Naïve Bayesian Classifier

Let **X** be the test vector. We need Pr(Occupied | **X**) and Pr(Unoccupied | **X**). We have already preprocessed the data by putting it into intervals and have calculated the probabilities of Occupied and Unoccupied for each of the intervals.

By Bayes's Theorem,

$$\Pr(\text{Occupied} \mid \mathbf{X}) = (\Pr(\text{Occupied}) * \Pr(\mathbf{X} \mid \text{Occupied})) / \Pr(\mathbf{X})$$

and

$$\Pr(\text{Unoccupied} \mid \mathbf{X}) = (\Pr(\text{Unoccupied}) * \Pr(\mathbf{X} \mid \text{Unoccupied})) / \Pr(\mathbf{X})$$

We can calculate Pr(Occupied) and Pr(Unoccupied) from the training data, so it is the other values that are important to us.

Since Pr(**X** | Occupied) is the probability of that particular **X** given that the room is Occupied, we need to find a way to calculate this based on the attributes we are using. The "naïve" part of the classifier assumes that all the attributes are independent, meaning that the value of one attribute has no effect on the values of other attributes. In real life this is often not the case, but this assumption allows us to create a quick model that is sometimes not too incorrect.

The independence assumption allows us to calculate Pr(**X** | Occupied) by considering each of the *n* attributes separately, so

$$\mathbf{OCC} = \Pr(\mathbf{X} \mid \text{Occupied}) = \Pr(\mathbf{X}\_1 \mid \text{Occupied}) * \Pr(\mathbf{X}\_2 \mid \text{Occupied}) * \ldots * \Pr(\mathbf{X}\_n \mid \text{Occupied})$$

The value of Pr(**X**_*i* | Occupied) for each *i* is available from the preprocessed data.

Similarly,

$$\mathbf{UNOCC} = \Pr(\mathbf{X} \mid \text{Unoccupied}) = \Pr(\mathbf{X}\_1 \mid \text{Unoccupied}) * \ldots * \Pr(\mathbf{X}\_n \mid \text{Unoccupied})$$

This gives

$$\Pr(\text{Occupied} \mid \mathbf{X}) = \Pr(\text{Occupied}) * \mathbf{OCC} / \Pr(\mathbf{X})$$

and

$$\Pr(\text{Unoccupied} \mid \mathbf{X}) = \Pr(\text{Unoccupied}) * \mathbf{UNOCC} / \Pr(\mathbf{X})$$

At this point, we do not know the left-hand side of the equation or Pr(**X**). However, since **X** is a particular outcome, Pr(**X**) is a constant. Furthermore, since it is in both equations, we can ignore Pr(**X**) and focus on the numerators.

Suppose Pr(Occupied) * **OCC** = 0.0125 and Pr(Unoccupied) * **UNOCC** = 0.0075. To calculate the desired probabilities, we normalize these values to yield

$$\Pr(\text{Occupied} \mid \mathbf{X}) = 0.0125 / (0.0125 * 0.0075) = 62.5\%$$

and

$$\Pr(\text{Unoccupied} \mid \mathbf{X}) = 0.0075 / (0.0125 + 0.0075) = 37.5\%$$

If we just want to give the probability of each category we can stop here, but if we need a single prediction then we select the most likely category and predict Occupied for this test point.