Classification: Bayesian Methods

1 Bayes' Rule

Bayes' rule states:

$$P(B = b_j | A = a_i) = \frac{P(A = a_i | B = b_j) P(B = b_j)}{P(A = a_i)}$$

Bayes' rule is useful in practice because there are many cases where we do have estimates of the three probabilities needed to compute the fourth.

For example, a doctor may know that: meningitis causes the patient to have a stiff neck 50% of the time; the (unconditional) probability that a patient has meningitis is 0.00002; and the (unconditional) probability that a patient has a stiff neck is 0.05.

$$\begin{aligned} P(\mathit{stiffNeck} = \text{true} | \mathit{meningitis} = \text{true}) &= 0.5 \\ P(\mathit{meningitis} = \text{true}) &= 0.00002 \\ P(\mathit{stiffNeck} = \text{true}) &= 0.05 \end{aligned}$$

By Bayes' rule,

$$\begin{aligned} P(\mathit{meningitis} = \text{true} | \mathit{stiffNeck} = \text{true}) &= \frac{P(\mathit{stiffNeck} = \text{true} | \mathit{meningitis} = \text{true}) P(\mathit{meningitis} = \text{true})}{P(\mathit{stiffNeck} = \text{true})} \\ &= \frac{0.5 \times 0.00002}{0.05} = 0.0002 \end{aligned}$$

i.e. we expect 1 in 5000 patients who have a stiff neck will have meningitis.

An obvious question is: why might we be likely to have an estimated conditional probability in one direction, but not the other? Why might we have an estimate of $P(\mathit{stiffNeck} = \text{true} | \mathit{meningitis} = \text{true})$ but not of $P(\mathit{meningitis} = \text{true} | \mathit{stiffNeck} = \text{true})$? The former (the probability of a stiff neck given that the patient has meningitis) reflects the way that meningitis works, i.e. that it causes stiff necks. This probability estimate is therefore largely unaffected by changing circumstances. The latter (the probability that the patient has meningitis given that s/he has a stiff neck — in other words how predictive a stiff neck is of meningitis), is more fragile knowledge: it depends on all the other, possibly changing, factors that affect why people are getting stiff necks.

2 Using Bayes' Rule in Classification

Recall from earlier that to classify a new instance, we do the following:

1. For each class label $cl \in L$, compute the conditional probability for that class given the description of the new instance:
$$P(\mathit{class} = cl | A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$$
2. Return the class label with the highest probability.

Previously, we computed the conditional probability from the joint probability distribution. But we are now assuming that this distribution is not available. But we can use Bayes' rule to rewrite the expression so that it uses probabilities that might be available:

$$P(\mathit{class} = cl | A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n) = \\ \frac{P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n | \mathit{class} = cl) P(\mathit{class} = cl)}{P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)}$$

Do we have these three probabilities?

- We can easily estimate $P(\mathit{class} = cl)$ from a dataset.
- As we saw in the previous lecture, we don't really need the divisor because it will be the same in every computation, i.e. for each $cl \in L$.
- But there is a problem when it comes to the conditional probabilities $P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n | \mathit{class} = cl)$. Again, the approach doesn't scale up. The number of conditional probabilities we need to have estimated is the number of possible instance descriptions multiplied by the number of classes. This will often be too large. To see this by way of an example, let's assume all n attributes are Boolean-valued; the number of different instance descriptions is then 2^n. We would need to estimate $2^n \times |L|$ conditional probabilities. We might as well go back to using the full joint probability distribution.

How can we rescue ourselves?

3 Bayesian Classification using the Naïve Bayes Distribution

In classification above, we had the problem of obtaining probabilities $P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n | \mathit{class} = cl)$. Maybe the notions of independence or conditional independence can help.

It is unlikely that statements about the attributes that describe instances will be independent, i.e. it is unlikely that $A_1 = a_1$ and $A_2 = a_2$ and ... and $A_n = a_n$ are independent.

But it may be the case that they are conditionally independent given $\mathit{class} = cl$. If so, we can rewrite:

$$P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n | \mathit{class} = cl) = \\ P(A_1 = a_1 | \mathit{class} = cl) \times P(A_2 = a_2 | \mathit{class} = cl) \times \ldots \times P(A_n = a_n | \mathit{class} = cl)$$

So, to recap, what we want to compute is:

$$P(\mathit{class} = cl | A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$$

which, by Bayes' rule, becomes

$$= \frac{P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n | \mathit{class} = cl) \times P(\mathit{class} = cl)}{P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)}$$

and, if the attributes are conditionally independent given the class, we get

$$= \frac{P(A_1 = a_1 | \mathit{class} = cl) \times P(A_2 = a_2 | \mathit{class} = cl) \times \ldots \times P(A_n = a_n | \mathit{class} = cl) \times P(\mathit{class} = cl)}{P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)}$$

Is it now more feasible to obtain the probabilities?

