In this lecture we introduce some of the mathematical tools that are at the base of the technical approaches we will consider in this class. The issues we will recall (1) Probability Theory (2) Large Deviation Bounds (3) Information Theory (4) Statistics.

1 Introduction

Consider the following examples:

1. I saw a girl *it* the park. (I saw a girl *in* the park.)
2. The *from* needs to be completed. (The form needs to be completed).
3. I *maybe* there tomorrow. (I may be there tomorrow.)
4. Nissan's car *plants* grew rapidly in the last few years.

Questions:

- How do we identify that there is a problem?
- How do we fix it?

In order to address these problems we would like to find a systematic way to study the problem.

- How do we model the problem?
- How do we use this model to solve the problem?

The problems described are new to you. You have never seen the sentences (1)-(4). How can you be so sure that you can recognize a problem and correct it? It is clear that there exist some regularities in language. How can we exploit these to detect and correct errors? There are general mathematical approaches that can be used to translate this intuition into a model that we can study.

Logical Organization of What's Coming Next:

It seems clear that in order to make decisions of this sort, one has to rely on previously observed data.

> *Statistics* is the art of learning from data. It is concerned with the collection of data, its description and its analysis, which often leads to the drawing of conclusions.

Sheldon Ross; Introduction to Statistics and Probability

In order to relate properties of the sample data to properties of the entire population, it is necessary to have some understanding of probability:

> By *chance*, we mean something like a guess. Why do we make guesses? We make guesses when we wish to make a judgment but have incomplete information or uncertain knowledge. We want to make a guess as to what things are, or what things are likely to happen. Often we wish to make a guess because we have to make a decision. Sometimes we make guesses because we wish, with our limited knowledge, to say as much as we can about some situation. Really, any generalization is in the nature of a guess. Any physical theory is a kind of a guesswork. There are good guesses and there are bad guesses. The theory of probability allows us to speak quantitatively about some situation which may be highly variable, but which does have some consistent average behavior.

Richard Feynman, 1963

In many cases, we want to be able to talk about generalization from the observed data to the whole population, but have very little knowledge as to **what is the model that governs the generation of the data.** This is some times called the study of non-parametric models. The main interest here is to be able to say something about the behavior of the data just by estimating some statistics of it. This mathematical theory, which is at the heart of Learning Theory, is based on several inequalities, and we will title this section **Large Deviation Bounds.**.

Another important branch of statistics, **statistical hypothesis testing** is relevant to us mostly when we run experiments. We would like to know what is the plausibility of a certain hypothesis – e.g. that one prediction we make is better than another.

Finally, I will mention a few of the key notions from **Information Theory** that are relevant in the study of NLP.

2 Some Notes on Statistics

In many cases we would like to collect data and learn some *interesting things* about it. This is what we did with the Tom Sayer book, and later with the NYT data. We just defined a few properties of the data, and did some counting to see how many times they occur. We *summarized the data* and described it via the measurements we defined. This branch of statistics is call *descriptive statistics.*.

In many cases, we also want to **draw conclusions from the data**, this is sometimes called *inferential statistics.* In order to draw logical conclusions from the the observed data, we must take into account the possibility of chance.

E.g., a coin toss.

To be able to draw conclusions from the data, we could make some assumption about the chances of different outcomes of experiments with the data. These assumptions are often referred to as a *probability model* for the data. That is, inferential statistics has at its base an understanding of the **theory of probability.**

Descriptive Statistics

is a field that most of you have seen in school, as early as elementary school. Basic concepts there include graphing data in different ways, estimating statistics that describe the **central tendencies of the data** (sample mean, median and mode) and estimating statistics that describe the **spread or variability of the data** (variance and standard deviation).

It is interesting to realize that descriptive statistics is not only about measurements; there are in fact some very fundamental relations between these measurements.

Recall that the sample mean of data set $x_1, x_2, \ldots x_n$, is defined as

$$\overline{x} = \sum_{i=1}^{n} x_i/n$$

and the sample standard deviation as

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2/(n-1)}.$$

Chebyshev's Inequality states that for any value of $k \geq 1$, more than $100(1 - 1/k^2)$ percent of the data is within ks from the mean. E.g., for $k = 2$, more than 75% of the data lies within $2s$ from the sample mean.

Theorem 2.1 (Chebyshev's Inequality for Sample Data) *Let $\overline{x}$ and s be the sample mean and sample standard deviation of the data set $x_1, x_2, \ldots x_n$, where $s > 0$. Let*

$$S_k = \{i, 1 \leq i \leq n : |x_i - \overline{x}| < ks\}$$

and let $N(S_k)$ be the number of elements in the set S_k. Then, for any $k \geq 1$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}.$$

Proof:

$$\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \sum_{i \in S_k}(x_i - \overline{x})^2 + \sum_{i \notin S_k}(x_i - \overline{x})^2 \\
&\geq \sum_{i \notin S_k}(x_i - \overline{x})^2 \\
&\geq \sum_{i \notin S_k} k^2 s^2 \\
&= k^2 s^2 (n - N(S_k)),
\end{aligned}$$

where the first inequality follows from since all terms being summed are nonnegative, and the second from the definition of S_k. Dividing both sides by nk^2s^2 proves the left inequality. The right one follows immediately. ■

Because Chebyshev's inequality holds universally, it might be expected that for a given data set, the actual percentage of the data that is within ks from the mean is quite a bit larger.

