NOTES FOR CS187

STUART M. SHIEBER

Contents

1. Overview

COMPUTATIONAL LINGUISTICS
NATURAL LANGUAGES

This course is about COMPUTATIONAL LINGUISTICS, the study of NATURAL LANGUAGES (the languages spoken by humans, as opposed to artificial languages like formal languages, logics, or programming languages) using tools and techniques of computer science with application to the automated processing of language. Computational linguistics has both a *scientific* side, which emphasizes the goal of gaining insight into how languages works by formal modeling using computational tools and formalisms, and an *engineering* side, which emphasizes the building or improving of useful computational artifacts that manipulate language. The latter enterprise is often referred to as NATURAL-LANGUAGE PROCESSING (NLP). Typical examples of such useful artifacts are systems for speech recognition, machine translation, natural-language interfaces, text retrieval, question answering, efficient text entry, and optical character recognition.

NATURAL-LANGUAGE PROCESSING

This year, the course will be organized around increasingly complex outputs of NLP tasks. Think of an NLP system as taking some natural-language input, some text or speech, and generating some output that usefully relates to that input.[1] We can taxonomize the systems by the complexity of the output. At the simplest level, the output might be just a simple classification of the input, drawn from a finite set of possible classes. At the most complex, it might be a formal representation of the meaning of the input.

Here are the categories of output we will consider in this class, in complexity (and chronological) order:

Classes — the classification problem: Texts can be classified into a fixed set of classes. For instance, emails might be classified as spam or non-spam, and questions may be classified as to the expected answer type (entity, location, date, amount, etc.).

[1] Some NLP systems take non-natural-language inputs and generate natural-language output, for instance, natural-language generation sysems. Although such cases don't match this pedagogical taxonomy, the techniques we discuss can apply to those systems too.

Integers — the counting problem: We may want to know how many of something there are, extrapolated from a sample of text. How many English words are there? Does this author have a larger vocabulary than that one? Is one presidential candidate more egocentric than another?

Sequences — classification in sequential context: Language is inherently sequential. We will often want to map natural-language sequences to corresponding sequential outputs. For instance, we may want to know the words spoken in a speech signal, or the parts of speech for each word in a sentence.

Structures — the parsing problem: The most fundamental truth of natural language, which has been known since the time of the Sanskrit grammarian Pāṇini almost three millennia ago, is that sentences of natural language have hierarchical structure. Recovering that structure from the observable word sequence is the parsing problem, a precursor to any further processing of sentences that rely on their meanings.

Meanings — the interpretation problem: The most sophisticated applications of natural-language processing, and even some of the simplest applications when at human levels of performance, require reconstruction of the meanings of the natural-language text being manipulated.

Along the way, we will introduce exemplar *problems* of each of these types, deploy *techniques* of general utility for these kinds of problems, and introduce certain *notions* fundamental to understanding and using these techniques.

The early stages of the course tend to make use of simple statistical models trained on samples of raw data. They fall within the EMPIRICIST APPROACH to NLP. As we move to more and more complex outputs, the techniques used require models of such complexity that statistical inference may not yet be up to the task. Human effort is needed to build these models, which are less statistical in nature. Such techniques fall within the RATIONALIST APPROACH to NLP.

EMPIRICIST APPROACH

RATIONALIST APPROACH

Different programming languages are better attuned to implementation of the empiricist and rationalist models. We will use the Python multi-paradigm programming language for the former and the Prolog logic-programming language for the latter.

2. Basic probability

I review some basic probability that will come up repeatedly in the course.

A random variable is an abstract object that we think of as being able to take on different values with different probabilities. The set of possible values for a random variable X is its domain, notated $Dom(X)$. By convention, we use capitalized symbols $X, Y, \ldots$ for random variables and sans serif symbols (for instance, $\mathsf{heads}, \mathsf{tails}$) for their values. Lower case symbols $x, y, \ldots$ (often subscripted) are used for metavariables that range over values and $A, B, \ldots$ as metavariables ranging over an event of one or more random variables taking on particular values.

random variable

domain

We might define a random variable *Flip* to take on values $Dom(Flip) = \{\mathsf{heads}, \mathsf{tails}\}$ as per a fair coin flip. Then the probability that $Flip = \mathsf{heads}$ is one half, which we notate as

$$\Pr(Flip = \mathsf{heads}) = 1/2 \quad .$$

When it is clear from context what random variable is being used, typically because the value it takes on makes that clear, we may drop the explicit mention of the random variable:

$$\Pr(\mathsf{heads}) = 1/2$$

The size of a random variable's domain, the number of values that it can take on, will be notated $|X|$. For instance, for the coin flip random variable, $|Flip| = 2$.

