Naive Bayes – Witten et al. (2011).

1. Cover: bla
2. Reading
3. Outline

4--- simple technique is to use all attributes and allow them to make contributions to the decision that are equally important and independent of one another, given the class. This is unrealistic, of course: what makes real-life datasets interesting is that the attributes are certainly not equally important or independent. But it leads to a simple scheme that again works surprisingly well in practice.
5--- Table shows a summary of the weather data obtained by counting how many times each attribute–value pair occurs with each value (yes and no) for play. The cells in the first row of the new table simply count these occurrences for all possible values of each attribute, and the play figure in the final column counts the total number of occurrences of yes and no. In the lower part of the table, we rewrote the same information in the form of fractions, or observed probabilities. For example, of the nine days that play is yes, outlook is sunny for two, yielding a fraction of 2/9. For play the fractions are different: they are the proportion of days that play is yes and no, respectively.
6--- Now suppose we encounter a new example with the values that are shown in Table. We treat the five features in Table 4.2—outlook, temperature, humidity, windy, and the overall likelihood that play is yes or no—as equally important, independent pieces of evidence and multiply the corresponding fractions. Looking at the outcome yes gives: (read) The fractions are taken from the yes entries in the table according to the values of the attributes for the new day, and the final 9/14 is the overall fraction representing the proportion of days on which play is yes. A similar calculation for the outcome no leads to (read) This indicates that for the new day, no is more likely than yes—four times more likely. The numbers can be turned into probabilities by normalizing them so that they sum to 1: probability of yes (read) probability of no (read)
7---- This simple and intuitive method is based on Bayes's rule of conditional probability. Bayes's rule says that if you have a hypothesis H and evidence E that bears on that hypothesis, then (read) We use the notation that Pr[A] denotes the probability of an event A and that Pr[A|B] denotes the probability of A conditional on another event B. The hypothesis H is that play will be, say, yes, and Pr[H|E] is going to turn out to be 20.5%, just as determined previously. The evidence E is the particular combination of attribute values for the new day, outlook = sunny, temperature = cool, humidity = high, and windy = true.
8---- Let's call these four pieces of evidence E1, E2, E3, and E4, respectively. Assuming that these pieces of evidence are independent (given the class), their combined probability is obtained by multiplying the probabilities: (read)
9--- Don't worry about the denominator: we will ignore it and eliminate it in the final normalizing step when we make the probabilities of yes and no sum to 1, just as we did previously. The Pr[yes] at the end is the probability of a yes outcome without knowing any of the evidence E, that is, without knowing anything about the particular day referenced—it's called the prior probability of the hypothesis H. In this case, it's just 9/14, because 9 of the 14 training examples had a yes value for play. Substituting the

fractions in Table 4.2 for the appropriate evidence probabilities leads to (read) just as we calculated previously. Again, the Pr[E] in the denominator will disappear when we normalize. This method goes by the name of Naïve Bayes, because it's based on Bayes's rule and "naïvely" assumes independence—it is only valid to multiply probabilities when the events are independent. The assumption that attributes are independent (given the class) in real life certainly is a simplistic one. But despite the disparaging name, Naïve Bayes works very well when tested on actual datasets, particularly when combined with some of the attribute selection procedures introduced shortly that eliminate redundant, and hence nonindependent, attributes.

10 --- One thing that can go wrong with Naïve Bayes is that if a particular attribute value does not occur in the training set in conjunction with every class value, things go badly awry. Suppose in the example that the training data was different and the attribute value outlook = sunny had always been associated with the outcome no. Then the probability of outlook = sunny given a yes, that is, Pr[outlook = sunny | yes], would be zero, and because the other probabilities are multiplied by this the final probability of yes would be zero no matter how large they were. Probabilities that are zero hold a veto over the other ones. This is not a good idea. But the bug is easily fixed by minor adjustments to the method of calculating probabilities from frequencies. For example, the upper part of Table 4.2 shows that for play = yes, outlook is sunny for two examples, overcast for four, and rainy for three, and the lower part gives these events probabilities of 2/9, 4/9, and 3/9, respectively. Instead, we could add 1 to each numerator and compensate by adding 3 to the denominator, giving probabilities of 3/12, 5/12, and 4/12, respectively. This will ensure that an attribute value that occurs zero times receives a probability which is nonzero, albeit small. The strategy of adding 1 to each count is a standard technique called the Laplace estimator after the great eighteenth-century French mathematician Pierre Laplace.