Indeed, if the data is distributed normally, we know that 95% of the data lies with $\overline{x} \pm 2s$.

Example 2.1 *Think about different histograms you can generate over natural langauge data. E.g., we looked at the number of possible pos tags each word can take. The average is actually 1.2. What are other histograms of interest?*

Descriptive statistics can allow us to say quite a few things about the data we observe. For example, assume that our data set consists of pairs of values that have some relationships to each other. The ith data point is represented as the pair (x_i, y_i). A question of interest might be whether the x_is are correlated with the y_is. That is, whether large x values go together with large y values, etc.

To obtain a quantitative measure of this relationship it is possible to develop a statistic that attempt to measure the degree to which larger x values go with larger y values and smaller x values with smaller y values. One of these is the *sample correlation coefficient.*

Definition 2.1 (sample correlation coefficient) *Let s_x and s_y denote, respectively, the sample standard deviation of the x values and the y values. The sample correlation coefficient, denoted r, of the data pairs (x_i, y_i), $i = 1, \ldots n$, is defined by:*

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

When $r > 0$ we say that the sample data pairs are positively correlated, and when $r < 0$ we say they are negatively correlated.

Properties or r

1. $-1 \leq r \leq 1$
2. If for constants a, b, with $b > 0$,
   $$y_i = a + bx_i, \qquad i = 1, \ldots n$$
   then $r = 1$
3. If for constants a, b, with $b < 0$,
   $$y_i = a + bx_i, \qquad i = 1, \ldots n$$
   then $r = -1$
4. If r is the sample correlation coefficient for the data pairs $(x_i, y_i) i = 1, , \ldots n$ then it is also the sample correlation coefficient for the data pairs $(a + bx_i, c + dy_i) i = 1, , \ldots n$ provided that b, d are both positive or both negative.

This are just examples (important ones) that exhibit what can be said about data. We will see later a few other statistics of interest for sample data.

3 Elements of Probability

Various meanings or interpretations can be given to the concept of "the probability of a particular event". The key interpretations are:

- *the frequency interpretation* in which the probability of a given outcome of an experiment is considered a "property" of that outcome. This outcome can be operationally determined by repeating the experiment and the probability of the outcome will be observable as the proportion of the experiments that result in this outcome.
- *the subjective interpretation* in which the probability is not a property of the outcome but rather a statement about the belief of the person who is quoting the probability, concerning the chances that the outcome will occur.

Regardless of the interpretation, the mathematics of probability is the same. Here is the list (no details) of key concepts used in probability theory.

3.1 Probability Spaces and Events

The set S of all possible outcomes of an experiment is the *sample space* of the experiment.

Example 3.1 *In sentence # 2 above $S = \{$from, form $\}$ could be one set.*

When the sample space is finite, things are conceptually simple. Any subset E of the finite sample space is called an *event*. That is, an event is a set consisting of possible outcomes of the experiment.

In the general case of of infinite sample space things become more complicated since not every subset of the sample space is *measurable*.

For us, we can think of the event space as the power set of S, with the algebra of sets defined over $\mathcal{E} = 2^S$.

A *probability distribution* is any function

$$P : \mathcal{E} \to [0, 1]$$

such that

- $P(S) = 1$
- For all $n = 1, 2, \ldots$, and for any collection of mutually exclusive events $E_1, E_2, \ldots E_n$ (that is, $E_i \cap E_j = \Phi$, when $i \neq j$)
$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i).$$

Note that all other rules of probability actually follow from these axioms. E.g.,

Sum Rule: Given two events A, B, the probability of their union is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The tuple $(S, \mathcal{E}, P)$ defines the **probability space**. It is crucially important to define properly it when we attempt to apply probability theory in any context.

3.2 Basic Principles of Counting

You know it. If not, go to any discrete Math book.

3.3 Conditional Probability

Conditional probability is one of the most important concepts in probability theory.

First, often we are interested in computing probabilities when some partial information concerning the result of the experiment are available, or in computing them in light of additional information. In such cases, the desired probabilities are conditional ones.

Second, as a bonus, often the easiest way to compute the probability of an event is to condition on the occurrence or the nonoccurrence of another event.

Example 3.2 *Consider sentence # 1 above:*

- *what is the probability to observe the word* it*?*
  Under the assumption that the probability space is a collection of words from a corpus (bag of words), this probability can also be estimated.
- *what is the probability that the word* it *follows the word* girl*?*
  Given a corpus of text, we can generate several reasonable probability spaces. E.g., the space defined over the collection of all ordered word pairs $\{w_1, w_2\}$ not separated by a sentence delimiter. Given that, the event above is well defined and the probability can be estimated. (E.g., consider all pairs in which it *is the second word. Find the fraction of those pairs in which the first word is* girl*.*

Example 3.3 *Consider rolling a pair of fair dice. What is the probability that the sum is 8? What is the probability that the sum is 8 given that the first one is 3? Try computing the first probability with and without conditioning.*

Definition 3.1 (Conditional Probability) *The conditional probability of event E given that event F has occurred ($P(F) > 0$) is defined to be*

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{1}$$

Even if $P(F) = 00$ we have that:

The Product Rule:

$$P(E \cap F) = P(E|F)P(F) = P(F|E)P(E).$$

This rule can be generalized to multiple events:

The Chain Rule:

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2) \cdots P(E_n| \bigcap_{i=1}^{n-1} E_i).$$

For events E, F, we can write:

$$P(E) = P(E \cap F) + P(E \cap F^c),$$

where F^c is the complementary event to F.

