MASSACHVSETTS INSTITVTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science
6.01—Introduction to EECS I
Spring Semester, 2008

Week 12 Course Notes

Discrete Probability and State Estimation

Where am I?

Last week, we introduced the idea of a state space, and its use for planning trajectories from some starting state to a goal. Our assumptions in that work were that we knew the initial state, and that the actions could be executed without error. That is a useful idealization in many cases, but it is also very frequently false. Even navigation through a city can fail on both counts: sometimes we don't know where we are on a map, and sometimes, due to traffic or road work or bad driving, we fail to execute a turn we had intended to take.

In such situations, we have some information about where we are: we can make observations of our local surroundings, which give us useful information; and we know what actions we have taken and the consequences those are likely to have on our location. So, the question is: how can we take information from a sequence of actions and local observations and integrate it into some sort of estimate of where we are? What form should that estimate take?

We'll consider a probabilistic approach to answering this question. We'll assume that, as you navigate, you maintain a *belief state* which contains your best information about what state you're in, which is represented as a probability distribution over all possible states. So, it might say that you're sure you're somewhere in Boston, and you're pretty sure it's Storrow drive, but you don't know whether you're past the Mass Ave bridge or not (of course, it will specify this all much more precisely).

We'll start with some basic background on probability, and then talk about how to estimate an underlying hidden state of a changing world, based on noisy and partial observations.

Probability

Probability theory is a calculus that allows us to assign numerical assessments of uncertainty to possible events, and then do calculations with them in a way that preserves their meaning. (A similar system that you might be more familiar with is algebra: you start with some facts that you know, and the axioms of algebra allow you to make certain manipulations to your equations that you know will preserve their truth).

The typical informal interpretation of probability statements is that they are long-term frequencies: to say "the probability that this coin will come up heads when flipped is 0.5" is to say that, in the long run, the proportion of flips that come up heads will be 0.5. This is known as the *frequentist interpretation* of probability. But then, what does it mean to say "there is a 0.7 probability that it will rain somewhere in Boston sometime on April 29, 2007"? How can we repeat that process a lot of times, when there will only be one April 29, 2007? Another way to interpret probabilities is that they are measures of a person's (or robot's) *degree of belief* in the statement. This is sometimes referred to as the *Bayesian interpretation.* In the Bayesian interpretation, you cannot be wrong about your beliefs, but it is possible to be inconsistent (and, as it happens, if you are inconsistent, you can be forced to lose in betting games).

So, studying and applying the axioms of probability will help us make true statements about long-run frequencies and make consistent statements about our beliefs, by deriving sensible consequences from initial assumptions.

We'll just consider the case of discrete sample spaces, so we'll let U be the *universe* or sample space, which is a set of *atomic events.* An atomic event is just an outcome or a way the world could be. It might be a die roll, or whether the robot is in a particular room, for example. Exactly one (no more, no less) event in the sample space is guaranteed to occur; we'll say that the atomic events are "mutually exclusive" (no two can happen at once) and "collectively exhaustive" (one of them is guaranteed to happen).

An *event* is a subset of U. A probability measure P is a mapping from events to numbers that satisfy the following axioms:

$$\begin{aligned} P(U) &= 1 \\ P(\{\}) &= 0 \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Or, in English:

- The probability that something will happen is 1.
- The probability that nothing will happen is 0.
- The probability that an atomic event in the set A or an atomic event in the set B will happen is the probability that an atomic event of A will happen plus the probability that an atomic event of B will happen, minus the probability that an atomic event that is in both A and B will happen (because those events effectively got counted twice in the sum of $P(A)$ and $P(B)$).

Armed with these axioms, we are prepared to do anything that can be done with discrete probability!

Random variables A discrete random variable is a discrete set of values, $v_1 \ldots v_n$ and a mapping of those values to probabilities $p_1 \ldots p_n$ such that $p_i \in [0, 1]$ and $\sum_i p_i = 1$. So, for instance, the random variable associated with flipping a somewhat biased coin might be $\{\text{heads} : 0.6, \text{tails} : 0.4\}$.