The total probability of all the values for the random variable must be one:

$$\sum_{x \in Dom(X)} \Pr(x) = 1 \tag{1}$$

Random variables that have a binary domain like this are given a special name, Bernoulli random variables. Bernoulli variables are completely characterized by a single probability $\Pr(X = \mathsf{true})$, since

Bernoulli random variables

the summation constraint (1) guarantees that

$$\Pr(X = \mathsf{false}) = 1 - \Pr(X = \mathsf{true}) \quad .$$

MULTINOMIAL RANDOM VARIABLES

Random variables with a finite domain are referred to as MULTINOMIAL RANDOM VARIABLES. Thus Bernoulli random variables are a special case, the simplest, of multinomial random variables.

JOINT PROBABILITY

The JOINT PROBABILITY of multiple random variables taking on particular values is indicated by a similar notation. If X is a random coin flip and Y a random die roll, then $\Pr(X = \mathsf{heads}, Y = 3)$ is the joint probability of both flipping heads and rolling a 3. Again, we may abbreviate this $\Pr(\mathsf{heads}, 3)$ in a context in which the random variables are clear.

CONDITIONAL PROBABILITY

We may be interested in the probability of an event A conditioned by (that is, under the assumption of occurrence of) an event B, for example, what is the probability that I will take an umbrella with me tomorrow given that it rains tomorrow. The CONDITIONAL PROBABILITY of an event A given another event A is written $\Pr(A \mid B)$. Conditional probabilities are defined by the following equation:

$$\Pr(A \mid B) \triangleq \frac{\Pr(A, B)}{\Pr(B)} \tag{2}$$

CHAIN RULE

By (2), we get the CHAIN RULE:

$$\Pr(A, B) = \Pr(A \mid B) \cdot \Pr(B) \tag{3}$$

Symmetrically,

$$\Pr(A, B) = \Pr(B \mid A) \cdot \Pr(A)$$

Thus,

$$\Pr(A \mid B) \cdot \Pr(B) = \Pr(B \mid A) \cdot \Pr(A)$$

and

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} \quad . \tag{4}$$

BAYES' THEOREM

This is BAYES' THEOREM, named for Thomas Bayes.

Probability distributions. For a given random variable X, the vector of probabilities $\Pr(X = x_i)$ for each $x_i \in Dom(X)$ characterizes a PROBABILITY DISTRIBUTION. We notate the distribution as $\Pr(X)$. Similarly, the joint probability distribution $\Pr(X, Y)$ gives values for each pair of values of the random variables X and Y. Many of the definitions and theorems about individual probabilities apply to distributions as well. For instance, we define conditional distributions

PROBABILITY DISTRIBUTION

$$\Pr(X \mid Y) = \frac{\Pr(X, Y)}{\Pr(Y)}$$

and the chain rule for distributions is

$$\Pr(X, Y) = \Pr(X \mid Y) \cdot \Pr(Y)$$

where the operations on the distributions are defined in terms of the normal matrix operations.

Independence. When the conditional probability $\Pr(X \mid Y)$ is identical to the unconditioned probability $\Pr(X)$, that is, the probability of X does not depend on Y, we say that X and Y are INDEPENDENT. For independent variables, then, the chain rule (3) becomes the PRODUCT RULE

INDEPENDENCE

PRODUCT RULE

$$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y) \quad .$$

Independence can be generalized to a conditional form. Event X is CONDITIONALLY INDEPENDENT of Y conditioned on Z when

CONDITIONAL INDEPENDENCE

$$\Pr(X, Y \mid Z) = \Pr(X \mid Z) \cdot \Pr(Y \mid Z)$$

in which case

$$\Pr(X \mid Y, Z) = \Pr(X \mid Z) \quad .$$

Marginals. Given a joint distribution $\Pr(X, Y)$, we can generate a new distribution by summing over all possible values of Y. Consider a particular value $x \in Dom(X)$. Then

$$\sum_{y \in Dom(Y)} \Pr(x, y) = \Pr(x) \quad .$$

Thus, considering the full joint distribution,

$$\sum_{y \in Dom(Y)} \Pr(X, y) = \Pr(X) \quad .$$

The distribution $\Pr(X)$ is thus retrievable from the joint probability $\Pr(X, Y)$ by this process of summing over Y, a process called MARGINALIZATION. We say that we have marginalized out Y from $\Pr(X, Y)$ to get the MARGINAL $\Pr(X)$.