In general we can get:

Total Probability: For mutually exclusive event $F_1, F_2, \ldots F_n$ with $\sum_{i=1,n} P(F_i) = 1$,

$$P(E) = \sum_i P(E|F_i)P(F_i).$$

The partitioning rule can be phrased also for conditional probabilities:

Total Conditional Probability For mutually exclusive event $F_1, F_2, \ldots F_n$ with $\sum_{i=1,n} P(F_i) = 1$,

$$P(E|G) = \sum P(E|F_i, G)P(F_i|G).$$

3.4 Bayesian Decision Theory

Let E, F be events. As a simple consequence of the product rule we get the ability to swap the order of dependence between events.

Bayes Theorem: (simple form)

$$P(F|E) = P(E|F)\frac{P(F)}{P(E)}. \tag{2}$$

For mutually exclusive events $F_1, F_2, \ldots F_n$ with $\sum_{i=1,n} P(F_i) = 1$,

Altogether we get the general form of **Bayes Theorem:**

$$P(F_i|E) = \frac{P(E|F_i)P(F_i)}{P(E)} = \frac{P(E|F_i)P(F_i)}{\sum_i P(E|F_i)P(F_i)}. \tag{3}$$

Bayes Rule, which is really a simple consequence of conditioning, is the basis for an important theory of decision making. The key observation is that Bayes rule can be used to determine, given a collection of events, which is the more likely event.

Let $B_1, B_2, \ldots B_k$ be a collection of events. Assume that we have observed the event A, and we would like to know which of the $B_i's$ is more likely given A. This is given by Bayes rule as:

$$i^* = argmax_i P(B_i|A) = argmax_i \frac{P(A|B_i)P(B_i)}{P(A)} = argmax_i P(A|B_i)P(B_i) \tag{4}$$

Notice that in order to choose the *most likely* B_i given A, you don't really need to know the probability of A.

3.5 Basics of Bayesian Learning

The decision rule 4 is all there is to probabilistic decision making. Everything else comes from the details of defining the probability space and the events A, B_i.

The way we model the problem, that is, the way we formalize the decision we are after depends on our definition of the probabilistic model.

- Words. What is the problem now? What is A? What are the B_i's?
- Pairs.
- The probability space can be defined over sequences of words and their POS tags. Given: A= a sequence of words; B_i - a possible pos tag

The key issues here would be (1) how to represent the probability distribution, and (2) how to represents the events. (Since we don't want to make the decision by computing it explicitly for all i.

- The probability space can be defined over sentences and their parse trees.
- The probability space can be defined over English sentences and their Chinese translation.
- The probability space can be over Boolean functions. For this purpose, represent a Boolean function as a collection of positive examples (all the rest are negative).

  A would be a collection of positive and negative examples that we observed and the B_is - a collection of 7 candidate functions. We want to know which of these functions is more likely to have generated the sample we observed.
-

Essentially all the probabilistic approaches we will address are instantiations of this rule. The key questions, and consequently the differences between models and computational approaches boil down to (1) what are the probability spaces and (2) how are they represented.

Here is how the Bayesian decision model (sometimes called Bayesian learning) is typically phrased:

We assume that we observe data S, and have a hypothesis space H, from which we want to choose the *Bayes Optimal* one, h.

Of course, in order to talk about the *most probable* hypothesis, we will have to assume a probability distribution over the class of hypotheses H. In addition, we will need to know how this distribution depends on the data observed.

1. $p(h)$ - the *prior probability* of an hypothesis h. This is the initial probability, before we observe any data. This reflects the background knowledge we have; if we have none, we can assign uniform distribution over H.
2. $P(S)$ - probability that this sample of the data is observed. Again, this in an initial probability, without any knowledge about which hypothesis holds.
3. $P(S|h)$ is the probability of observing the sample S, given that the hypothesis h holds.
4. $P(h|S)$ is the *posterior probability* of h. It is the probability of h, given that we have seen the sample S.

The term $P(S)$ is sometimes called the *generative model*. It reflects our understanding of how data is generated in the world. In many cases it is unlikely that we have a good knowledge of that but we will see that in order to make decisions, in many cases we do not actually to know this model.

Typically, we are interested in computing $P(h|S)$ or, at least, providing a ranking of $P(h|S)$ for different hypotheses $h \in H$.

The key technical issue underlying Bayesian decision making is Bayes Theorem, phrased here in terms of h and S as:

$$P(h|S) = P(S|h)\frac{P(h)}{P(S)}.$$

Some intuitions: $P(h|S)$ increases with $P(h)$ and with $P(S|h)$. It decreases with $P(S)$, since the more probable it is that D is observed independently of h, the less evidence it provides in support of h.