- As above, we can easily estimate $P(\mathit{class} = cl)$ from a dataset.
- As above, in classification we don't really need the divisor.
- Due to conditional independence, we no longer need $P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n | \mathit{class} = cl)$ for each class. Instead, we need $P(A_1 = a_1 | \mathit{class} = cl)$, $P(A_2 = a_2 | \mathit{class} = cl)$, ..., $P(A_n = a_n | \mathit{class} = cl)$ for each class. These are easily computed from datasets. Are there fewer probabilities to estimate? Yes! Consider n Boolean-valued attributes and $|L|$ classes. Originally, we needed $2^n \times |L|$ conditional probabilities. Now we need $2n \times |L|$ conditional probabilities.

Of course, this all relies on the attribute-value pairs being conditionally independent given the class. However, this approach is often used even when the attribute-value pairs are not conditionally independent. It can work surprisingly well, even in these applications. Classifiers that work this way are often called *naïve Bayes classifiers*: they are 'naïve' because of the simplifying assumption.

4 Summary of Naïve Bayes Classifiers

Learning step. This requires a dataset. For each instance, the dataset must give the instance's attribute values and its class. Learning that uses data of this kind (i.e. where each instance is already associated with its corresponding output) is called *supervised learning*. This is a phrase we will define properly in a future lecture.

From the dataset, we must estimate

- $P(\mathit{class} = cl)$ for each class $cl \in L$.
- $P(A = a_i | \mathit{class} = cl)$ for each attribute-value pair $A = a_i$ and class $cl \in L$.

Classification step. A new instance $\{A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n\}$ is classified as follows:

1. Compute, for each class $cl \in L$,

   $P(\mathit{class} = c | A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$

   which, by Bayes' rule and the simplifying assumption of conditional independence, is computed as

   $$= \frac{P(A_1 = a_1|\mathit{class} = cl) \times P(A_2 = a_2|\mathit{class} = cl) \times \ldots \times P(A_n = a_n|\mathit{class} = cl) \times P(\mathit{class} = cl)}{P(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)}$$

   (And, remember, we don't need the divisor.)
2. Return the class label with the highest probability.

The learning step is done once-and-for-all, in advance, to obtain the probabilities. In principle, once we have obtained the probabilities, the dataset can be thrown away. Then, as each new instance arises, we can classify it using these probabilities.

5 Example

We are given the following dataset:

shape	*colour*	*size*	*class*
circle	blue	large	+
circle	red	medium	-
circle	red	large	-
square	blue	small	-
square	red	small	-
square	red	medium	+
square	blue	medium	+
square	blue	large	-
triangle	red	small	-
triangle	red	large	+
triangle	blue	medium	+

In the lecture we will use the naïve Bayes classifier to classify the new instance $\{\mathit{shape} = \mathit{circle}, \mathit{colour} = \mathit{blue}, \mathit{size} = \mathit{medium}\}$.

In the lecture we will also answer:

- Are $\mathit{colour} = \mathit{red}$ and $\mathit{size} = \mathit{large}$ independent?
- Are $\mathit{colour} = \mathit{red}$ and $\mathit{size} = \mathit{large}$ conditionally independent given $\mathit{class} = +$?

6 Discussion

- The naïve Bayes classifier is competitive with other classification methods in many cases. In some cases, it even outperforms other methods. For classifying textual documents (e.g. into spam and ham), it is among the more effective algorithms. It is also scalable: it can easily handle thousands of attributes.
- It is possible that some of the conditional probabilities $P(A = a_i|\mathit{class} = cl)$ that we compute from the dataset will be zero. If zero probabilities are ever used in classification, then the whole probability of that class will evaluate to zero and that class will be ruled out. A simple solution is to replace the zero by a small constant, e.g. $0.5/n$ or $P(\mathit{class} = cl)/n$ where n is the number of instances in the dataset. Others apply a correction to all probabilities. For example, instead of computing

  $$\frac{n^{cl}_{A=a_i}}{n^{cl}}$$

  where n^{cl} is the number of instances in which $\mathit{class} = cl$ and $n^{cl}_{A=a_i}$ is the number *of these* in which $A = a_i$, they compute

  $$\frac{n^{cl}_{A=a_i} + m}{n^{cl} + m}$$

  where m is a small positive integer.
- If you have a large number of attributes, then the product of the conditional probabilities, is likely to underflow in floating point maths. To avoid this problem, instead of computing and comparing the following

  $$P(A_1 = a_1|\mathit{class} = cl) \times P(A_2 = a_2|\mathit{class} = cl) \times \ldots \times P(A_n = a_n|\mathit{class} = cl) \times P(\mathit{class} = cl),$$

  (from above but without the divisor), you use logarithms

  $$\log P(A_1 = a_1|\mathit{class} = cl) + \log P(A_2 = a_2|\mathit{class} = cl) + \ldots + \log P(A_n = a_n|\mathit{class} = cl) + \log P(\mathit{class} = cl)$$

  (This is another reason why we cannot allow probabilities to be zero: you cannot take the log of zero.)
- We've assumed discrete values for our attributes. Where this does not hold, you can use discretisation. There are also methods for handling real-valued attributes directly.
- If there are many candidate attributes, it may be useful to restrict attention to only a subset of the attributes. Various methods have been proposed for automatically determining which subset to use.
- There is an enormous amount of work now on *Bayesian networks*, which provide a way of explicitly representing dependencies, thus allowing the removal of some of the independence assumptions made by naïve Bayesian methods.