E(X^2) - E^2(X) = \sigma^2
\]

- \(\sigma\) is called the standard deviation.
In general, for language events, $P$ is unknown.

We need to estimate $P$, (or model $M$ of the language).

We’ll do this by looking at evidence about what $P$ must be based on a sample of data.
Estimation of $P$

- Frequentist statistics
- Bayesian statistics
Frequentist Statistics

- Relative frequency: proportion of times an outcome $u$ occurs

$$f_u = \frac{C(u)}{N}$$

- $C(u)$ is the number of times $u$ occurs in $N$ trials
- For $N \to \infty$ the relative frequency tends to stabilize around some number: probability estimates
- Difficult to estimate if the number of different values $u$ is large
Frequentist Statistics (cont)

- Two different approaches:
  - Parametric
  - Non-parametric (distribution free)
Parametric Methods

- Assume that some phenomenon in language is acceptably modeled by one of the well-known family of distributions (such binomial, normal)
- We have an explicit probabilistic model of the process by which the data was generated, and determining a particular probability distribution within the family requires only the specification of a few parameters (less training data)
Non-Parametric Methods

- No assumption about the underlying distribution of the data
- For ex, simply estimate $P$ empirically by counting a large number of random events is a distribution-free method
- Less prior information, more training data needed
Binomial Distribution (Parametric)

- Series of trials with only two outcomes, each trial being independent from all the others
- Number \( r \) of successes out of \( n \) trials given that the probability of success in any trial is \( p \):

\[
b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r}
\]
Normal (Gaussian) Distribution (Parametric)

- **Continuous**
- **Two parameters:** mean $\mu$ and standard deviation $\sigma$

$$n(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
Consider sampling the height of 15 male dwarfs:

- Heights (in cm): 114, 87, 112, 76, 102, 72, 89, 110, 93, 127, 86, 107, 95, 123, 98.

How to model the distribution of dwarf heights?

E.g. what is the probability of meeting a dwarf more than 130cm high?
Parametric vs. non-parametric - example - cont

- Non parametric estimation: Smoothing

Histogram

Smoothing
Parametric vs. non-parametric - example - cont

- parametric estimation: modeling heights as a normal distribution. Only needs to estimate $\mu$ and $\sigma$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu = 99.4$
$\sigma = 16.2$
Frequentist Statistics

- **D**: data
- **M**: model (distribution $P$)
- **$\Theta$**: model parameters (e.g. $\mu$, $\sigma$)

For $M$ fixed: *Maximum likelihood estimate*: choose $\hat{\Theta}$ such that

$$\hat{\Theta} = \arg\max_{\Theta} P(D | M, \Theta)$$
Frequentist Statistics

- Model selection, by comparing the maximum likelihood: choose $\hat{M}$ such that

$$
\hat{M} = \arg\max_M P\left(D \mid M, \theta(M)\right)
$$

$$
\hat{\theta} = \arg\max_{\theta} P(D \mid M, \theta)
$$
Estimation of $P$

- Frequentist statistics
  - Parametric methods
    - Standard distributions:
      - Binomial distribution (discrete)
      - Normal (Gaussian) distribution (continuous)
        - Maximum likelihood
  - Non-parametric methods
- Bayesian statistics
Bayesian Statistics

- Bayesian statistics measures degrees of belief.
- Degrees are calculated by starting with prior beliefs and updating them in face of the evidence, using Bayes theorem.
Bayesian Statistics (cont)

\[ "^* \]
\[
\hat{M} = \arg\max_{M} P(M | D) \\
= \arg\max_{M} \frac{P(D | M)P(M)}{P(D)} \\
= \arg\max_{M} P(D | M)P(M)
\]

MAP is maximum a posteriori
Bayesian Statistics (cont)

- $M$ is the distribution; for fully describing the model, I need both the distribution $M$ and the parameters $\theta$

$$
\hat{M} = \text{argmax}_M P(D | M)P(M)
$$

$$
P(D | M) = \int P(D, \theta | M) d\theta
$$

$$
= \int P(D | M, \theta)P(\theta | M) d\theta
$$

$P(D | M)$ is the marginal likelihood
Frequentist vs. Bayesian

- **Bayesian**
  \[ M^* = \arg\max_M P(M) \int P(D | M, \theta) P(\theta | M) d\theta \]

- **Frequentist**
  \[ \theta^* = \arg\max_\theta P(D | M, \theta) \]
  \[ M^* = \arg\max_M P(D | M, \theta^*(M)) \]

\( P(D | M, \theta) \) is the likelihood
\( P(\theta | M) \) is the parameter prior
\( P(M) \) is the model prior
Bayesian Updating

- How to update $P(M)$?
- We start with a priori probability distribution $P(M)$, and when a new datum comes in, we can update our beliefs by calculating the posterior probability $P(M|D)$. This then becomes the new prior and the process repeats on each new datum.
Bayesian Decision Theory