3.5.1 Learning scenario:

The learning step here is inducing the "best" h from the data we observe, given what we know on the probability distribution. The learner considers a set of candidate hypotheses H, and is trying to find the most probably one $h \in H$, given the observed data.

1. Such maximally probable hypothesis (it is not necessarily unique) is called *maximum a posteriori (MAP* hypothesis.

   $$h_{MAP} = argmax_{h \in H} P(h|S) = argmax_{h \in H} P(S|h)\frac{P(h)}{P(S)} = argmax_{h \in H} P(S|h)P(h).$$

   Notice that we have dropped the dependence on $P(S)$ because it is a constant and does not depend on h.

2. In some cases, we assume that the hypotheses are equally probable a priori.

   $$P(h_i) = P(h_j), \quad for \;\; all \;\; h_i, h_j \in H.$$

   In this case, we can further simplify the above equation:

   $$h_{ML} = argmax_{h \in H} P(S|h).$$

   This is referred to as the *Maximum Likelihood* hypothesis.

Reiterating: we will use these concepts at many levels. From estimating parameters to smoothing to learning complex probabilistic models to decision making.

Example 3.4 *A given coin is either fair or has a* 60% *bias in favor of Head. Decide what is the bias of the coin.*

Two hypotheses: $h_1 : P(H) = 0.5; h_2 : P(H) = 0.6$

Prior: $P(h) : P(h_1) = 0.75P(h_2) = 0.25$

Now observe the data: Toss the coin once; say the outcome is H.
$P(D|h) : P(D|h_1) = 0.5; P(D|h_2) = 0$
$P(D) : \quad P(D) = P(D|h_1)P(h_1) + P(D|h_2)P(h_2) = 0.5 \cdot 0.75 + 0.6 \cdot 0.25 = 0.525$

$P(h|D)$:
$P(h_1|D) = P(D|h_1)P(h_1)/P(D) = 0.5 \cdot 0.75/0.525 = 0.714$
$P(h_2|D) = P(D|h_2)P(h_2)/P(D) = 0.6 \cdot 0.25/0.525 = 0.286$

Try it assuming 100 *tosses;* 60 *heads; try it in the most likelihood case.*

Example 3.5 *Given a corpus; what is the probability of the word "*it*" in the corpus?*

Bag-of-word model: let m be the number of words in the corpus. Let k be the number of "it"s in the corpus. The probability is p/k

Modeling: Assume a bag-of-words model, with a binomial distribution model. The problem becomes that of:

> *you toss a $(p, 1-p)$ coin m times and get k Heads, $m-k$ Tails. What is p?*

The hypothesis space is $p \in [0,1]$. A continuous space...

If p is the probability of Head, the probability of the data observed is:

$$P(D|p) = p^k(1-p)^{m-k}.$$

We want to find the value of p that maximizes this quantity. Easier to work in log space: The Log Likelihood: To maximize, set the derivative w.r.t. p equal to 0:

$$p^* = argmax_p P(D|p) = argmax_p log P(D|p) = argmax_p[klogp + (m-k)log(1-p)].$$

$$\frac{dlogP(D|p)}{dp} = \frac{dkp + (m-k)(1-p)}{dp} = \frac{k}{p} - \frac{m-k}{1-p}.$$

The derivative with respect to p *is* 0 *when* $p = k/p$.

Solving this for p, gives: p=k/m

3.6 Independence

Two event E, F are *independent* of each other if

$$P(E \cap F) = P(E)P(F).$$

If $P(F) \neq 0$, this is equivalent to saying that:

$$P(E) = P(E|F).$$

We can also say that E and F are *conditionally independent* given G, when:

$$P(E \cap F|G) = P(E|G)P(F|G).$$

Example 3.6 *Develop an example of three events A, B, C such that A, B are in dependent but are dependent given C and vice versa.*

3.7 Random Variables

A function

$$X : S \rightarrow \Re$$

that is used as a convenient way to defined events over S.

Assume that the range R of X is discrete. For $r \in R$ we define the event

$$X^{-1}(r) = \{s \in S | X(s) = r\}$$

and can talk about the probability of this event.

The *probability mass function* p of the random variable X is defined as

$$\forall r \in R : p(r) \equiv p(X = r) = P[X^{-1}(r)].$$

And, of course,

$$\sum_r p(r) = P(S) = 1.$$

3.7.1 Expectation

3.7.2 Variance

3.7.3 Covariance

3.8 Probability Distributions

3.8.1 Joint and conditional distributions

We can define a collection of random variables over a probability space. The distribution space is then called the *join probability space.*

3.8.2 Representation of probability distributions

Probability distributions can be represented in many different way. The representation is crucially important since it determined what we can say about it and what can be (efficiently) computed about it.

- A distribution can be represented explicitly, by specifying the probability of each element in the space.
- It can be specified parametrically: Binomial distribution; Normal distribution; geometric distribution, etc.
- It can be represented graphically.
- It can be specified by specifying a *process* that generates elements in the space. Depending on the process, it may be easy or hard to actually compute the probability of an event in the space.