MARGINALIZATION

MARGINAL

Name	Gender	Region	Law degree?	Party
Michele BACHMANN	F	Midwest	Yes	Republican
Karen BASS	F	West	No	Democratic
Judy CHU	F	West	No	Democratic
Barbara BOXER	F	West	No	Democratic
Max Sieben BAUCUS	M	West	Yes	Democratic
Xavier BECERRA	M	West	Yes	Democratic
Gus BILIRAKIS	M	South	Yes	Republican
Robert (Rob) BISHOP	M	West	No	Republican
Rodney ALEXANDER	M	South	No	Republican
Bill CASSIDY	M	South	No	Republican
Charles W. BOUSTANY	M	South	No	Republican
Dave LOEBSACK	M	Midwest	No	Democratic

TABLE 1. Some data about some members of Congress

3. Classification via the Naive Bayes method

Many problems in natural-language processing can be thought of as classification problems: They involve classifying words, phrases, sentences, paragraphs, or full documents into one of a small set of classes. For instance, we may want to classify emails as spam or non-spam; movie reviews as positive, neutral, or negative; product reviews as authentic or deceptive. We introduce here a simple and flexible but often quite effective machine learning algorithm for classification problems, the Naive Bayes method.

Let's start with some data, as shown in Table 1. You discover that Susan Davis is a congresswoman from the West without a law degree. Based on the data in the table, what would you guess Ms. Davis's party affiliation is: Democratic or Republican?

You might reason as follows: Republicans are rarely from the West and skew male, with a slight tendency not to have law degrees. Democrats on the other hand are frequently from the West and are slightly less likely to be male, with a similar slight tendency not to have law degrees. So given that otherwise there's no preponderance of Republicans over Democrats, she's probably a Democrat.

What is going on in your reasoning? You are using probabilities gleaned from the TRAINING DATA to (informally) estimate the probability that a certain feature of the person is observed given the particular class, and combining these bits of evidence to get an overall picture. This is essentially the Naive Bayes method. Let's characterize the method formally.

TRAINING DATA

We want to classify an object into one of a set of CLASSES $c_1, \ldots, c_k$. The object might be anything. Eventually, we'll be exploring documents as the objects in question, but here, the objects are Congress members, and the classes are Democratic and Republican. We classify the objects based on their observable FEATURES. Here the observable features are gender, region, and possession of a law degree. Each object is characterized by a vector of values for these features $\bar{x} = \langle x_1, \ldots, x_m \rangle$, for instance, for Ms. Davis $\langle \mathsf{female}, \mathsf{West}, \mathsf{no} \rangle$.

CLASSES

FEATURES

The fact that our initial reasoning seems inescapably tied to likelihood intimates that probabilities are the appropriate way of thinking about the issue: What is the probability that the person is in one or another class given the observed features? This is a conditional probability $\Pr(c_i \mid \bar{x})$. We can use Bayes theorem (4) to convert this probability to a combination that will turn out to be easier to generate estimates for:

$$\Pr(c_i \mid \bar{x}) = \frac{\Pr(\bar{x} \mid c_i) \cdot \Pr(c_i)}{\Pr(\bar{x})}$$

The term $\Pr(c_i)$ is called the PRIOR PROBABILITY for c_i, the probability that the class is c_i prior to having any information about the particular $\bar{x}$ we are considering. The term $\frac{\Pr(\bar{x} \mid c_i)}{\Pr(\bar{x})}$ constitutes the evidence for the observed $\bar{x}$ that the class c_i provides. The term $\Pr(c_i \mid \bar{x})$ is the POSTERIOR PROBABILITY, the prior updated with the evidence.