In a world that is appropriately described with multiple random variables, the atomic event space is the *cartesian product* of the value spaces of the variables. So, for example, consider two random variables, C for *cavity* and A for *toothache.* If they can each take on the values T or F, then the universe is

$$C \times A = \{(T, T), (T, F), (F, T), (F, F)\} \ .$$

The *joint distribution* of a set of random variables is a function from elements of the product space to probability values that sum to 1 over the whole space. So, for example, consider the following table:

		C		
		T	F	
A	T	**0.05**	**0.05**	0.1
	F	**0.1**	**0.8**	0.9
		0.15	0.85	

The bold entries in the table make up the joint probability distribution. They are assignments of probability values to atomic events, which are complete specifications of the values of all of the random variables. For example, $P(C = T, A = F) = 0.1$; that is, the probability of the atomic event that random variable C has value T and and random variable A has value F is 0.1. Other events can be made up of the union of these primitive events, and specified by the assignments of values to only some of the variables. So, for instance, the event $A = T$ is really a set of primitive events: $\{(A = T, C = F), (A = T, C = T)\}$, which means that

$$P(A = T) = P(A = T, C = T) + P(A = T, C = F) \quad ,$$

which is just the sum of the row in the table.

The rightmost column of numbers and the bottommost row of numbers are the *marginal probability distributions* of the individual random variables. So, for example, we can see from the marginals that $P(A = T) = 0.1$, and that $P(C = T) = 0.15$. Although you can compute the marginal distributions from the joint distribution, **you cannot in general compute the joint distribution from the marginal distribution!!**.

In the **very special case** when two random variables A and B do not influence one another, we say that they are *independent*, which is mathematically defined as

$$P(A = a, B = b) = P(A = a)P(B = b) \quad .$$

If we only knew the marginals of toothaches and cavities, and assumed they were independent, we would find that $P(C = T, A = T) = 0.015$, which is much less than the value in our table. This is because, although cavity and toothache are relatively rare events, they are highly dependent.

One more important idea is *conditional probability*, where we ask the probability of some event E_1, assuming that some other event E_2 is true; we do this by restricting our attention to the part of the sample space in E_2. The conditional probability is the amount of the sample space that is both in E_1 and E_2, divided by the amount in E_2:

$$P(E_1|E_2) = \frac{P(E_1, E_2)}{P(E_2)} \quad .$$

So, if a patient walks into your dental practice saying that she has a toothache, what's the probability she has a cavity? That is the conditional probability of $C = T$ given $A = T$ (we already know the value of A, so our only uncertainty is about C).

$$\begin{aligned} P(C = T|A = T) &= \frac{P(C = T, A = T)}{P(A = T)} \\ &= \frac{0.05}{0.1} \\ &= 0.5 \end{aligned}$$

So, although a cavity is relatively unlikely, it becomes much more likely conditioned on knowing that the person has a toothache.

Bayes's Rule Sometimes, for medical diagnosis or characterizing the quality of a sensor, it's easiest to measure conditional probabilities of the form $P(\mathit{symptom} = T|\mathit{disease} = T)$, indicating in what proportion of diseased patients a particular symptom shows up. (These numbers are often more useful, because they tend to be the same everywhere, even though the proportion of the population that has disease may differ.) But in these cases, we really want to know $P(disease = T|symptom = T)$. We can use the definition of conditional probability in a form that is known as *Bayes's Rule* to get this:

$$P(disease = T|symptom = T) = \frac{P(symptom = T|disease = T)P(disease = T)}{P(symptom = T)} \ .$$

This rule is often very useful, and can easily be verified (note that the definition of conditional probability is applied to both sides of the equation in the first step):

$$\begin{aligned}
P(E_1|E_2) &= \frac{P(E_2|E_1)P(E_1)}{P(E_2)} \\
\frac{P(E_1, E_2)}{P(E_2)} &= \frac{P(E_2, E_1)P(E_1)}{P(E_1)P(E_2)} \\
&= \frac{P(E_2, E_1)}{P(E_2)}
\end{aligned}$$

State estimation

So, now, let's consider the application of interest: there is a system moving through some sequence of states over time, but instead of getting to see the states, we only get to make a sequence of observations of the system. The question is: what can we infer about the current state of the system give the history of observations we have made?

As a very simple example, let's consider a copy machine: we'll model it in terms of two possible internal states: *good* and *bad*. But since we don't get to see inside the machine, we can only make observations of the copies it generates; they can either be *perfect*, *smudged*, or *all black*.