  Example 3.7 (Model of Language) *Consider the following model. We have five characters, A,B,C,D,E. At any point in time we can generate: A, with probability 0.4, B, with probability 0.1, C, with probability 0.2, D, with probability 0.1, E, with probability 0.2. E represents end of sentence.*

  A graphical representation: a flower model. A sentence in this language could be:

  $$AAACDCDAADE$$

  An unlikely sentence in this language would be:

  $$DDDBBBDDDBBBDDDBBB.$$

  The is an issue of normalization that we don't need to deal with if all we care about is comparing likelihoods.

It is clear that given the model we can compute the probability of each string and make a judgment as to which sentence is more likely.

Example 3.8 (Another model of Language) *Models can be more complicated; e.g., a probabilistic finite state model. Start state is A, with probability 0.4; B, with probability 0.4, C, with probability 0.2. From A you can go to A, B or C, with probabilities 0.5, 0.2, 0.3. From B you can go to A, B or C, with probabilities 0.5, 0.2, 0.3. From C you can go to A, B or C, with probabilities 0.5, 0.2, 0.3. Etc.*

- In this way we can easily define models that are intractable.
- When you think about language, you need, of course to think about what do $A, B, ...$ represent. words, pos tags, structures of some sort, ideas?...

- Sometimes we define a probability distribution by defining *properties* of it, or *conditions* over events, or both. E.g., define the probability of some events.
  - This may or may not define the probability distribution completely.
  - This may of may not be sufficient to compute to probability of some other events, or other properties of the distribution.
- Sometimes we are only given a *family* of probability distribution (define parametrically, as a process or by specifying some conditions as above. A key problem in this case might be to choose one specific distribution in this family. This can be done by observing data, specifying some property of the distribution that we care about the might distinguish a unique element in the family, or both.

3.9 Probabilistic Inequalities and the law of Large Numbers

4 Large Deviation Bounds

As we will see, large deviation bounds are at the base of the Direct Learning approach. When one wants to directly learn how to make predictions the approach is basically as follows:

- Look at many examples
- Discover some regularities in the language
- Use these regularities to construct a prediction policy

The key question the direct learning approach needs to address, given the distribution free assumption, is that of generalization. Why would the prediction policy induced from the empirical data be valid on new, previously unobserved data. The key assumption in this theory is that throughout the process, we observe examples sampled by some fixed, but unknown, distribution. That is, the distribution the governs the generation of examples during the learning process is the one that governs it during the evaluation process. Two key basic results from probability theory are useful in developing the justification.

4.1 Chernoff Bounds

Consider a binary random variable X (e.g., the result of a coin toss) which has probability p of being *head (1)*, and probability $1-p$ of being *tail (0)*. Two important questions are:

- How do we estimate p?
- How confident are we in this estimate?

Consider m samples, $\{x_1, x_2, \ldots\}$, drawn independently from X. A natural estimate of the underlying probability p is the relative frequency of $x_i = 1$. That is

$$\hat{p} = \sum_i x_i/m.$$

Notice that this is also the maximum likelihood estimate according to the binomial distribution.

Now, by the law of large numbers $\hat{p}$ will converge to p as the sample size m goes to infinity. The second question now becomes: how quickly does this estimate converge to p?

The Chernoff bounds go a step further than the law of large numbers, which is an asymptotic result (a result concerning what happens as the sample size goes to infinity).

Theorem 4.1 (Chernoff Bounds) *For all $p \in [0, 1], \epsilon > 0$,*

$$P[|p - \hat{p}|] \leq 2e^{-2m\epsilon^2} \tag{5}$$

where the probability P is taken over the distribution of training samples of size m generated with underlying parameter p.

The bound states that for all values of p, and for all values of ϵ, if we repeatedly draw training samples of size m of a binary variable with underlying probability p, the relative frequency estimate $\hat{p}$ converges to the true value p at a rate that is exponential in m.

The bounds can be phrases in a once sided way (the coefficient 2 disappears) and in more subtle way that depend on the value of p.

Some Numbers:

Take $m = 1000$, $\epsilon = 0.05$, then $e^{-2m\epsilon^2} = e^{-5} \approx 1/148$. That is: if we repeatedly take samples of size 1000 and compute the relative frequency, then in 146 out of 148 samples, our estimate will be within 0.05 from the true probability. We need to be quite unlucky for that to happen.

4.2 Union Bounds

A second useful result, a very simple one in fact, is that of the Union Bounds.

Theorem 4.2 (Union Bounds) *For any n events $\{A_1, A_2, \dots A_n\}$ and for any distribution P whose sample space includes all A_i,*

$$P[A_1 \cup A_2 \cup \dots A_n] \leq \sum_i P[A_i] \tag{6}$$

The Union Bound follows directly from the axioms of probability theory. For example, for $n = 2$

$$P[A_1 \cup A_2] = P[A_1] + A[A_2] - P[A_1 \cap A_2] \leq P[A_1] + A[A_2].$$

The general result follows by induction on n.

There two results are applied to derive the basic generalization bounds in learning theory.

5 Information Theory

5.1 Noisy Channel Model

One of the important computational abstractions made in NLP is the of the Noisy Channel Model. Technically, this is no more than an application of Bayes rule, but it is sometimes a useful way to think about and model problems.