11 ---- Although it works well in practice, there is no particular reason for adding 1 to the counts: we could instead choose a small constant m and use (read) The value of m, which was set to 3, effectively provides a weight that determines how influential the a priori values of 1/3, 1/3, and 1/3 are for each of the three possible attribute values. A large m says that these priors are very important compared with the new evidence coming in from the training set, whereas a small one gives them less influence. Finally, there is no particular reason for dividing m into three equal parts in the numerators: we could use (read) instead, where p1, p2, and p3 sum to 1. Effectively, these three numbers are a priori probabilities of the values of the outlook attribute being sunny, overcast, and rainy, respectively. This is now a fully Bayesian formulation where prior probabilities have been assigned to everything in sight. It has the advantage of being completely rigorous, but the disadvantage that it is not usually clear just how these prior probabilities should be assigned. In practice, the prior probabilities make little difference provided that there are a reasonable number of training instances, and people generally just estimate frequencies using the Laplace estimator by initializing all counts to one instead of to zero.

12--- One of the really nice things about the Bayesian formulation is that missing values are no problem at all. For example, if the value of outlook were missing in the example of Table 4.3, the calculation would simply omit this attribute, yielding (read) These two numbers are individually a lot higher than they were before, because one of the fractions is missing. But that's not a problem because a fraction is missing in both cases, and these likelihoods are subject to a further normalization process. This yields probabilities for yes and no of 41% and 59%, respectively. If a value is missing in a training instance, it is

simply not included in the frequency counts, and the probability ratios are based on the number of values that actually occur rather than on the total number of instances.
13--- Numeric values are usually handled by assuming that they have a “normal” or “Gaussian” probability distribution. Table 4.4 gives a summary of the weather data with numeric features from Table 1.3. For nominal attributes, we calculated counts as before, and for numeric ones we simply listed the values that occur. Then, whereas we normalized the counts for the nominal attributes into probabilities, we calculated the mean and standard deviation for each class and each numeric attribute. Thus the mean value of temperature over the yes instances is 73, and its standard deviation is 6.2. The mean is simply the average of the preceding values, that is, the sum divided by the number of values. The standard deviation is the square root of the sample variance, which we can calculate as follows: subtract the mean from each value, square the result, sum them together, and then divide by one less than the number of values. After we have found this sample variance, find its square root to determine the standard deviation. This is the standard way of calculating mean and standard deviation of a set of numbers (the “one less than” is to do with the number of degrees of freedom in the sample, a statistical notion that we don’t want to get into here). The probability density function for a normal distribution with mean m and standard deviation s is given by the rather formidable expression:(read)
14--- But fear not! All this means is that if we are considering a yes outcome when temperature has a value, say, of 66, we just need to plug x = 66, m = 73, and s = 6.2 into the formula. So the value of the probability density function is (read) By the same token, the probability density of a yes outcome when humidity has value, say, of 90 is calculated in the same way: (read)
15--- Using these probabilities for the new day in Table 4.5 yields (read) which leads to probabilities (read) These figures are very close to the probabilities calculated earlier for the new day in Table 4.3, because the temperature and humidity values of 66 and 90 yield similar probabilities to the cool and high values used before. The normal-distribution assumption makes it easy to extend the Naïve Bayes classifier to deal with numeric attributes. If the values of any numeric attributes are missing, the mean and standard deviation calculations are based only on the ones that are present.
16 --- The probability density function for an event is very closely related to its probability. However, it is not quite the same thing. If temperature is a continuous scale, the probability of the temperature being exactly 66—or exactly any other value, such as 63.14159262—is zero. The real meaning of the density function f(x) is that the probability that the quantity lies within a small region around x, say, between x - e/2 and x + e/2, is e f(x).What we have written above is correct if temperature is measured to the nearest degree and humidity is measured to the nearest percentage point. You might think we ought to factor in the accuracy figure e when using these probabilities, but that’s not necessary. The same e would appear in both the yes and no likelihoods that follow and cancel out when the probabilities were calculated.
17--- One important domain for machine learning is document classification, in which each instance represents a document and the instance’s class is the document’s topic. Documents might be news items and the classes might be domestic news, overseas news, financial news, and sport. Documents are characterized by the words that appear in them, and one way to apply machine learning to document classification is to treat the presence or absence of each word as a Boolean attribute. Naïve Bayes is a popular technique for this application because it is very fast and quite accurate. However, this does not take into account the number of occurrences of each word, which is potentially

