CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Machine Learning

Machine Learning

Lecture: Bayesian Classification

Ted Scully

Contents

Review of Basic Concepts

- Over the next few slides we will review some basic probability concepts and the use the following sample dataset to help illustrate.

ID	Headache	Fever	Vomiting	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

Machine Learning for Predictive Analytics – Kelleher et al.

Review of Basic Concepts and Terminology

- An **event** defines an **assignment of values** to the **features** in the domain; these assignments may define values for all the features in the domain (e.g. a full row in the dataset) or just to one or more features (Fever = True).

- A **probability function** is a function that takes an event (an assignment of values to features) as a parameter and returns the likelihood of that event (P(Fever = True)).

- The value returned by a probability function for an event is simply the **relative frequency** of that event in the dataset
- In other words, how often the event happened divided by how often it could have happened.

Review of Basic Concepts

- **Prior probability** (**Unconditional probability**) : The probability of an event that is not dependent on any other feature.
  - The count of all the rows in the dataset where the feature is assigned the relevant value divided by the number of rows in the dataset.

- **Joint probability**: The probability of two or more events happening together.
  - The number of rows in the dataset where the set of assignments listed in the joint event holds divided by the total number of rows in the dataset.

- **Posterior Probability** (**Conditional Probability**): The probability of an event where one or more other events are known to have happened.
  - The number of rows in the dataset where both events are true divided by the number of rows in the dataset where just the given event is true.

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

$P(h) = ?$

$P(m|h) = ?$

$P(m, h) = ?$

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

$P(h) = ?$

$P(m|h) = ?$

$P(m, h) = ?$

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

$P(h) = ?$

$P(m|h) = ?$

$P(m, h) = ?$

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

$P(h) = ?$

$P(m|h) = ?$

$P(m, h) = ?$

$$P(h) = \frac{|\{\mathbf{d}_{11}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{37}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{54}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{89}, \mathbf{d}_{92}\}|} = \frac{7}{10} = 0.7$$

$$P(m|h) = \frac{|\{\mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{92}\}|} = \frac{2}{7} = 0.2857$$

$$P(m, h) = \frac{|\{\mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{37}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{54}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{89}, \mathbf{d}_{92}\}|} = \frac{2}{10} = 0.2$$

Review of Basic Concepts

- **Probability Distribution** : For all the possible values of a feature it describes the probability of the feature taking that value.
- A probability distribution of a categorical feature is a vector that lists the probabilities associated with the values in the domain of the feature.

- **Joint Probability Distribution** : A matrix where each cell in the matrix lists the probability for one of the events in the sample space defined by the combination of feature values. Put another way it give us an exhaustive list of probabilities for all possible combination of the feature values.

- A **full joint probability distribution** is simply a joint probability distribution over all the features in a domain.

Review of Basic Concepts

$$\mathbf{P}(H,F,V,M) = \begin{bmatrix} P(h,f,v,m), & P(\neg h,f,v,m) \\ P(h,f,v,\neg m), & P(\neg h,f,v,\neg m) \\ P(h,f,\neg v,m), & P(\neg h,f,\neg v,m) \\ P(h,f,\neg v,\neg m), & P(\neg h,f,\neg v,\neg m) \\ P(h,\neg f,v,m), & P(\neg h,\neg f,v,m) \\ P(h,\neg f,v,\neg m), & P(\neg h,\neg f,v,\neg m) \\ P(h,\neg f,\neg v,m), & P(\neg h,\neg f,\neg v,m) \\ P(h,\neg f,\neg v,\neg m), & P(\neg h,\neg f,\neg v,\neg m) \end{bmatrix}$$

We can use the joint probability distribution to calculate conditional probabilities. For example, calculate the **P(h | f)** then you could sum up all values from the distribution where h and f are True. These are

- P(h, f, v, m)
- P(h, f, v, ¬m)
- P(h, f, ¬v, m)
- P(h, f, ¬v, ¬m)

Review of Basic Concepts

- Can you think of any problem we might encounter with calculating a joint probability distribution as we increase the number of features (and the number of values for those features also increases).

Review of Basic Concepts

- Unfortunately the size of a **full joint probability distribution** grows exponentially as the number of features and the number of values in the domain of the features grow. Consequently, they are difficult to generate.

- Remember computing each probability value in the joint probability distribution requires a set of instances.

- As we add additional features the size of the **distribution grows exponentially** but so too does the **size of the dataset required** to generate the joint probability distribution.

- Therefore, for domains of any reasonable complexity it is **not tractable** to build a full joint probability distribution.