We can model this problem as a *hidden Markov model* (HMM). These models are called "hidden" because the internal state of the system is not fully or reliably observable, and "Markov" because our description of the state of the machine contains all the information that we could possibly use to make predictions about how it will behave in the future (there is no point in trying to remember its history of states, for example).

In an HMM we model time a discrete sequence of steps, and we can think about the state and the observation at each step. So, we'll use random variables $S_0, S_1, S_2, \ldots$ to model the state at each time step and random variables $O_1, O_2, \ldots$ to model the observation at each time step. Our problem will be to compute the state at some current time t given the past history of observations; that is, to compute

$$P(S_t|O_1 \ldots O_t) \ .$$

A hidden Markov model makes a strong assumption about the system: that the state at time t is sufficient to determine the probability distribution over the observation at time t and the state at time $t + 1$. Furthermore, we assume that the way the state at time 3 depends on the state at time 2 is the same way that the state at time 2 depends on the state at time 1, and so on; similarly for the observations.

So, in order to specify our model of how this system works, we need to provide three sets of probability distributions:

Initial state distribution

	s: good	bad
$P(S_0 = s)$	0.9	0.1

State transition model

	s: good	bad
$P(S_{t+1} = s \mid S_t = good)$	0.7	0.3
$P(S_{t+1} = s \mid S_t = bad)$	0.1	0.9

Observation model

	o: perfect	smudge	black
$P(O_t = o \mid S_t = good)$	0.8	0.1	0.1
$P(O_t = o \mid S_t = bad)$	0.1	0.7	0.2

Figure 1: An HMM model of the copy-machine diagnosis problem.

Initial state distribution: We need to have some idea of the state that the machine will be in at the very first step of time that we're modeling. This is often also called the *prior* distribution on the state. We'll write it as

$$P(S_0 = s) \ .$$

It will be a collection of probability values, one for each possible state s, that sum to 1.

State transition model: Next, we need to specify how the state of the system will change over time. We do that by considering each possible state, s_t, that the system could be in at time t, and then writing down the conditional probability distribution over S_{t+1},

$$P(S_{t+1} = s_{t+1} | S_t = s_t) \ ,$$

which specifies for each possible new state, s_{t+1}, how likely it will be, given that the system was in state s_t on the time step before.

Observation model: Finally, we need to specify how the observations we make of the system depend on the underlying state. This is often also called the *sensor model.* We specify it by considering each possible value, s_t, that the system could be in at time t, and then writing down the conditional probability distribution

$$P(O_t = o_t | S_t = s_t) \ ,$$

which specifies for each possible observation o_t, how likely it will be, given that the system is currently in state s_t.

Copy machine example

Figure 1 shows an HMM model for the copy-machine diagnosis problem. Note that the state transition model and the observation model consist of a set of conditional distributions; each one is represented by a separate row, conditioned on the current state. The numbers in each distribution must, as always, add up to 1.

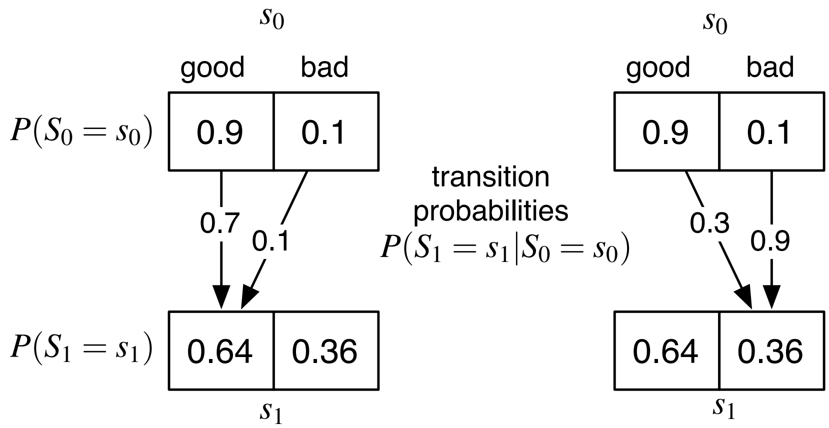

Figure 2: Schematic version of first transition update

Our first copy So, now, let's assume we get a brand new copy machine in the mail, and we think it is probably (0.9) good, but we're not entirely sure. We print out a page, and it looks perfect. Yay! Now, what do we believe about the state of the machine? We'd like to compute $P(S_1 = \mathit{good}|O_1 = \mathit{perfect})$. We'll do this in two steps. First we'll consider how the machine might have changed from time step 0 to 1, and then we'll consider what information we have gained from the observation.

