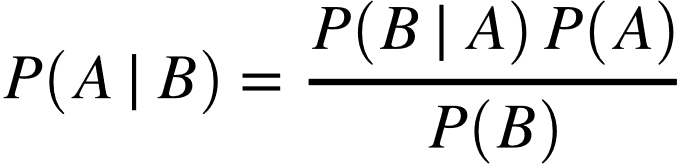
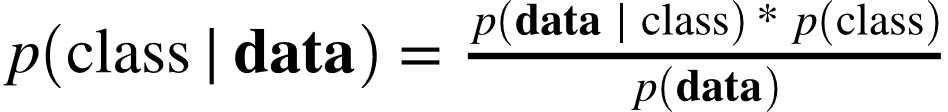
**Bayes Theorem**

Bayes theorem is a famous equation that allows us to make predictions based on data. Here is the classic version of the Bayes theorem:



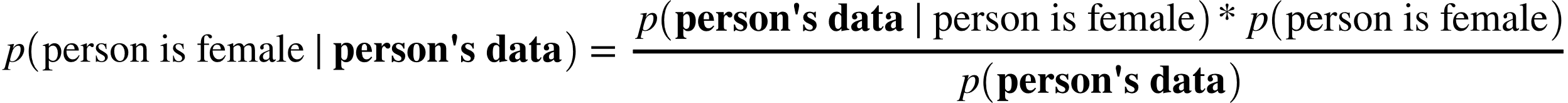
This might be too abstract, so let us replace some of the variables to make it more concrete. In a bayes classifier, we are interested in finding out the class (e.g. male or female, spam or ham) of an observation given the data:



Where :

* Class is a particular class (e.g - male)
* Data is an **observation’s data**
* P(Class | Data) is called the **posterior**
* P(data | class) is called the **likelihood**
* P(class) is called the **prior**
* P(data) is called the **marginal probability** or **evidence**

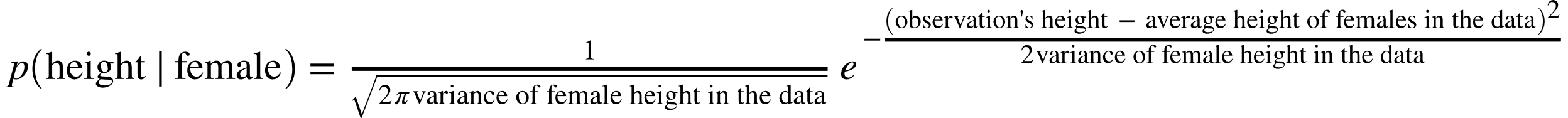
In a bayes classifier, we calculate the posterior (technically we only calculate the numerator of the posterior, but ignore that for now) for every class for each observation. Then, classify the observation based on the class with the largest posterior value. In our example, we have one observation to predict and two possible classes (e.g. male and female), therefore we will calculate two posteriors: one for male and one for female.



Now let us unpack the top equation a bit:

* P(male) is the prior probabilities. It is, as you can see, simply the probability an observation is male. This is just the number of males in the dataset divided by the total number of people in the dxataset.
* P(height | female) p(weight | female) p(foot size | female) is the likelihood. Notice that we have unpacked **person’s data** so it is now every feature in the dataset. The "gaussian" and "naive" come from two assumptions present in this likelihood:

1. If you look each term in the likelihood you will notice that we assume each feature is uncorrelated from each other. That is, foot size is independent of weight or height etc.. This is obviously not true, and is a "naive" assumption - hence the name "naive bayes."
2. Second, we assume have that the value of the features (e.g. the height of women, the weight of women) are normally (gaussian) distributed. This means that p(height | female) is calculated by inputing the required parameters into the probability density function of the normal distribution:



* Marginal probability is probably one of the most confusing parts of bayesian approaches. In toy examples (including ours) it is completely possible to calculate the marginal probability. However, in many real-world cases, it is either extremely difficult or impossible to find the value of the marginal probability (explaining why is beyond the scope of this tutorial). This is not as much of a problem for our classifier as you might think. Why? Because we don't care what the true posterior value is, we only care which class has a the highest posterior value. And because the marginal probability is the same for all classes 1) we can ignore the denominator, 2) calculate only the posterior's numerator for each class, and 3) pick the largest numerator. That is, we can ignore the posterior's denominator and make a prediction solely on the relative values of the posterior's numerator.

Steps for calculating all the different parts of the bayes equations :-

* Calculate Priors
* Calculate Likelihood
* Apply Bayes classifier to new data point