The abstraction is rooted in Information Theory and is due originally to Shannon. He modeled the process of communicating across a channel. The goal is to optimize, in terms of throughput and accuracy, the transmission of messages across a noisy channel.

The goal is to encode the transmitted message in such a way that the receiver will be able to decode the message accurately. The key tradeoff is between *compression* and *redundancy*.

In the most trivial instantiation of it – if you send each character 17 times, and the receiver decodes by majority over each group of 17 characters, than (under the assumption the the channel can only change one character to another) is it enough for 9/17 to reach the receiver intact. Of course, better encoding schemes might be possible, e.g., taking into account some properties of the message of the langauge in which they are generated.

The assumption is that there is some probabilistic process that governs the generation of language (messages) and that there is a corruption process, modeled via the notion of a channel that affects the generated language structure before it reaches the listener, say.

Let I be the input space, a space of some structures in the language. ("structure" may mean different things in different contexts).

Let P be a probability distribution over I; that is, this is our *language model.* For each $i \in I$, $P(i)$ is the probability of observing this structure.

The noisy channel has a probability distribution $P(o|i)$ of outputting the structure o when receiving the input i.

```
  Input ---- i ---> [ Noisy Channel ] ---- o ----> [ decoder ]----> i'
```

Notice that the real information theoretic model has another stage, an encoder, which is the most crucial stage. It allows one to study how to encode the input as a function of properties of the channel.

The goal is to decode the most likely input given the observed output. Formally, that means

$$I = argmax_{i \in I} P(i|o) = argmax_{i \in I} P(o|i)P(i)/P(o) = argmax_{i \in I} P(o|i)P(i).$$

So, in addition to the language model, we also need to know the distribution of the noisy channel.

Example 5.1 Translation*: According to this model, English is nothing but corrupted French.*

True text: French (i), English (O). Given a structure (sentence) in English, find the most likely French sentence. Unfortunately, the French sentence went through a pretty noisy channel, and came out as English. In order to translate it, we need to know something about the noisy channel.

Examples:

Application	Input	Output	P(i)	P(o|i)
Speech Recognition.	text	speech	language model	acoustic model P(signal|text)
Optical Character Recognition (OCR)	actual text	observed text (with mistakes)	language model	mistake model
POS tagging	Seq. of POS tags	Seq. of English Words	Prob. of POS seq.	P(w|t) (Seq? Single?)
Lexical Disambiguation	English Sent.	Incorrect Sent.	lang model.	P(o|peace)= P(peach|o)P(o)/P(peace

The key achievement of Shannon was the ability to quantify the theoretically best data compression possible and the best transmission rate possible and the relation between them.

Mathematically, these correspond to the *Entropy* of the transmitted distribution and the *Capacity* of the channel.

Basically, he showed that each channel is associated with a *capacity* and if you transmit your information at a rate that is lower then it, the decoding error can be as small as possible.

The channel capacity is given using the notion of *the mutual information* between the random variables I and O.

That is the reason that Entropy and Mutual Information became important notions in NLP.

5.2 Entropy

For a give random variable X, how much information is conveyed in the message that $X = x$?

In order to quantify this statement we can first agree that the amount of information in the message that $X = x$ should depend on how likely it was the X would equal x.

In addition, it seems reasonable to assume that the more unlikely it was that X would equal x, the more informative would be the message.

For instance, if X represents the sum of two fair dice, then there seems to be more information in the message that $X = 12$ than there would be in the message that $X = 7$ since the former events happens with probability 1/36 and the latter 1/6.

Let's denote by $I(p)$ the amount of information contained in the message that an event whose probability is p has occurred. Clearly $I(p)$ should be non negative, decreasing function of p.

To determine it's form, let X and Y be independent random variables, and suppose that

$$P\{X = x\} = p \qquad P\{Y = y\} = q.$$

How much information is contained in the message that X equals x and Y equals y?

Note that since knowledge of the fact that X equals x does not affect the probability that Y will equal y (since X, Y are independent), it seems reasonable that the additional amount of information contained in the statement the $Y = y$ should equal $I(q)$. Therefore, the amount of information in the message that $X = x$ and $Y = y$ is $I(p) + I(q)$.

On the other hand:

$$P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\} = pq$$

which implies that the the amount of information in the message that $X = x$ and $Y = y$ is $I(pq)$.

Therefore, the function I should satisfy the identity

$$I(pq) = I(p) + I(q)$$

If we define

$$I(2^{-p}) = G(p)$$

we get from the above that

$$G(p + q) = I(2^{-(p+q)}) = I(2^{-p}2^{-q}) = I(2^{-p}) + I(2^{-q}) = G(p) + G(q)$$

However, it can be shown that the only monotone functions G that satisfy the this functional relationship are those of the form

$$G(p) = cp$$

for some constant c. Therefore, we must have that

$$I(2^{-p}) = cp,$$

or, letting $z = 2^{-p}$,

$$I(z) = -clog_2(q)$$

for some positive constant c. It is traditional to assume that $c = 1$ and say that the information is measured in units of *bits*.

Consider now a random variable X, which must take on one of the values $x_1, \dots x_n$ with respective probabilities $p_1, \dots p_n$.

