CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Machine Learning

Machine Learning

Lecture: Introduction to Machine Learning

Ted Scully

Different Categories of ML

- *Machine learning algorithms can be divided into five main categories*

- *Supervised Learning Algorithms*
  - *Labelled Data Only*
- *Unsupervised Learning Algorithms*
  - *Unlabelled Data Only*
- *Semi Supervised Learning Algorithms*
  - *Labelled and Unlabelled*
- *Self Supervised Machine Learning Algorithms*
  - *Unlabelled data but supervised techniques*
- *Reinforcement Learning Algorithms*
  - *Learns an optimal policy of actions through interaction with it's environment*

Challenges in Machine Learning

- Insufficient amount of training data (particularly an issue with supervised algorithms).
- Non-representative training data.
- Issues with your data (missing values, feature importance, imbalance, etc).
- Issues with your models (which models and which hyper-parameters).
- Methodology for assessing the performance of the model.

Overfitting

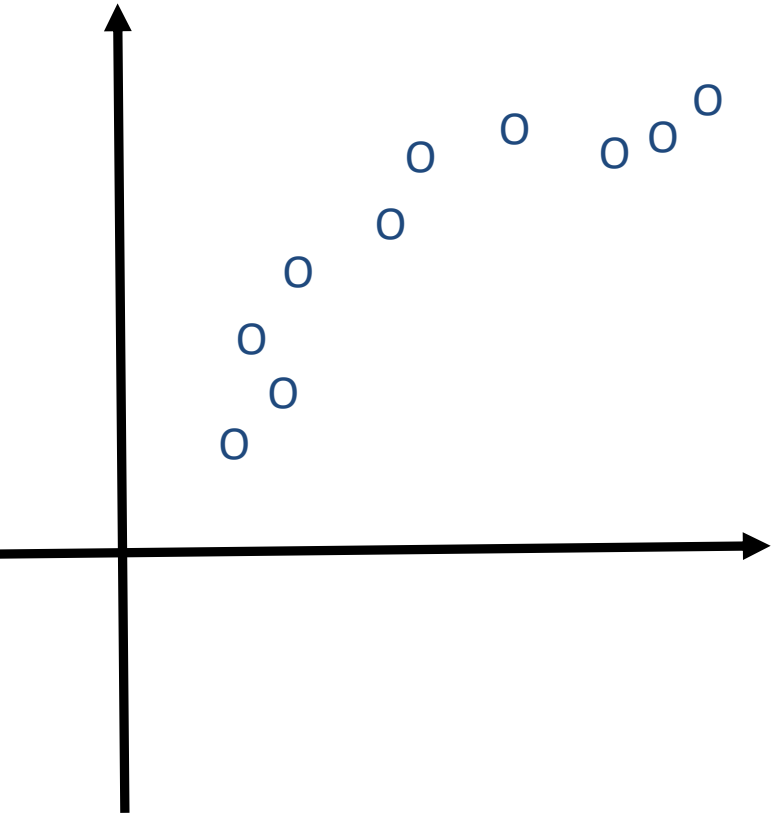
- Overfitting generally occurs when a **model/function is excessively complex and has fit too tightly to the training data.**

