

Probability and Naïve Bayes

The following identities will be useful for these exercises:

Marginalization or “summing out”. Note that Y and Z can represent sets of variables. Given a table with the full joint PD we can calculate the probability of Y by adding up all of the rows that match.	$P(Y) = \sum_{z \in Z} P(Y, z)$
Definition of conditional probability	$P(X Y) = \frac{P(X, Y)}{P(Y)}$
Chain rule	$P(X, Y) = P(Y X)P(X) = P(X Y)P(Y)$
Bayes Theorem	$P(Y X) = \frac{P(X Y)P(Y)}{P(X)}$
Naïve Bayes: attributes conditionally independent given the class Y .	$P(Y X_1, \dots, X_d) \propto P(Y) \prod_{i=1}^d P(X_i Y)$

1. Estimating Joint Probability Distributions

It is your mission to develop a probabilistic model that will enable the JMU administration to uncover North Korean secret agents among the student body. The JMU secret police have randomly pulled in 20 individuals and determined their agent status as well as whether or not they golf, and what sort of hat they wear (Fedora or Fez). The data set below represents the raw data.

(spy, golf, fedora)	(loyal, golf, fez)
(spy, golf, fez)	(loyal, golf, fez)
(spy, golf, fez)	(loyal, ¬golf, fedora)
(spy, golf, fez)	(loyal, ¬golf, fedora)
(spy, ¬golf, fedora)	(loyal, ¬golf, fedora)
(loyal, golf, fedora)	(loyal, ¬golf, fedora)
(loyal, golf, fedora)	(loyal, ¬golf, fedora)
(loyal, golf, fedora)	(loyal, ¬golf, fedora)
(loyal, golf, fedora)	(loyal, ¬golf, fez)
(loyal, golf, fez)	(loyal, ¬golf, fez)

(a) Use the data above to fill in the following joint probability table.

Spy	Golfer	Fedora	Probability
T	T	T	
T	T	F	
T	F	T	
T	F	F	
F	T	T	
F	T	F	
F	F	T	
F	F	F	

- (b) Do you think these probabilities are likely to be accurate (in our hypothetical world)? I.e. do you think they represent the true probability of seeing these outcomes if we draw a random student from the population? Why or why not?
- (c) According to your probability distribution, what is the probability that a randomly selected student is a spy who golfs and wears a fez?
In other words, what is $P(Spy = T, Golf = T, Fedora = F)$?
- (d) What is the marginal probability $P(Spy = T)$? Show your work.
- (e) What is the marginal probability $P(Golf = T, Fedora = F)$? Show your work.
- (f) Use Bayes Theorem to calculate $P(Spy = T \mid Golf = T, Fedora = F)$ Show your work.

2. Scaling Up: the Naïve Bayes Assumption

Our table in Question 1 needed one row for every combination of the three variables.

- (a) Suppose the secret police had recorded 10 binary attributes for each student instead of two. How many rows would the full joint probability table have? (As in Question 1, the table has one row for every combination of the class *and* the attributes.) Why is this a problem if we still only have 20 samples?

- (b) One solution is to assume that the attributes are conditionally independent given the class:

$$P(Golf, Fedora | Spy) = P(Golf | Spy) P(Fedora | Spy)$$

This is the “naïve” assumption. Now we can estimate each class-conditional probability separately. Fill in the table below from the raw data.

	<i>Spy</i> = T (5 students)	<i>Spy</i> = F (15 students)
$P(Golf = T Spy)$		
$P(Golf = F Spy)$		
$P(Fedora = T Spy)$		
$P(Fedora = F Spy)$		

- (c) Bayes Theorem tells us

$$P(Spy | Golf, Fedora) = \frac{P(Golf, Fedora | Spy) P(Spy)}{P(Golf, Fedora)}$$

The denominator is the same for both classes, so to decide which class is more likely we only need to compare the numerators. Using the naïve assumption, compute

$$P(Spy) P(Golf = F | Spy) P(Fedora = F | Spy)$$

for each class, for a new student who does **not** golf and wears a **fez**. Which class wins?

- (d) The two numerators from part (c) do not sum to 1, so they cannot be read as probabilities. **Normalize** them: divide each one by the sum of both. The results do sum to 1, and are the naïve Bayes model’s estimate of how likely each class is. What probability does the model assign to this student being a spy?

3. Smoothing

- (a) Suppose the secret police had also recorded whether each student owns a cat, and that *none* of the 20 students, spy or loyal, owned one. A cat owner who does not golf and wears a fez is brought in. Compute both naïve Bayes numerators as in Question 2(c), now including the cat term, and then try to normalize them as in Question 2(d). What goes wrong?

- (b) The usual fix is **Laplace smoothing**: pretend we saw one extra example of every value before we started counting.

$$P(X_i = x \mid Y = y) = \frac{\text{count}(X_i = x, Y = y) + 1}{\text{count}(Y = y) + |V_i|}$$

where $|V_i|$ is the number of possible values for attribute X_i . The table below applies smoothing to the values from Question 2(b) and adds the cat attribute. Fill in the missing cat entries.

	<i>Spy</i> = T (5 students)	<i>Spy</i> = F (15 students)
$P(\textit{Golf} = \text{T} \mid \textit{Spy})$	$\frac{4+1}{5+2} = \frac{5}{7}$	$\frac{7+1}{15+2} = \frac{8}{17}$
$P(\textit{Golf} = \text{F} \mid \textit{Spy})$	$\frac{1+1}{5+2} = \frac{2}{7}$	$\frac{8+1}{15+2} = \frac{9}{17}$
$P(\textit{Fedora} = \text{T} \mid \textit{Spy})$	$\frac{2+1}{5+2} = \frac{3}{7}$	$\frac{10+1}{15+2} = \frac{11}{17}$
$P(\textit{Fedora} = \text{F} \mid \textit{Spy})$	$\frac{3+1}{5+2} = \frac{4}{7}$	$\frac{5+1}{15+2} = \frac{6}{17}$
$P(\textit{Cat} = \text{T} \mid \textit{Spy})$		
$P(\textit{Cat} = \text{F} \mid \textit{Spy})$		

- (c) Redo part (a) using your smoothed values. After normalizing, what probability does the smoothed model assign to the cat owner being a spy?

- (d) In one sentence: what happens to the effect of smoothing as the amount of training data grows?