As $-log(p_i)$ represents the information conveyed by the message that X is equal to x_i, it follows that the expected amount of information that will be conveyed when the value of X is transmitted is given by

$$H(X) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

This quantity is know in information theory as the **entropy** of the random variable X.

Definition 5.1 (Entropy) *Let p be a probability distribution over a discrete domain X. the entropy of p is*

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

Notice that we can think of the entropy as

$$H(p) = -E_p \log p(x).$$

Since $0 \leq p(x) \leq 1$, $1/p(x) > 1$, $\log 1/p(x) > 0$ and therefore $0 < H(p) < \infty$.

Intuitively, the entropy of the random variable measures the uncertainty of the random variable (or: how much we know about its value when we know the distribution p). The value of $H(p)$ is thus maximal and equal to $\log |X|$ when p is the uniform distribution.

We will see the notion of Entropy in three different contexts.

- As a measure for information content in a distributional representation.
- As a measure for information; how much information about X do we get from knowing about Y?
- As a way to characterize and *choose* probability distributional ultimately, classifiers.

Example 5.2 *Consider a simple language L over a vocabulary of size* 8*. Assume that the distribution over L is uniform:* $p(l) = 1/8, \forall l \in L$.

Then,

$$H(L) = -\sum_{l=1}^{8} p(l) log p(l) = -log\frac{1}{8} = 3$$

Indeed, if you want to transmit a character in this language, the most efficient way is to encode each of the 8 *characters in* 3 *bits. There isn't a more clever way to transmit these messages. An optimal code sends a message of probability p in* $-log p$ *bits.*

On the other hand, if the distribution over L is:

$$\{1/2, 1/81/8, 1/8, 1/32, 1/32, 1/32, 1/32\}$$

Then:

$$H(L) = -\sum_{l=1}^{8} p(l) log p(l) = [1/2 \cdot 1 + 3(1/8 \cdot 3) + 4(1/32 \cdot 5) = 1/2 + 9/8 + 20/32 = 2.25.$$

Definition 5.2 (Joint and Conditional Entropy) *Let p be a joint probability distribution over a pair of discrete random variables* X, Y*. The average amount of information needed to specify both their values is the joint entropy:*

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

The conditional entropy of Y *given* X*, for* $p(x, y)$ *expresses the average additional information one needs to supply about* Y *given that* X *is known:*

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) = \sum_{x \in X} p(x)[-\sum_{y \in Y} p(y|x) log p(y|x)] = -\sum_{y \in Y} \sum_{x \in X} p(x, y) \log p(y|x).$$

Definition 5.3 (Chain Rule) *The chain rule for entropy is given by:*

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

More generally:

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

5.2.1 A simple example

Consider the problem of selecting a probability distribution over two variables, $Z = X \times C$, where $X = \{x, y\}$ and $C = \{0, 1\}$.

There could be many probability distributions over this space of 4 elements. Assume, though, that we know something about this distribution. E.g., by observing a sample S we determine that:

$$p\{c = 0\} \equiv p(x, 0) + p(y, 0) = 0.6.$$

Now, there are still many probability distributions with this marginal.

We need to fill this table:

X \C	0	1	
x			
y			
total	0.6		1.0

One way to do it is:

X \C	0	1	
x	0.3	0.2	
y	0.3	0.2	
total	0.6		1.0

In this case, the entropy is

$$\begin{aligned} H(p) &= -\sum_{x \in X} p(x) \log p(x) && (7) \\ &= 0.3 \log 0.3 + 0.3 \log 0.3 + 0.2 \log 0.2 + 0.2 \log 0.2 && (8) \\ &= 1.366/\ln 2 = 1.97 && (9) \end{aligned}$$

Another way is:

X \C	0	1

```
  x    | 0.5 | 0.3 |
  y    | 0.1 | 0.1 |
------------------------
total    0.6 |     | 1.0
```

Here we get:

$$
\begin{aligned}
H(p) &= -\sum_{x \in X} p(x) \log p(x) && (10)\\
&= 0.5 \log 0.5 + 0.3 \log 0.3 + 0.1 \log 0.1 + 0.1 \log 0.1 && (11)\\
&= 1.16/\ln 2 = 1.68 && (12)
\end{aligned}
$$

Notice that the maximum entropy you can get for a distribution over a domain of size 4 is $-log 1/4 = log 4 = 2$.

One plausible way to choose a distribution that satisfies given constrains is to choose the one that has the maximum entropy. Intuitively, this one minimizes the number of *additional constraints* one assumes on the distribution.

In general, though, there is no *closed form solution* to finding the distribution that satisfies the constrains and has the maximum entropy and there will be a need to resort to an iterative algorithm.

The problem above can also be viewed as a way to solve a decision problem, if we are looking for the conditional probability $P(C|X)$, rather than the joint. Constraints might be, again, marginals, like: $P(X = x|C = 0) = 0.2$

5.3 Mutual Information

We showed:

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

which implies:

$$H(X) - H(X|Y) = H(Y) - H(Y|X).$$

This difference is called the mutual information between X and Y and is denoted $I(X : Y)$. It is the reduction in the uncertainty of one RV due to knowing about the other.