Suppose we have 2 models $M_1$ and $M_2$; we want to evaluate which model better explains some new data.

\[
\frac{P(M_1 | D)}{P(M_2 | D)} = \frac{P(D | M_1)P(M_1)}{P(D | M_2)P(M_2)}
\]

if \( \frac{P(M_1 | D)}{P(M_2 | D)} > 1 \) i.e. \( P(M_1 | D) > P(M_2 | D) \)

$M_1$ is the most likely model, otherwise $M_2$
Essential Information Theory

- Developed by Shannon in the 40s
- Maximizing the amount of information that can be transmitted over an imperfect communication channel
- Data compression (entropy)
- Transmission rate (channel capacity)
Entropy

- **X**: discrete RV, p(X)
- Entropy (or self-information)
  \[ H(p) = H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \]

- Entropy measures the amount of information in a RV; it’s the average length of the message needed to transmit an outcome of that variable using the optimal code.
Entropy (cont)

\[ H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \]

\[ = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \]

\[ = E \left( \log_2 \frac{1}{p(x)} \right) \]

\[ H(X) \geq 0 \]

\[ H(X) = 0 \iff p(X) = 1 \]

i.e. when the value of \( X \) is determinate, there is a value \( x \) with \( p(x) = 1 \).
Joint Entropy

- The joint entropy of 2 RV $X, Y$ is the amount of the information needed on average to specify both their values

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$
Conditional Entropy

The conditional entropy of a RV $Y$ given another $X$, expresses how much extra information one still needs to supply on average to communicate $Y$ given that the other party knows $X$.

$$H(Y | X) = \sum_{x \in X} p(x)H(Y | X = x)$$

$$= - \sum_{x \in X} p(x) \sum_{y \in Y} p(y | x) \log p(y | x)$$

$$= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y | x) = -E(\log p(Y | X))$$
Chain Rule

\[ H(X, Y) = H(X) + H(Y | X) \]

\[ H(X_1, ..., X_n) = H(X_1) + H(X_2 | X_1) + ... + H(X_n | X_1, ..., X_{n-1}) \]
Mutual Information

\[ H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y) \]
\[ H(X) - H(X | Y) = H(Y) - H(Y | X) = I(X, Y) \]

- \( I(X, Y) \) is the mutual information between \( X \) and \( Y \). It is the reduction of uncertainty of one RV due to knowing about the other, or the amount of information one RV contains about the other.
Mutual Information (cont)

\[ I(X, Y) = H(X) - H(X | Y) = H(Y) - H(Y | X) \]

- I is 0 only when X, Y are independent: 
  \[ H(X|Y) = H(X) \]

- \( H(X) = H(X) - H(X|X) = I(X,X) \) Entropy is the self-information

- May be written as
  \[ I(X, Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]
Entropy and Linguistics

- Entropy is a measure of uncertainty. The more we know about something the lower the entropy.

- If a language model captures more of the structure of the language, then the entropy should be lower.

- We can use entropy as a measure of the quality of our models.
Entropy and Linguistics

\[ H(p) = H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \]

- \( H \): entropy of language; we don’t know \( p(X) \); so..?
- Suppose our model of the language is \( q(X) \)
- How good estimate of \( p(X) \) is \( q(X) \)?
Relative entropy or KL (Kullback-Leibler) divergence applies to two distributions $p$ and $q$

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

$$= E_p \left( \log \frac{p(X)}{q(X)} \right)$$
Entropy and Linguistics

- $D_{\text{kl}}(p||q)$ measures how different two probability distributions are.
- Average number of bits that are wasted by encoding events from a distribution $p$ with a code based on a not-quite-right distribution $q$.
- Goal: minimize relative entropy $D(p||q)$ to have a probabilistic model as accurate as possible.
The entropy of English - character entropy

Given 27 characters (a-z + space) measure the cross entropy $H(p,q) = -\sum p(x) \log q(x)$ of English.

How well does $q$ model the true distribution $p$.

Here $q$ is modeled as Markov model

$$H(P, q) \sim 1/n \sum_{i=1}^{n} -\log q(w_i|h)$$
The entropy of English – character entropy

$q$ is modeled as a Markov chain also know as n-gram model

<table>
<thead>
<tr>
<th>Model</th>
<th>cross entropy (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0^{th}$ order</td>
<td>4.76</td>
</tr>
<tr>
<td>$1^{st}$ order</td>
<td>4.03</td>
</tr>
<tr>
<td>$2^{nd}$ order</td>
<td>3.32</td>
</tr>
<tr>
<td>$2^{nd}$ order</td>
<td>3.1</td>
</tr>
<tr>
<td>$2^{nd}$ order</td>
<td>2.8</td>
</tr>
<tr>
<td>Shannon exp.</td>
<td>1.34</td>
</tr>
</tbody>
</table>