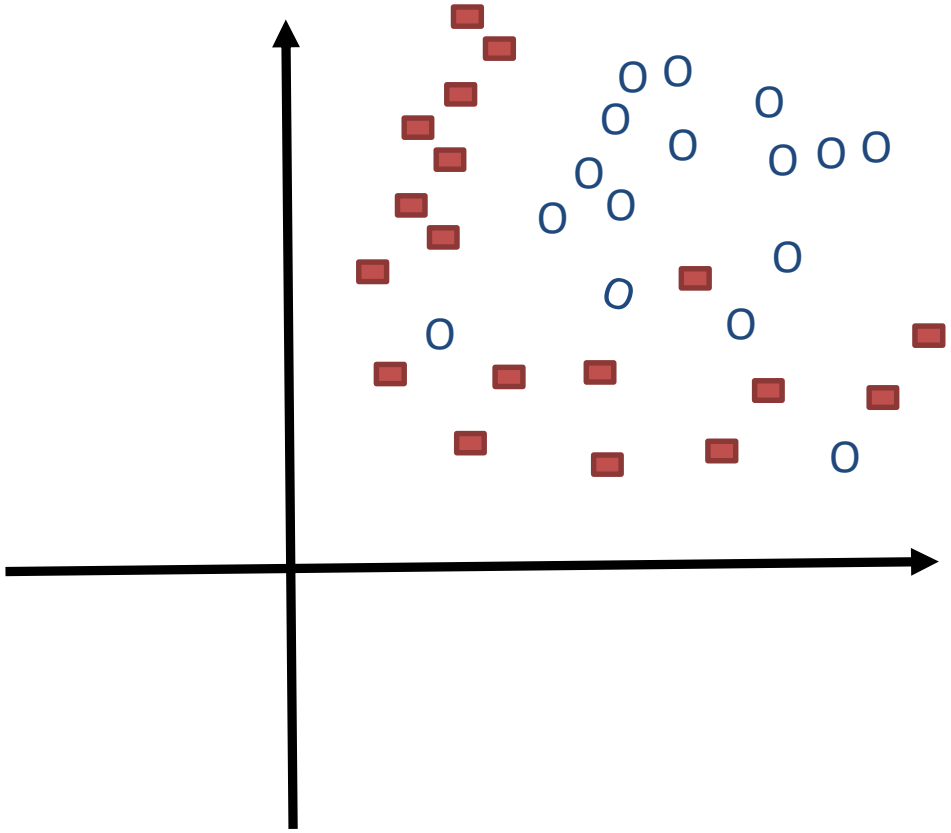
- A model/**function which has been overfit will generally have poor predictive performance on unseen data (it doesn't generalize well to unseen examples)**, as it can exaggerate minor fluctuations in the data.
  - A model is typically **trained** by maximizing its performance on some set of training data.
  - However, its overall performance is determined not by its performance on the training data but by its ability to perform well on **unseen data**.

- You can think of this as the difference between memorizing the data and generalizing from the data.

Underfitting

- As you might guess, underfitting is the opposite of overfitting: it occurs when your model is unable to learn the underlying structure of the data.





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

- **Unconditional probability** : Unconditional probability is simply the probability of the event occurring (not impacted by previous or future events). 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

$P(h) = ?$
$P(m, h) = ?$
$P(m|h) = ?$

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

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

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

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** : It gives us an exhaustive list of probabilities for all possible combinations of values for a specified set of features.
- A **full joint probability distribution** is simply a joint probability distribution over all the features in a domain.

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

Review of Basic Concepts

- Can you think of any problem we might encounter with calculating a full 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.

$P(h) = 0.7$

$P(m|h) = 0.2857$

Product Rule

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

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

- Using the product rule let's calculate the joint probability of P(m, h)

- P(m,h)

Bayes Rule

- Bayes rule can be derived from the product rule

$$P(c, d) = P(c|d)\,P(d) = P(d|c)\,P(c)$$

ML Workflow

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

P(Female | Joe) = P(Joe | Female)P(Female) / P(Joe)

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

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

P(Female | Joe) = P(Joe | Female)P(Female) / P(Joe)

= (1/4)(4/8) / (3/8) = **0.333**

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

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

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)

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

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

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**

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

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 there is a **redundant component**.

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

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

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

$$\underset{c \in C}{argmax} P(c|d)$$

$$\underset{c \in C}{argmax} (P(d|c)P(c))/P(d)$$

$$\underset{c \in C}{argmax} P(d|c)P(c)$$

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 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

$$\underset{c \in C}{argmax}(P(d|c)P(c))$$

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

– We denote the n features $x_1, x_2, \ldots x_n$

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

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

- 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.

$$P(x_1, x_2, \ldots x_n \,|c) = P(x_1|c).\, P(x_2|c).\, P(x_3|c) \;\ldots . P(x_n|c)$$

Naïve Bayes

- Therefore, we can reformulate our Bayesian classification equation as:

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

$$\underset{c \in C}{argmax}(P(c) \prod_{x \in X} P(x|c))$$

- Let's return to the previous example (with the multi-feature dataset).
- Assume we pick an employee with the following feature values

Name= "Joe", Weight = Light, Height = Tall.

  - Should we class this individual as a male or female? How would we work this out?

Assume we pick an employee with the following attribute values **Name= "Joe", Weight = Light, Height = Tall**.

Should we class this individual as a male or female? How would we solve this?

$$\underset{c \in C}{argmax} \left(P(c) \prod_{x \in X} P(x|c) \right)$$

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

P(Female | Name= "Joe", Weight = Light, Height = Tall) = P(Name = "Joe" | Female) * P(**Weight = Light** | Female) * P(Height = Tall | Female) * P(Female)

P(Male | Name= "Joe", Weight = Light, Height = Tall) = P(Name = "Joe" | Male) * P(**Weight = Heavy**| Male) * P(Height = Tall | Male) * P(Male)