Mathematical Foundations

Elementary Probability Theory

Essential Information Theory

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.

Motivations (Cont)

- 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 Ω is listing of all possible outcome of an experiment
- Event A is a subset of Ω
- 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)

$$P(A,B) = P(A \mid B)P(B)$$
$$= P(B \mid A)P(A)$$

- 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

$$P(A,B) = P(A|B)P(B)$$
$$= P(B|A)P(A)$$

$$P(A,B,C,D\ldots) = P(A)P(B|A)P(C|A,B)P(D|A,B,C..)$$

(Conditional) independence

- Two events A e 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 $P(A|B) = \frac{P(A,B)}{P(B)}$
- Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example

- S:stiff neck, M: meningitis
- P(S|M) =0.5, P(M) = 1/50,000 P(S)=1/20
- I have stiff neck, should I worry?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

$$= \frac{0.5 \times 1/50{,}000}{1/20} = 0.0002$$

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

$$X : \Omega \rightarrow \Re$$

$$X : \Omega \rightarrow \Im$$

Expectation

$$p(x) = p(X = x) = p(A_x)$$

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

$$\sum_x p(x) = 1 \qquad 0 \le p(x) \le 1$$

- The Expectation is the *mean* or *average* of a RV

$$E(x) = \sum_x x p(x) = \mu$$

Variance

- The *variance* of a RV is a measure of whether the values of the RV tend to be consistent over trials or to vary a lot

$$Var(X) = E((X - E(X))^2)$$
$$= E(X^2) - E^2(X) = \sigma^2$$

- σ is the *standard deviation*

Back to the Language Model

- 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 $N \rightarrow \infty$ the relative frequency tends to stabilize around some number: probability estimates

Frequentist Statistics (cont)

- Two different approach:
  - 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 μ and standard deviation σ

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

- Used in clustering

Frequentist Statistics

- D: data
- M: model (distribution P)
- Θ: parameters (es μ, σ)
- For M fixed: *Maximum likelihood estimate*: choose $\overset{*}{\Theta}$ such that

$$\overset{*}{\theta} = \underset{\theta}{\operatorname{argmax}} P(D \mid M, \theta)$$

Frequentist Statistics

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

$$\overset{*}{M} = \underset{M}{\operatorname{argmax}}\, P\left(D \mid M, \overset{*}{\theta}(M) \right)$$

$$\overset{*}{\theta} = \underset{\theta}{\operatorname{argmax}} P(D \mid M, \theta)$$

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)

$$\overset{*}{M} = \underset{M}{\operatorname{argmax}} P(M \mid D)$$

$$= \underset{M}{\operatorname{argmax}} \frac{P(D \mid M)P(M)}{P(D)}$$

$$= \underset{M}{\operatorname{argmax}} P(D \mid M)P(M)$$

MAP!

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 θ

$$M^* = \underset{M}{\operatorname{argmax}} P(D \mid M) P(M)$$

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

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

P(D|M) is the marginal likelihood

Frequentist vs. Bayesian

- Bayesian

$$\overset{*}{M} = \underset{M}{\operatorname{argmax}} P(M) \int P(D \mid M, \theta) P(\theta \mid M) d\theta$$

- Frequentist

$$\overset{*}{\theta} = \underset{\theta}{\operatorname{argmax}} P(D \mid M, \theta) \qquad \overset{*}{M} = \underset{M}{\operatorname{argmax}} P\left(D \mid M, \overset{*}{\theta}(M) \right)$$

$P(D \mid M, \theta)$ is the likelihood

$P(\theta \mid M)$ 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.

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1)P(M_1)}{P(D \mid M_2)P(M_2)}$$

if $\dfrac{P(M_1 \mid D)}{P(M_2 \mid D)} > 1$ i.e $P(M_1 \mid D) > P(M_2 \mid D)$

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)

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

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

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

$$H(X) \geq 0$$

$$H(X) = 0 \Leftrightarrow p(X) = 1$$

i.e when the value of X is determinate, hence providing no new information

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

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

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

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

Chain Rule

$$H(X,Y) = H(X) + H(Y|X)$$

$$H(X_{1,},\ldots,X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1,\ldots X_{n-1})$$

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)

$$I(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

- 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

Entropy and Linguistics

- Entropy is 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)?

