Bayes' Rule (as will be used here) :

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$ is the *prior* probability. This reflects *background knowledge*.
- $P(D)$ is the probability that the data happens, *given no knowledge of the hypothesis*.
- $P(D|h)$ probability of the data, given that the hypothesis holds
- But we are more interested in $P(h|D)$, the *posterior* probability. This is our confidence that the hypothesis holds, after observing D.

Question 1. We often want to find the best hypothesis that fits the data. What does that even mean?

Bayes' Rule (more general form) :
Given two events E and F with non-zero probability, then:

$$p(F|E) = \frac{p(E|F)p(F)}{p(E|F)p(F) + p(E|\overline{F})p(\overline{F})}$$

Question 2. *My company is going to introduce a new product. Every time we release a new product, we do market research to predict if it will be a success or failure. 60% of the products we release have been successes. Of the products that were successful, 70% were predicted to be a success. Of the products that were failures, 40% were predicted to be a success. Our new product has been predicted to be a success.*

Suppose we want to use Bayes' Rule to answer the question: What is the probability our new product is successful?

What are the values we plug into the formula for the following?

- $p(E|F)$
- $p(F)$
- $p(E|\overline{F})$
- $p(\overline{F})$

Question 3. There are two boxes. The first contains two gold marbles and seven blue marbles. The second contains four gold marbles and three blue marbles. You choose a box at random, then you choose a marbles at random. You draw out a blue marble. What is the probability that the marble came from the first box?

Question 4. There is a rare disease which infects only 1 out of $100,000$ people. You can detect it with a very accurate diagnostic test. If someone has the disease, it correctly identifies it 99% of the time. If someone does not have the disease, it correctly states this 99.5% of the time.

- Suppose the test comes out negative. What is the probability the person does not have the disease?
- Suppose the test comes out positive. What is the probability the person does have the disease?

As we seek to connect Bayes Theorem to Concept Learning, we will start by thinking about a brute force MAP (*maximum a posteriori*) hypothesis learning program. Note that the initial brute force algorithm is *impractical* as far as implementation and usage, although it will be useful for other reasons.

The brute force algorithm proceeds as follows: for each hypothesis $h \in H$, calculate $P(h|D)$. Then, output the hypothesis with the highest calculated $P(h|D)$ value.

An algorithm is a **consistent learner** if it always outputs a hypothesis that commits zero training error for the data it is given (with a reasonable assumption here).

Question 5. Name an algorithm we have seen so far that is a consistent learner. There are at least two.

Question 6. Name an algorithm we have seen so far that is **not** a consistent learner. There is at least one.

Bayes Optimal Classifier

We were asking: most probable *hypothesis*?

We should ask: most probable *classification* of new instance?

Question 7. Are these the same thing?

$P(v_j|D)$: probability correct classification for new instance is v_j:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Bayes Optimal Classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_j)P(h_i|D)$$

A less optimal algorithm is the Gibbs Algorithm, which works as follows:

1. Choose $h \in H$ at random, proportional to the posterior distribution over H
2. Use h to classify next instance

Surprisingly, this can be shown to have no worse than twice the expected error of the Bayes Optimal Classifier, while being far more efficient.
Question 8. Suppose during Concept Learning, we apply this idea: when we want to classify a query posed to us, we choose a hypothesis uniformly at random from the current version space and use that to classify the instance. How does this behave compared to an optimal classifier?

Naive Bayes Classifier

Let's revisit the PlayTennis? example data set.

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Question 9. Estimate $P(\text{PlayTennis} = \text{Yes})$ and $P(\text{PlayTennis} = \text{No})$

Question 10. Estimate $P(\text{Wind=strong}|\text{PlayTennis} = \text{Yes})$ and similarly for no.

Question 11. Do we play tennis in the following circumstance?

Outlook	Temperature	Humidity	Wind
Sunny	cool	high	strong

Question 12. In Naive Bayes, we estimated $P(x|y)$ as

$$\frac{\#\text{ observed } x \text{ and } y \text{ together}}{\#\text{ observed } y}$$

When might this be a poor idea, and what can we do differently?

Building a Spam Filter

Goal: Build a classifier that, given an email, classifies it as spam ("yes") or not ("no").

Training Phase

During the training phase, we have lots of emails that are labeled, each as "spam" or "not spam."
Question 13. What are the priors we compute, and how?

Question 14. What do we compute for each word w_i?

Question 15. What is $P(x|\text{spam})$?

Classification Phase

Given an email that we wish to classify:

- Count the words
- Apply weights of spam, not-spam categories
- Make a decision: if spam score is higher, it's spam.

Question 16. What does it mean when we say that a Naive Bayes are **generative** classification models?

This is an artificial question intended to help you review Naive Bayes in anticipation of more advanced Bayesian techniques. You should feel free to use a computer or calculator (or phone, etc) for this problem.

Consider a binary classification problem with variable $X_1 \in \{0, 1\}$ and label $Y \in \{0, 1\}$. The true generative distribution $P(X_1, Y) = P(Y)P(X_1|Y)$ is shown below:

$Y = 0$	$Y = 1$
0.8	0.2

	$X_1 = 0$	$X_1 = 1$
$Y = 0$	0.7	0.3
$Y = 1$	0.3	0.7

1. Now suppose we have trained a Naive Bayes classifier, using infinite training data generated according to those tables. Now fill in Table 3. In particular, fill in the probabilities in the first two columns, and fill in the prediction of Y in the last column of the table. The process is sufficient (i.e., 0.8×0.7 is fine, as would be 0.56)

	$\hat{P}(X_1, Y = 0)$	$\hat{P}(X_1, Y = 1)$	$\hat{Y}(X_1)$
$X_1 = 0$			
$X_1 = 1$			

2. What is the expected error rate of this classifier on training examples generated according to the first two tables? In other words, what is $P(Y \neq \hat{Y}(X_1))$?
(Hint: $P(Y \neq \hat{Y}(X_1)) = P(Y \neq \hat{Y}(X_1), X_1 = 1) + P(Y \neq \hat{Y}(X_1), X_1 = 0))$

3. Now we add a feature to this data X_2 such that X_2 is an exact duplicate of X_1. Suppose we have trained Naive Bayes classifier using infinite training data that are generated by following the first two tables, and then add the additional duplicate feature X_2. Please fill in the following tables.

	$X_2 = 0$	$X_2 = 1$
$Y = 0$	0.7	0.3
$Y = 1$	0.3	0.7

Fill in the probabilities for the following table and write down the predictions of Y for different X_1 and X_2 value combinations.

	$\hat{P}(X_1, X_2, Y = 0)$	$\hat{P}(X_1, X_2, Y = 1)$	$\hat{Y}(X_1, X_2)$
$X_1 = 0, X_2 = 0$			
$X_1 = 1, X_2 = 0$			
$X_1 = 0, X_2 = 1$			
$X_1 = 1, X_2 = 1$			

4. What is the expected error rate of this Naive Bayes classifier on this data?

5. Compare the error rate in d to the error rate in b. What is the reason for the difference?