Product Rule and Chain Rule

- The product rule is shown below. In this form it allows us to calculate the joint probability of two events.

$$P(a, b) = P(a|b)\,P(b) = P(b|a)\,P(a)$$

- So let's calculate the joint probability of P(m, h)

- P(m,h) = P(m|h) * P(h)

$$P(h) = 0.7$$

$$P(m|h) = 0.2857$$

$$P(m, h) = 0.2$$

Product Rule and Chain Rule

- *P(a, b) = P(a|b) P(b) = P(b|a) P(a)*
- We can extend the product rule to a more general form called the chain rule.
- P(a, b,c) = P(a|b,c) P(b,c) = P(a|b,c) P(b|c) P(c)
- So this can be extended
- P(a, b, c..., z) = P(a| b,c, ..z) P(b| c, d, ..z) …P(x|y,z) P(y|z)P(z)
- The chain rule is useful because it facilitates the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities.

Bayes Rule

- Over the next few slides we are going to look at Bayes rule.
- We are going to look at this in a classification setting and I'm going to use the following notation.
- When using the **notation** ***c*** we refer to a specific class
- When using ***d*** we are referring to a new data instance (data to be classified).

Bayes' Rule

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Bayes' Rule can be easily derived from the product rule

P(c , d) = **P(c | d) P(d) = P(d|c)P(c)**

Divide by P(d): **P(c | d) = P(d | c) P(c)/P(d)**

Contents

Classification Example

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

- The table below shows a database of Christian names for workers that work in a particular company. If I pick an employee from the company and their name is Joe, **what is the probability that this individual is male or female**. The class in this problem is the Sex and the attribute/feature is the Name of the individual. This is a trivial problem but we can use Bayes to help solve it.

Name	Sex
Joe	**Female**
Jim	Male
Joe	**Male**
Ted	Male
Mary	Female
Joe	**Male**
Carol	Female
Emily	Female

Classification Example

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Name	Sex
Joe	**Female**
Jim	Male
Joe	**Male**
Ted	Male
Mary	Female
Joe	**Male**
Carol	Female
Emily	Female

- We must calculate the probability for each class and then adopt the class with the highest probability

Classification Example

Name	Sex
Joe	Female
Jim	Male
Joe	Male
Ted	Male
Mary	Female
Joe	Male
Carol	Female
Emily	Female

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

- **P(Female | Joe)** = P(Joe | Female)P(Female) / P(Joe)
  = (1/4)(4/8) / (3/8) = **0.333**

- **P(Male | Joe)** = P(Joe | Male)P(Male) / P(Joe)
  = (2/4)(4/8) **/ (**3/8) = **0.666**

- Probability of Joe being a Male is higher, therefore we can classify the individual Joe as being a Male

Classification Example

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

- P(Female | Joe) = P(Joe | Female)P(Female) / P(Joe)
  = (1/4)(4/8) / (3/8) = 0.333

- P(Male | Joe) = P(Joe | Male)P(Male) / P(Joe)
  = (2/4)(4/8) / (3/8) = 0.666

In both of the above calculations I have included a **redundant component** as part of the calculation.

Classification Example

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

- P(Female | Joe) = P(Joe | Female)P(Female) / P(Joe)
  = (1/4)(4/8) / **(3/8)** = 0.333

- P(Male | Joe) = P(Joe | Male)P(Male) / P(Joe)
  = (2/4)(4/8) / **(3/8)** = 0.666

You might notice that for all these calculations, the denominators are identical—P(d). Thus, they are independent of the hypotheses.

Classification Example

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

- P(Female | Joe) = P(Joe | Female)P(Female) / P(Joe)
  = (1/4)(4/8) = 0.125

- P(Male | Joe) = P(Joe | Male)P(Male) / P(Joe)
  = (2/4)(4/8) = 0.25

More Formally

- More formally we examine each class and identify the class with the highlighted probability based on Bayes theorem.
- There we search for the class that maximises the probability of that class given the observed feature value d

$$= \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Bayes with Multiple Attributes/Features

- In the previous slides we considered Bayes classification when we had only a single feature (for example the name of an individual).
- But if it is to be useful then we have to apply in in situations where we have many other features as well such as age, height, weight etc.

Height	Weight	Name	Sex
X	X	Joe	Female
X	X	Jim	Male
X	X	Joe	Male
X	X	Ted	Male
X	X	Mary	Female
X	X	Joe	Male
X	X	Carol	Female
x	X	Emily	Female

Bayes with Multiple Attributes/Features

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

– We denote the n features
  $x_1, x_2, .., x_n$

Height	Weight	Name	Sex
X	X	Joe	Female
X	X	Jim	Male
X	X	Joe	Male
X	X	Ted	Male
X	X	Mary	Female
X	X	Joe	Male
X	X	Carol	Female
x	X	Emily	Female

Naïve Bayes

$$P(x_1, x_2, \ldots, x_n \mid c)$$

- We make the naïve assumption of conditional independence
  - We assume the feature probabilities $P(x_i \mid c_j)$ are independent given a class c.
  - Whether one features occurs given a class and whether another feature occurs given a class are independent

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$