PRIOR PROBABILITY

POSTERIOR PROBABILITY

We want the class c_i that maximizes this posterior probability, that is,

$$\begin{aligned} \operatorname*{argmax}_i \Pr(c_i \mid \bar{x}) &= \frac{\Pr(\bar{x} \mid c_i) \cdot \Pr(c_i)}{\Pr(\bar{x})} \\ &= \operatorname*{argmax}_i \Pr(\bar{x} \mid c_i) \cdot \Pr(c_i) \end{aligned}$$

this last step because the denominator does not involve the i that we are maximizing over; it is essentially constant for all i.

4. Determining the probabilities of the observed features

The update involves estimating $\Pr(\bar{x} \mid c_i)$, the probability of the feature set $\bar{x}$ given the class. We can simplify this using the chain rule (3) for probabilities, provided here in a more general form:

$$\Pr(X, Y \mid \theta) = \Pr(X \mid \theta) \cdot \Pr(Y \mid X, \theta) \qquad .$$

Using this rule repeatedly, we have

$$\begin{aligned}
\Pr(\bar{x} \mid c_i) &= \Pr(x_1, \ldots, x_m \mid c_i) \\
&= \Pr(x_1 \mid c_i) \cdot \Pr(x_2, \ldots, x_m \mid x_1, c_i) \\
&= \Pr(x_1 \mid c_i) \cdot \Pr(x_2 \mid x_1, c_i) \cdot \Pr(x_3, \ldots, x_m \mid x_1, x_2, c_i) \\
\cdots &= \prod_{j=1}^{m} \Pr(x_j \mid x_1, \ldots, x_{j-1}, c_i) \qquad .
\end{aligned}$$

The individual elements in the product $\Pr(x_j \mid x_1, \ldots, x_{j-1}, c_i)$ correspond to the bits of evidence in our previous reasoning: "Democrats are frequently from the West", "Republicans skew male", "Republicans have a slight tendency not to have law degrees", and so forth. These are informal ways of saying that the probability that a congress member is from the West given that the member is a Democrat ($\Pr(\mathsf{West} \mid \mathsf{Democratic})$) is close to 1, or the probability that a member is male given that the member is Republican ($\Pr(\mathsf{male} \mid \mathsf{Republican})$) is higher than .5, etc.

But notice that this reasoning has the probability conditioned on c_i (being Democrat, say) but not on the other "earlier" features. For example, we weren't looking at $\Pr(\mathsf{West} \mid \mathsf{female}, \mathsf{Democratic})$. But the derivation above says that we should have. Implicitly, our reasoning was treating each of these pieces of evidence as *conditionally independent* of each other given the class. Each $\Pr(x_j \mid x_1, \ldots, x_{j-1}, c_i)$ was being approximated by $\Pr(x_j \mid c_i)$.

The assumption of independence is what makes the Naive Bayes method naive. It allows the further approximation

$$\begin{aligned}\Pr(\bar{x} \mid c_i) &= \prod_{j=1}^{m} \Pr(x_j \mid x_1, \ldots, x_{j-1}, c_i) \\ &\approx \prod_{j=1}^{m} \Pr(x_j \mid c_i) \qquad .\end{aligned}$$

Put all together, the Naive Bayes method selects the class c_i defined by

Naive Bayes method

$$\operatorname*{argmax}_i \Pr(c_i) \cdot \prod_{j=1}^{m} \Pr(x_j \mid c_i)$$

5. Maximum likelihood estimation of probabilities

Now all we need to do is to estimate the component probabilities $\Pr(c_i)$ and $\Pr(x_j \mid c_i)$ for each i and j. These estimates naturally come from looking at the training data that we were given, the table of Congress members and their features. When we said "Republicans tend to be from the South", we were really looking at the count of southern Republicans relative to the Republicans overall. We were estimating

$$\Pr(\mathsf{South} \mid \mathsf{Republican}) \approx \frac{\sharp(\mathsf{South}, \mathsf{Republican})}{\sharp(\mathsf{Republican})} = 4/6 \approx .66$$

COUNT NOTATION

(Here, we use the COUNT NOTATION $\sharp(\cdot)$ to indicate the number of objects in the training data with the specified features.)