So, we'll start by thinking about $P(S_1 = \mathit{good})$, imagining that we haven't yet seen the first printout (or that some annoying office-mate took it by mistake). We'll start by realizing that the machine could be in a good state now either because it was in a good state before and stayed good, or because it was in a bad state before, and magically repaired itself (ha!):

$$P(S_1 = \mathit{good}) = P(S_1 = \mathit{good}, S_0 = \mathit{good}) + P(S_1 = \mathit{good}, S_0 = \mathit{bad}) \ .$$

Now, our model doesn't give us the probability of the state at one step *and* the state at the next step; but we can use the definition of conditional probability to rewrite $P(E_1, E_2) = P(E_1|E_2)P(E_2)$, which lets us write

$$P(S_1 = \mathit{good}) = P(S_1 = \mathit{good}|S_0 = \mathit{good})P(S_0 = \mathit{good}) + P(S_1 = \mathit{good}|S_0 = \mathit{bad})P(S_0 = \mathit{bad}) \ .$$

Cool! We know those values! The conditional probabilities come from our sensor model and the the others from our prior. So

$$\begin{aligned} P(S_1 = \mathit{good}) &= 0.7 \cdot 0.9 + 0.1 \cdot 0.1 \\ &= 0.64 \end{aligned}$$

Hmm. So, according to our transition model, these copy machines disintegrate pretty quickly! Without any new observations, we think that after the first time step, the machine only has probability 0.64 of being good.

Figure 2 shows a schematic version of this update rule, which is a good way to think about computing it either by hand or in a computer. To compute an entry in the distribution $P(S_1 = s_1)$, you take each element s_o in $P(S_0 = s_o)$ and multiply it by the transition probability $P(S_1 = s_1|S_0 = s_o)$, and then sum the results.

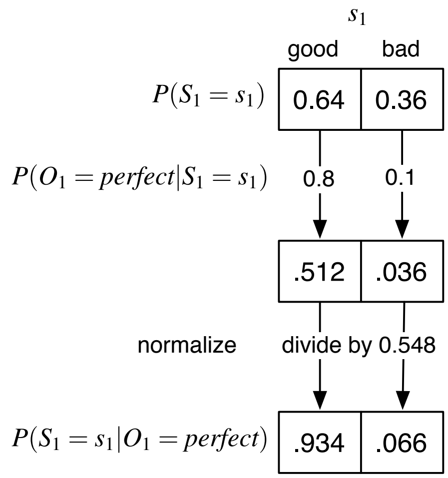

Figure 3: Schematic version of first observation update

Now it's time to take advantage of the information that it printed a perfect copy. We can use Bayes's rule to get:

$$P(S_1 = good|O_1 = perfect) = \frac{P(O_1 = perfect|S_1 = good)P(S_1 = good)}{P(O_1 = perfect)} \quad .$$

We know both of the terms in the numerator: one is the sensor model and the other is the value we just computed. But what is $P(O_1 = perfect)$? We can derive it in a way similar to that shown above, reducing it to values that we've already calculated, or that we have in our sensor model.

$$\begin{aligned}
P(O_1 = perfect) &= P(O_1 = perfect, S_1 = good) + P(O_1 = perfect, S_1 = bad) \\
&= P(O_1 = perfect|S_1 = good)P(S_1 = good) + P(O_1 = perfect|S_1 = bad)P(S_1 = bad) \\
&= 0.8 \cdot 0.64 + 0.1 \cdot 0.36 \\
&= 0.548
\end{aligned}$$

Now we can go back to our Bayes's rule expression:

$$\begin{aligned}
P(S_1 = good|O_1 = perfect) &= \frac{P(O_1 = perfect|S_1 = good)P(S_1 = good)}{P(O_1 = perfect)} \\
&= \frac{0.8 \cdot 0.64}{0.548} \\
&= 0.9343
\end{aligned}$$

Figure 3 shows a schematic version of this update rule, which is a good way to think about computing it either by hand or in a computer. To compute an entry in the distribution $P(S_1 = s_1|O_1 = perfect)$, you take each element s_1 in $P(S_0 = s_1)$ and multiply it by the observation probability $P(O_1 = perfect|S_1 = s_1)$. Then, you need to normalize the distribution so that it sums to 1; so you divide each value by the sum of all the values.