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) = \\
&= H(X) + H(Y) - H(X,Y) = \\
&= \sum_x p(x)log\frac{1}{p(x)} + \sum_x p(y)log\frac{1}{p(y)} - \sum_{x,y} p(x,y)log\frac{1}{p(x,y)} = \\
&= \sum_{x,y} p(x,y)log\frac{1}{p(x)} + \sum_{x,y} p(x,y)log\frac{1}{p(y)} - \sum_{x,y} p(x,y)log\frac{1}{p(x,y)} = \\
&= +\sum_{x,y} p(x,y)log\frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

This allows you to see the relation to independence. I(X;Y) can be best thought as a measure of *independence* between X and Y.

- If X and Y are independent than $I(X;Y) = 0$
- If X and Y are dependent than $I(X;Y)$ grows with the degree of dependence but also with the entropy.
- $I(X;X) = H(X) - H(X|X) = H(X)$

It is also possible to talk about *conditional mutual information*:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z).$$

5.3.1 Preliminary comments on Mutual Information

We have defined mutual information between random variables. In many applications this is being abused and people use it to measure *point-wise mutual information.* Let x, y be two points in a probability space. The point-wise mutual information between these two points can be defined to be:

$$I(x, y) = log\frac{p(x, y)}{p(x)p(y)}.$$

This can be thought of as some statistical measure relating the events x and y.

For example, consider looking at pair of consecutive words w, w'. We can define:

- $p(w)$ - the frequency of the word w in a corpus,
- $p(w')$ - the frequency of the word w' in a corpus
- $p(w, w')$ - frequency of the pair w, w' in the corpus.

and then, consider $I(w, w')$ as *the amount of information* about the occurrence of w in location i given that we know that w' occurs at $i + 1$.

It turns out the this is not a good measure, and we will see examples for it. Other statistics, like χ^2 represent better what we want here.

As an example, consider two extreme cases. Assume the w, w' occurs only together. Then:

$$I(w, w') = log\frac{p(w, w')}{p(w)p(w')} = log\frac{p(w)}{p(w)p(w')} = log\frac{1}{p(w')}$$

So, the mutual information *increases* among perfectly dependent bigrams, when they become rarer.

On the other hand, if they are completely independent,

$$I(w, w') = log\frac{p(w, w')}{p(w)p(w')} = log 1 = 0.$$

So, it can be thought of as a good measure for independence, but not a good measure for dependence, since it depends on the score of individual words.

6 Relative Entropy and the Kullback-Liebler Distance

Definition 6.1 (Relative Entropy; Kullback-Liebler Distance) *Let p, q be two probability distributions over a discrete domain X. The relative entropy between p and q is*

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

Notice that $D(p||q)$ is not symmetric. It could be viewed as

$$D(p||q) = E_p \log \frac{p(x)}{q(x)},$$

the expectation according to p of $\log p/q$. It is therefore unbounded and not defined if p gives positive support to instances q does not. It is 0 when $p = 0$.

Lemma 6.1 (KL-Divergence) *For any probability distributions p, q, $D(p||q) \geq 0$ and equality holds if and only if $p = q$.*

The KL divergence is not a *metric*, being non symmetric and does not satisfy the triangle inequality (but it has some other nice properties we will see later). We can still think about it as a *distance between distributions.*

The mutual information can be thought of as the KL divergence between the joint probability distribution on X, Y and the "product" distribution defined on X, Y.

$$I(X;Y) = D(p(x,y)||p(x)p(y)).$$

Phrased otherwise, it is the measure of how far the true joint is from independence.

7 Cross Entropy

Going back to the example of the simple language we started with, we can now think back on the notion of the *entropy* of the langauge.

Assume that you know the distribution; then you can think of the entropy as a measure of how surprised you are when you observe the next character in the language.

Let p the true distribution of a random variable X, and q some model that we estimate for X.

Definition 7.1 (Cross Entropy) *The Cross Entropy between a random variable X distributed according to p and another probability distribution q is given by:*

$$H(X,q) = H(X) + D(p||q) = -\sum_x p(x)\log p(x) + \sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_x p(x) \log q(x).$$

Notice that this is just the expectation of $1/q(x)$ with respect to $p(x)$:

$$H(X,q) = E_p \frac{1}{\log q(x)}.$$

It seems like we need to know p in order to estimate this. However, under some assumptions, this can be simplified. If we assume that p (our language) is *stationary* (probabilities assigned to sequences are invariant with respect to shifts) and *ergodic* (process the is not sensitive to initial conditions) then Shannon-McMillan-Breiman theorem shows that:

$$H(X,q) \approx lim_{n\rightarrow\infty} - \frac{1}{n} \log q(w_1 w_2 \dots w_n).$$

With this approximation, we can now try to imagine models q, and try to compute the entropy of a language model. The assumption is then that the lower the entropy is, the better the model is.

Think about how would you do it in practice.

Model	Cross Entropy (bits)
0-th order	4.76 (= log 27; uniform over 27 characters)
1-st order	4.03
2-nd order	2.8
Shannon Experiment	1.3

When people estimate a langauge model they sometime use the cross entropy, or the *perplexity*, defined as

$$2^H$$

as a quality measure.

Reducing the perplexity thus means improving the language model. But, notice that that does not always mean better predictions.