useful information when determining the category of a document. Instead, a document can be viewed as a bag of words—a set that contains all the words in the document, with multiple occurrences of a word appearing multiple times (technically, a set includes each of its members just once, whereas a bag can have repeated elements). Word frequencies can be accommodated by applying a modified form of Naïve Bayes that is sometimes described as multinominal Naïve Bayes. Suppose n1, n2, . . . , nk is the number of times word i occurs in the document, and P1, P2, . . . , Pk is the probability of obtaining word i when sampling from all the documents in category H. Assume that the probability is independent of the word's context and position in the document. These assumptions lead to a multinomial distribution for document probabilities. For this distribution, the probability of a document E given its class H—in other words, the formula for computing the probability Pr[E|H] in Bayes's rule—is (read) where N = n1 + n2 + . . . + nk is the number of words in the document. The reason for the factorials is to account for the fact that the ordering of the occurrences of each word is immaterial according to the bag-of-words model. Pi is estimated by computing the relative frequency of word i in the text of all training documents pertaining to category H. In reality there should be a further term that gives the probability that the model for category H generates a document whose length is the same as the length of E (that is why we use the symbol ª instead of =), but it is common to assume that this is the same for all classes and hence can be dropped.

18 --- For example, suppose there are only the two words, yellow and blue, in the vocabulary, and a particular document class H has Pr[yellow|H] = 75% and Pr[blue|H] = 25% (you might call H the class of yellowish green documents). Suppose E is the document blue yellow blue with a length of N = 3 words. There are four possible bags of three words. One is {yellow yellow yellow}, and its probability according to the preceding formula is (read) The other three, with their probabilities, are (read) Here, E corresponds to the last case (recall that in a bag of words the order is immaterial); thus its probability of being generated by the yellowish green document model is 9/64, or 14%. Suppose another class, very bluish green documents (call it H¢), has Pr[yellow | H¢] = 10%, Pr[blue | H¢] = 90%. The probability that E is generated by this model is 24%. If these are the only two classes, does that mean that E is in the very bluish green document class? Not necessarily. Bayes's rule, given earlier, says that you have to take into account the prior probability of each hypothesis. If you know that in fact very bluish green documents are twice as rare as yellowish green ones, this would be just sufficient to outweigh the preceding 14% to 24% disparity and tip the balance in favor of the yellowish green class. The factorials in the preceding probability formula don't actually need to be computed because—being the same for every class—they drop out in the normalization process anyway. However, the formula still involves multiplying together many small probabilities, which soon yields extremely small numbers that cause underflow on large documents. The problem can be avoided by using logarithms of the probabilities instead of the probabilities themselves. In the multinomial Naïve Bayes formulation a document's class is determined not just by the words that occur in it but also by the number of times they occur. In general it performs better than the ordinary Naïve Bayes model for document classification, particularly for large dictionary sizes.

19--- Naïve Bayes gives a simple approach, with clear semantics, to representing, using, and learning probabilistic knowledge. Impressive results can be achieved using it. It has often been shown that Naïve Bayes rivals, and indeed outperforms, more sophisticated classifiers on many datasets. The moral is, always try the simple things first. Repeatedly in machine learning people have eventually, after an extended struggle, obtained good

results using sophisticated learning methods only to discover years later that simple methods such as 1R and Naïve Bayes do just as well—or even better. There are many datasets for which Naïve Bayes does not do so well, however, and it is easy to see why. Because attributes are treated as though they were completely independent, the addition of redundant ones skews the learning process. As an extreme example, if you were to include a new attribute with the same values as temperature to the weather data, the effect of the temperature attribute would be multiplied: all of its probabilities would be squared, giving it a great deal more influence in the decision. If you were to add 10 such attributes, then the decisions would effectively be made on temperature alone. Dependencies between attributes inevitably reduce the power of Naïve Bayes to discern what is going on. They can, however, be ameliorated by using a subset of the attributes in the decision procedure, making a careful selection of which ones to use. Chapter 7 shows how. The normal-distribution assumption for numeric attributes is another restriction on Naïve Bayes as we have formulated it here. Many features simply aren't normally distributed. However, there is nothing to prevent us from using other distributions for the numeric attributes: there is nothing magic about the normal distribution. If you know that a particular attribute is likely to follow some other distribution, standard estimation procedures for that distribution can be used instead. If you suspect it isn't normal but don't know the actual distribution, there are procedures for "kernel density estimation" that do not assume any particular distribution for the attribute values. Another possibility is simply to discretize the data first.

4. the end: re-list the topics, final words?