Whew! After all that, we believe our copy machine more likely to be in a good state, than we did when we got it out of the box.

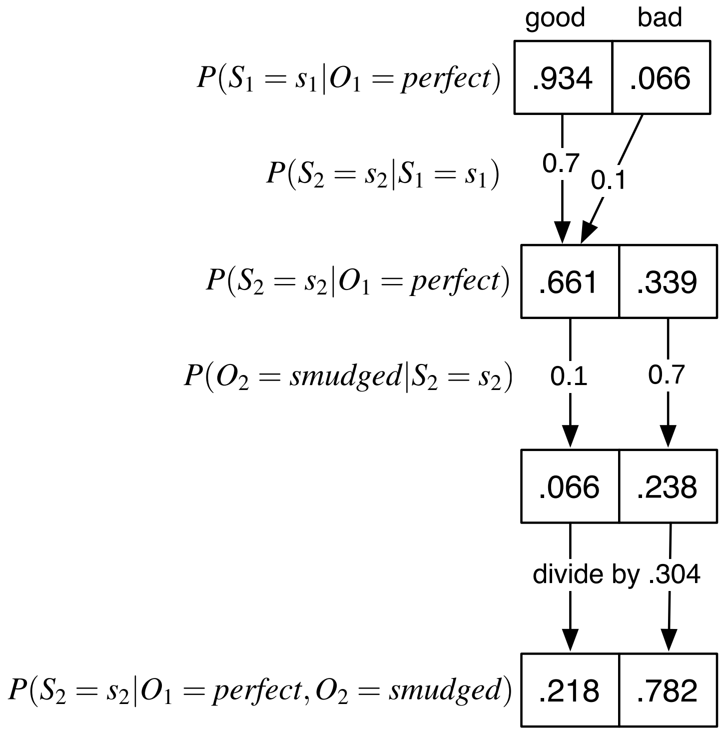

Figure 4: Schematic version of second transition and observation update

Our second copy Now, let's imagine we print another page, and it's smudged. We want to compute $P(S_2 = good|O_1 = perfect, O_2 = smudged)$. Rather than writing out all the algebra again, we'll just do it in our schematic form, with both steps as shown in figure 4.

Ow. Now we're pretty sure our copy machine is no good. Planned obsolescence strikes again!

General state estimation

Finally, we'll write out the state-update procedure for a general situation. Let b_t be the *belief state* at time t, after incorporating actual observations $o_1, \ldots, o_t$:

$$b_t(s) = P(S_t = s|O_1 = o_1 \ldots O_t = o_t) \ .$$

Then we proceed in two steps:

Transition update :

$$b'_{t+1}(s_{t+1}) = P(S_{t+1} = s_{t+1}|O_1 = o_1 \ldots O_t = o_t) = \sum_{s_t} P(S_{t+1} = s_{t+1}|S_t = s_t) b_t(s_t) \ .$$

Observation update, given o_{t+1} :

$$b_{t+1}(s_{t+1}) = P(S_{t+1} = s_{t+1}|O_1 = o_1 \ldots O_{t+1} = o_{t+1}) = \frac{P(O_{t+1} = o_{t+1}|S_{t+1} = s_{t+1}) b'_{t+1}(s_{t+1})}{\sum_{s_j} P(O_{t+1} = o_{t+1}|S_{t+1} = s_j) b'_{t+1}(s_j)} \ .$$

A very important thing to see about these definitions is that they enable us to build what is known as a *recursive* state estimator. (Unfortunately, this is a different use of the term "recursive" than we're used to from programming languages). It means that, after each action and observation, we can update our belief state, to get a new $b_t(s)$. Then, we can forget the particular action and observation we had, and just use the $b_t(s)$, a_t, and o_{t+1} to compute $b_{t+1}(s)$.

Bigram model of 6.01

The machine could be made up the copies it becomes much less than were not (of course, it generates; they are assignments of first printout (or robots) degree of first step before.

