Naïve Bayesian Classifier

To apply the Naïve Bayesian Classifier to the Room Occupancy data, we must first organize the data in a useful manner.

There are two categories for this data: Occupied or Not Occupied.

There are five or six attributes to consider. The five that you must use are Temperature, Humidity, Light, CO2, and HumidityRatio. If you use time as an attribute, then you probably should use only time of day and not the rest of the date.

Be aware that using time suggests that the room schedule is the same each day during the period the data was collected and that the schedule will continue in future. If the schedule was fixed during training but changes later, then the predictions may be incorrect. If the schedule remains the same, then using time will likely greatly improve the predictions. Notice that we don't need to try to determine the room schedule ourselves. The models can determine that for us if it exists.

What is the interval of values in the training data?

1. Determine the interval of values for each attribute. This is just [low_value, high_ value] for each attribute.

2. You might want to consider the interval of values for each attribute separated by category as well. Note: If there is no overlap in the intervals for each category, then we would be able to predict that test data in a given interval belongs to the corresponding category without doing any additional calculations. This would happen if and only if the data were linearly separable.

3. Separate each interval from **1** into *equal-length* subintervals. With our data, five subintervals for each attribute are probably sufficient.

   For example, temperature ranges from 19 to 23.18. Using three ***poorly*** chosen intervals, we create the following table:

Interval	Occupied	Not Occupied	Pr(Occupied)	Pr(Not Occupied)
[19, 19.5)	0	1280	0	1
[19.5, 23.1)	1725	5134	0.251494	0.748506
[23.1, 23.18]	4	0	1	0

   Notice that the intervals do not overlap. All except the last go up to but *not including* the data value at the right of the interval. Based just on this data, temperatures below 19.5 indicate that the room is Not Occupied but temperatures at or above 23.1 indicate the room is Occupied. Temperatures in between those values are more unclear, since the probability that the room is Not Occupied is dominated by the fact that there is so much more Not Occupied data. However, with more evenly spaced intervals, temperature distinctions should become more apparent.

4. Consider a similar table for each of the five (or six) attributes. Each table would have data from the five subintervals used for that attribute.

In our example, we see that Pr(Occupied | [19.5, 23.1)) = 0.251494. This is one of the values Pr(C | A) referred to in the lab.

5. In Bayes's Theorem, we use the information to switch between Pr(X | Y) and Pr(Y | X). In most cases, one of those values is easier to calculate than the other. Part of our task is to determine which is important to our needs and how to fill in the remaining components in order to use the theorem appropriately.

6. For **2c** in the lab, we need to use *all* the attributes to determine our prediction. Given a test vector ***t***, we have (possibly conflicting) information from each of the attributes. We wish to use all that information, and the corresponding probabilities, to predict the category.

To determine the prediction, which is based on the probability of being in category C given the data that we have observed, we need to calculate Pr(C | ***t***). This means we must combine information from all the attributes.

It is useful to think of Pr(C | ***t***) as Pr(C | [a0, a1, a2, a3, a4, a5]) where each of the values in the list corresponds to the value from the particular attribute. (If you do not use time, then there are five rather than six attribute values.)

At this point, we apply the formula from Bayes's Theorem. You might want to follow our "bad code" example from class or **A more complicated example** on the Wikipedia page for Bayes's Theorem.

The "naïve" part of the name comes from the fact that we assume the attributes are independent, which is probably not accurate since lights being on most likely raises the temperature in the room at least somewhat. You can probably think of other connections between some of the attributes as well.

7. From **5**, we have probabilities for which of the categories a particular data point might belong. Using MAP, we simply predict the most likely of the two. Some methods return only the probability. Other methods return a prediction along with its probability.

8. As in other classification problems, we then compare our predictions to the correct answers for the test data to see how well our algorithm performs.