MAXIMUM LIKELIHOOD ESTIMATE

This estimate for the probability is called the MAXIMUM LIKELIHOOD ESTIMATE (MLE). Suppose we have N samples of a random variable X that can take on different values $x_1, \ldots, x_{|X|}$. As above, we use the notation $\sharp(X = x_i)$ to indicate the count of events in the sample for which X takes on the value x_i (abbreviating as $\sharp(x_i)$ as usual). Then the maximum likelihood estimate of $\Pr(x_i)$ is given by

$$\widehat{\Pr}(X = x_i) \triangleq \frac{\sharp(X = x_i)}{N}$$

It follows that the MLE for conditional probabilities is

$$\widehat{\Pr}(A \mid B) = \frac{\sharp(A, B)}{\sharp(B)}$$

The maximum likelihood estimate is so named because of all possible estimates of the probability, this one generates the data with the highest likelihood. Suppose we flip a weighted coin, with a probability of heads of p, 6 times and it lands heads 4 times. What is the most likely value for p, the value that makes observing heads 4 times out of 6 most probable? We start by considering the probability q that a p-weighted coin lands heads 4 times out of 6.

$$q = \binom{6}{4} p^4 (1-p)^2 = 15(p^6 - 2p^5 + p^4)$$

Setting the derivative q' to zero,

$$q' = 15(6p^5 - 10p^4 + 4p^3) = 30p^3(p-1)(3p-2) = 0$$

we find inflection points at 0, 1, and 2/3. The first two are minima, the last the maximum. Thus if p is 2/3, we are most likely to see the pattern actually found in the data, 4 heads out of 6. And of course, 2/3 is exactly the empirical proportion of heads 4/6.[2]

Using MLE with Naive Bayes. We can compute maximum likelihood estimates for each of the needed probabilities, in order to find the class with the highest probability according to the Naive Bayes method. The estimated probability that Ms. Davis is a Republican is

$$\begin{aligned}
&\Pr(\mathsf{Republican} \mid \mathsf{female}, \mathsf{West}, \mathsf{no}) \\
&\qquad \propto \Pr(\mathsf{Republican}) \cdot \Pr(\mathsf{female} \mid \mathsf{Republican}) \\
&\qquad\quad \cdot \Pr(\mathsf{West} \mid \mathsf{Republican}) \cdot \Pr(\mathsf{no} \mid \mathsf{Republican}) \\
&\qquad \approx \frac{6}{12} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{4}{6} \\
&\qquad = \frac{1}{108}
\end{aligned}$$

The corresponding probability for Ms. Davis as a Democrat gives us

$$\Pr(\mathsf{Democratic} \mid \mathsf{female}, \mathsf{West}, \mathsf{no}) \propto \frac{5}{36}$$

We thus predict that Ms. Davis is a Democrat (as she in fact is).

[2] This may seem like a long run for a short slide. But while the point is apparently obvious, it is not actually obvious.

Let it be when it is mine to be sure let it be when it is mine when it is mine let it be to be sure when it is mine to be sure let it be let it be let it be to be sure let it be to be sure when it is mine to be sure let it to be sure when it is mine let it be to be sure let it be to be sure to be sure let it be to be sure let it be to be sure to be sure let it be to be sure let it be to be sure let it be to be sure let it be mine to be sure let it be to be sure to be mine to be sure to be mine to be sure to be mine let it be to be mine let it be to be sure to be mine to be sure let it be to be mine let it be to be sure let it be to be sure to be sure let it to be sure mine to be sure let it be mine to let it be to be sure to let it be mine when to be sure when to be sure to let it to be sure to be mine.

FIGURE 1. A sentence from Stein's "An Acquaintance with Description", 1929.

6. WORDS, TYPES, AND TOKENS

As we turn to processing of text, some standard terminology about words is useful, starting with the word "word" itself. The question of what is a word is itself somewhat fraught, as we'll see in more detail later. For the time being, let's just consider the words in a text to be the maximal sequences of alphabetic characters separated by whitespace. (As it turns out, this is an exceptionally poor definition, but sufficient for the time being.)

TYPES
TOKENS

We distinguish word TYPES from word TOKENS. A text w is made up of a series of word tokens $w_1 w_2 w_3 \cdots w_{|w|}$. Each word token belongs to a word type; by convention, we will use the symbol t to range over word types. Consider the text corpus in Figure 1, a sentence from Gertrude Stein's 1929 poem "An Acquaintance With Description". This corpus has 225 word tokens, which are instances of just eight word types (if we conflate case). The eight types, in decreasing order of frequency, are: "be", "to", "it", "sure", "let", "mine", "when", "is".