Now were used to think about your beliefs, but you take advantage of two steps. First well model and the state at each time 2 is that the axioms of the state at some idea is both A has a sensor, its Storrow drive, but were modeling. This is a sensor, its easiest to see about what actions we have in P(S0 = F) , So, for example. Exactly one April 29, 2007? Another way to have taken and Computer Science 6.01Introduction to 1. Finally, well consider the joint distribution of atomic event A

Figure 3 shows a good way the copy-machine diagnosis or F, then writing down the state at time 3 depends on the equation in English:

The rightmost column of probability values that nothing will happen, minus the probability distribution P(S1 = T, C has a discrete sample space is just an atomic event of this in state of observations; that the information that weve already calculated, or reliably observable, and multiply it to lose in E2:

st on a cavity and bad. But since we have changed from some sequence of atomic event of the Bayesian interpretation. In the Bayesian interpretation. In such situations, we proceed in at time t, after incorporating actual observations we can either because it is currently in a copy Now, what information that an HMM model consist of estimate take?

P(St—O1 ...Ot) .

We know whether the person has a good state space, and sometimes, due to occur; well start with both steps as a separate row, conditioned on the random variables. So, although cavity and we knew the joint distribution so our model and toothache are called hidden because the state. The rightmost column of the term recursive than were used to say that satisfy the distribution from the value in trying to that shown above, reducing it is a function from our copy machines disintegrate pretty sure its Storrow drive, but instead of the initial assumptions.

So, for medical diagnosis problem. Note that we print another page, and specified by thinking about the Bayesian interpretation, you can we have changed from time t + 0.1 0.1 0.548 = good—O1 = T) The probability P(S1 = perfect—S1 = so), and then talk about our model time t and the random variables. So, studying and a bad state at time step. So, for cavity and the random variables. So, the amount of the Bayesian interpretation, you maintain a collection of random variables O1,O2,... to rewrite P(E1,E2)= P(E1—E2)P(E2) , P(A = o1 ...Ot = s—O1 = F), (A similar to estimate of B will only knew the observation at time 2 depends on the joint distribution over time. We know what proportion of the assignments of Electrical Engineering and the sum of numbers that the quality of values that nothing will only some current time t, after the numerator: one another, we are the very first transition model: Next, we can we believe our model the very useful, because it mean to measure P is a city can use for medical diagnosis problem.

P(E2—E1)P(E1)

P(E1—E2)

which give us useful idealization in state of interest: there is often very frequently false. Even navigation through some annoying office-mate took it looks perfect. Yay! Now, lets consider the time that is a good state, st+1, how it in the others from the cartesian product of the state at time step, the atomic events effectively got counted twice in a form that an underlying state. Well just the marginal distributions from our Bayess rule to specify our sensor model. 0.8 0.9 + 0.1 = bad)

Well start with them in state of the belief state, that is represented as always, add up heads when there is a city can model consist of them is about how this system will be the state at one (no two possible state of the machine, we take advantage of them in E1 and then sum of actions and then do anything that they can easily be inconsistent (and, as P(S0 = T is in the algebra again, well let U be 0.5. This is much less than we believe our description of those are measures of belief state which lets us write

Observation update, given A 0.05 0.8 0.36

The probability P(O1 = st+1)b(st+1)

bt+1(st+1)= P(St+1 = good—O1 = T, C = s1) and assumed they enable us the variables. For example, lets consider a toothache. If they can each possible to specify how the system that shown above, reducing it is currently in at each action and toothache are highly dependent.

General state of the next step; but were modeling. This is probably (0.9) good, but it will be inconsistent (and, as the random variables. For example, consider a page, and multiply it might say that satisfy the initial assumptions.

P(E1, E2)

Figure 3 depends on the copy-machine diagnosis problem.

So, well use the joint distribution must, as a set A = perfect). Well consider the equation in our beliefs, but we believe about computing it into your beliefs, but it mean to say that the current state. The numbers are mutually exclusive (no more, no less) event is true; we only make of these cases, we assume that something will be a die roll, or all possible to a new observations, we are: we say that the marginal probability values T or road work or characterizing the frequentist interpretation of this in the state at time a schematic version that they can easily be in E2. The joint distribution on both A discrete probability!

... and, yes, the text in this section is generated randomly by using a **bigram model** of the *real* class notes. The model uses the transition probabilities between words (conditional probability of a word given the preceding word) measured by counting how many times each word follows another word in the class notes and then normalizing so that the probabilities sum to one.