Entropy and Linguistic Kullback-Leibler Divergence

- Relative entropy or KL (Kullback-Leibler) divergence

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

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

Entropy and Linguistic

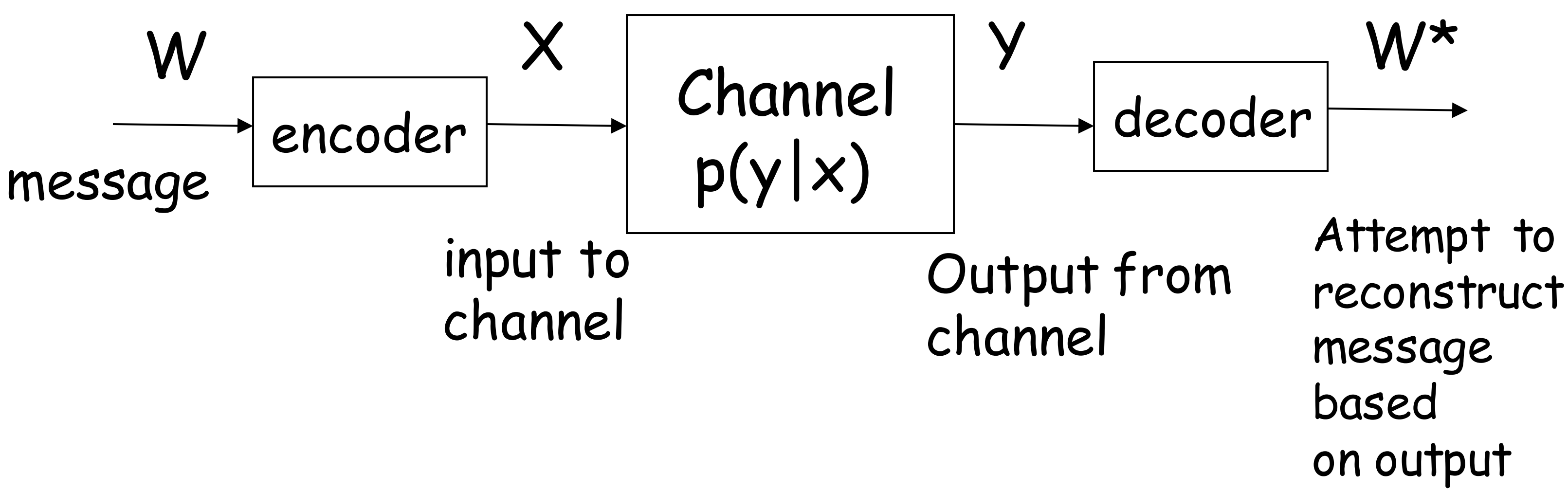
- Measure of how different two probability distributions are
- Average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite right distribution q
- Goal: minimize relative entropy D(p||q) to have a probabilistic model as accurate as possible

The Noisy Channel Model

- The aim is to optimize in terms of throughput and accuracy the communication of messages in the presence of noise in the channel
- Duality between compression (achieved by removing all redundancy) and transmission accuracy (achieved by adding controlled redundancy so that the input can be recovered in the presence of noise)

The Noisy Channel Model

- Goal: encode the message in such a way that it occupies minimal space while still containing enough redundancy to be able to detect and correct errors

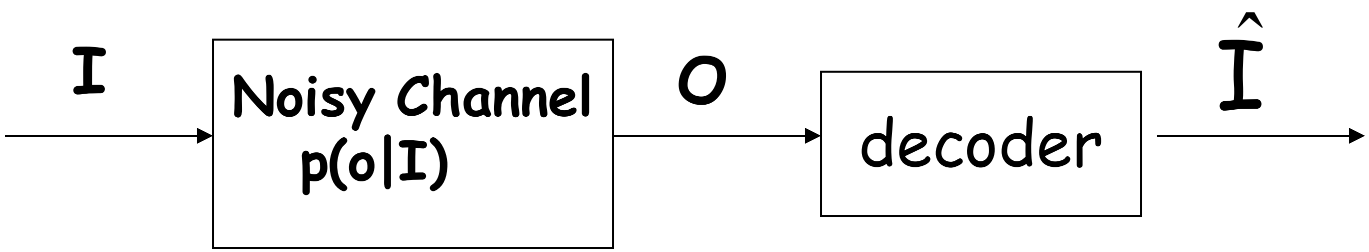


The Noisy Channel Model

- Channel capacity: rate at which one can transmit information through the channel with an arbitrary low probability of being unable to recover the input from the output
- $$C = \max_{p(X)} I(X;Y)$$
- We reach a channel capacity if we manage to design an input code X whose distribution p(X) maximizes I between input and output

Linguistics and the Noisy Channel Model

- In linguistic we can't control the encoding phase. We want to decode the output to give the most likely input.



$$\hat{I} = \underset{i}{\operatorname{argmax}}\ p(i|o) = \underset{i}{\operatorname{argmax}} \frac{p(i)p(o|i)}{p(o)} = \underset{i}{\operatorname{argmax}}\ p(i)p(o|i)$$

The noisy Channel Model

$$\hat{I} = \underset{i}{\operatorname{argmax}}\ p(i|o) = \underset{i}{\operatorname{argmax}} \frac{p(i)p(o|i)}{p(o)} = \underset{i}{\operatorname{argmax}}\ p(i)p(o|i)$$

- $p(i)$ is the language model and $p(o|i)$ is the channel probability
- Ex: Machine translation, optical character recognition, speech recognition